$$T_m = \bigcap_{n \ge m}^{\infty} \Sigma_n .$$

$$p_1 p_2 \sum_{m=0}^{k} q_1^m q_2^{k-m}$$

$$\int_A$$

$$\int_\Omega d\mathbf{P} = 1$$

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] = 1$$

# Probability and Measure in Public Health

Lem Moyé, M.D., Ph.D.

# Probability and Measure in Public Health

Lem Moyé

Other books by Lem Moyé

- *Statistical Reasoning in Medicine: The Intuitive P–Value Primer*.
- *Difference Equations with Public Health Applications* (with Asha S. Kapadia).
- *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*
- *Finding Your Way in Science: How You Can Combine Character, Compassion, and Productivity in Your Research Career*
- *Probability and Statistical Inference: Applications, Computations, and Solutions* (with Asha S. Kapadia and Wenyaw Chan)
- *Statistical Monitoring of Clinical Research: Fundamentals for Investigators*
- *Statistical Reasoning in Medicine: The Intuitive P–Value Primer. 2nd Edition*
- *Face to Face with Katrina Survivors: A First Responder's Tribute*
- *Elementary Bayesian Biostatistics*
- *Saving Grace – A Novel*
- *Weighing the Evidence: Duality, Set, and Measure Theory in Clinical Research*
- *Catching Cold: Vol .1 Breakthrough – A Pre-COVID Novel*
- *Finding Your Way in Science: How You Can Combine Character, Compassion, and Productivity in Your Research Career- 2nd Edition*
- *Catching Cold: Vol 2. Redemption*

*To Dixie and the DELTS*

# Acknowledgments

My grandmother descended from the few North Carolina Cherokee who chose not to take the *Nunadautsun't* in the 18th century overland to the Oklahoma reservations. At the feet of her rocker, her grandchildren learned that life would be harder, yet better if we served as caretakers, both looking after – and being looked after – by others.

This philosophy infused the many classes where I taught probability in one format or another for over thirty years at the University of Texas School of Public Health in Houston. Every student for whom I have prepared those innumerable lectures deserves a measure of credit for this treatise. Teaching is a phenomenal way to learn.

During this time, my colleagues at the school taught and reminded me that public health did not reside wholly in the purview of probability. Demography, epidemiology, health care economics, environmental and behavioral sciences, as well as the frightening yet fascinating disease process were also core components of public health.

A special thanks to Purdue University Emeritus Professors Louis J. Cote, Burgess Davis, and KCS Pillai. During a critical time in my development, each went out of his way to ensure that I took all of the time that I actually required in order to absorb challenging theoretical concepts in probability. If not for them, I could not write this.

If probability was to maintain its relevance in public health, it must work within and support the interconnections of these fields. My thanks goes to Robert Hardy, Asha Kapadia, Barry Davis, Ralph Frankowski, Mort Hawkins, Fred Annegers, Rick Shekelle, Darwin Labarthe, Wenyaw Chan, Palmer Beasley, Guy Parcel, and Eric Boerwinkle for their steadfastness.

My work in cell therapy exposed me to new applications and problems in probability. While many of these ideas produce nothing of tangible value, they sharpened my critical thinking and articulation of mathematical ideas, while driving my conviction that probability was essential for clinical research and for public health deeper. My enduring thanks go to Robert Simari and the Cardiovascular Cell Therapy Research Network (CCTRN), NIH, as well as all of the physician scientists with and for whom I worked. They reminded me that the application of probability to public health while relevant also needed to be comprehensible while providing new perspectives of palpable value.

Three colleagues, Hulin Wu, Dejian Lai, and Hongzian Zhu encouraged me to give full treatment to the theoretical aspects of probability. Hongzian and I have several important discussions about the potential applications of measure theory.

One of my project managers, Shelly Sayre, raised probing questions about this work leading to its unique design. Her ideas, as always were both unanticipated and illuminating. All failings in their execution belong to me.

Finally, my dearest thanks go to Dixie, my wife, on whose personality, character, love, and common sense I have come to rely, and to my daughters Flora Ardon and Bella Scalise who, in their own irresistible manners and movements help to shape the new world for which we hunger.

# Preface

September 19th, 1979

On my 27th birthday, I sat in a probability class at Purdue University. The windows in the old wooden classroom were open, the blazing colors of the campus trees announcing the arrival of another vibrant Midwest fall.

Behind me, I saw five years of medical school and internship. I held a new license to practice medicine in Indiana. During that grinding time I had watched as, year after year, classmates and friends choose their career specialties. General surgery. Pediatrics. Internal medicine with a view to cardiology. Family practice. Ophthalmology.

And I had selected a path in ─ probability and statistics?

After patiently listening to my decision, my earnest medical friends and colleagues wrangled with me over this startling choice. Why, they asked, abandon medicine, a field that was so full of scientific promise?

"Have you really endured five years of medical training to not use what you have learned?" they asked,

Then, the waves of our career choices carried us to our respective postgraduate training centers.

I chose a path that was not only iconoclastic but, to my knowledge at the time, solitary and completely untried.*

It was likely unworkable, and probably a dead end.

Yet, I was compelled to walk it.

So, with my fellow physicians starting their first post internship year of residency in far off venues, here I sat in a probability class with students five years younger.

And that fall morning, my first in graduate school, as I glanced from the professor's discussion on the nature of sample spaces through the wide open window into a blue sky populated with cotton ball clouds, one thought resonated.

I had come home.

I have always enjoyed probability. Enjoyed learning about it, working through its innumerable exercises, gaining experience, always gathering the kindle that would ignite some illumination. Even as I fought my way through measure theoretic probability theory (without

* As of this date, I specifically know of only two other physicians who also had chosen to obtain a PhD in mathematics or statistics.

having taken the prerequisite analyses course!) I enjoyed its direction and was strengthened by my early futile exertions that at first yielded so little fruit.

This treatise is powered by several motivations, but the principal, most powerful one is that I cherish this field.

## Navigating this book

A second reason for this treatise is to incisively utilize some of the technological advances that have come with information management.

Technology has not yet substantially influenced the utility of modern day texts. I believe that a principal limiting feature of traditional texts is the fixed sequence of pages. There is typically only one path through a book ─ especially a math book – starts at the beginning and work your way to the end. Just as we read words sequentially, so we too read pages.

But not here, where it is hyperlinks and not page sequences that serve as bridges between sections.

In this the third decade of the twenty-first century, we are accustomed to navigating the web through these links. Here, we use them to move from section to section. And these different "linked sequences" generate many different paths through the material; each set of connections represents a different learning track. To a great extent, hyperlink sequences personalize the learning experience, permitting this text to meet the needs of different audiences. This is discussed in detail in here.

## Measure theory

Another one of my motivations for developing yet one more work in this field is the belief that the measure theoretic component of probability could be better integrated into the subject matter.

First, I must confess that you can learn much from this book while wholly avoiding the sections devoted to the measure theoretic treatment of probability. However, measure theory has great value. This is true even for the most elementary reader – if presented appropriately.

When I was in graduate school, I felt ambushed by measure theory. It was presented as its own field, separate and apart from traditional probability. It did not appear value added. After all, my fellow students and I had become comfortable with many computations in probability involving both discrete and continuous distributions and was able to compute some very complicated probabilities. The Riemann integral had been good enough so far! From this perspective, "measure theory" seemed like an unnecessary, disorienting, demanding affair.

After completing its introduction I did not feel like I was any better at computing probabilities, only better at computing integrals that I would never use (e.g., functions which uncountably many discontinuities). I was more confused than ever despite my attempts to learn it.

This negative if not harmful experience of mine and many biostatistical colleagues was spawned by the assumption that two years of calculus and real analysis was required as a prerequisite to understand the basics of measure theory. This heavy mathematical preamble fosters the sense that measure theory is a distant and obscure way to look at the notion of integration and summation when it is not.

The thesis of this treatise is that one needn't be a specialists in real and complex analysis to learn and use measure theory, any more than one need have a doctorate in electrical engineering to flip a light switch.

The concept of measure theory is at its heart simple, and can be understood by introductory students if explained using plain language and simple examples. This is what I have attempted in several of the tracts of this text.

Therefore this treatise will introduce measure theory without requiring a preceding course in real analysis. It is not my goal to supplant analysis – only to implement its basics in order to demonstrate that students can understand the concepts of measure theory sufficient for its application to health care probability calculation.

## History and Probability

Finally, in order to indulge my penchant for history, I have included hyperlinked essays on many of the early giants of probability. I do this because sometimes taking a break from the intense mathematics is needed (of course, there are already ample distractions on our personal devices).

But, it is also true that, as one reviews these brief biographies, it becomes clear that the early probability giants were not giants at all but mere mortals like you and I. They had their share of family, financial, and political issues like we do. One lost the use of his writing arm, but during this time, produced half of the papers of his career. Another had been left to swing from a door for hours at a time, day after day as a child. Another, who became the force behind integral calculus, respected his father so much that, when he was in his twenties and considered changing his field from religious studies to mathematics, asked his father's permission. Yet another interrupted his studies to work as a railroad conductor, writing mathematical papers in his spare time.

These "giants" in probability struggled, got distracted, were deflected, and made mistakes.

Just as we do.

And they all needed help. Some found the right teacher. Other had their teachers raise money for them. Yet another met an influential friend while in the army. There is no doubt that they were intelligent, but that was not enough. Only when the environment was right, and they had adequate support did the intelligence gain traction and lead to the content that we all admire. Without that support, the result was sometimes lethal.

What was true for them is also true for us.

Finally, although I have done my best to purge this text of errors, my best is never good enough. If you find them, please let me know, and I will correct them and can quickly issue a new version. Such is the result of my own fallibility and the resilience and power of electronic documents.

Lem Moyé
Chandler, AZ
August, 2022
Probability@PrincipalEvidence.com

# Contents

# Introduction

My answer to the omnipresent question all authors of probability texts must face, "Why do we need yet one more book on probability?" is that this is not a traditional book.

Typically a book is a set of sequential chapters that flows logically and that all readers are directed to follow.

This book uses hypertexting to concatenate different mathematical developments in a way that makes sense to the reader, permitting them to put together a sequence at their level that will guide their mathematical development. Hypertexting permits student-specific resequencing of chapters.

For example, one such path covers the discrete probability distributions on an elementary level. Another path lies through more intense discussions. Each path begins with sections that serve as prerequisites, and ends with links that function well as sequels, enabling the reader to trace out their own path of exposure through this material.

Other hyperlinks allow the reader to refer to important prerequisites and refresher material, for example, the Poisson based death process provides links back to the binomial probability generating function for review.

The implication is that *Probability, Measure, and Public Health* is written to be a wholly electronic experience. My advice is to stay away from printed copies – they will only confuse you if you try to read more than a couple of contiguous sequences. This treatise was not written to be perused in the standard "just turn the page" format.

## Measure theory

The eBook format with its hyperlinks permits measure theory to be more completely integrated into the body of probability.

The observation that, while the topic of measure theory is so intimidating, its concepts at their heart are so simple, speak to our inability to teach this topic effectively. Measure theory can be understood by introductory students if explained using plain language and simple examples. This is what I have attempted in several of the tracts of this text.

This treatise will introduce measure theory without requiring a preceding course in real analysis. It is not my goal to supplant analysis but instead to only include sections on analysis in so far as they support the goal of learning measure theory.

In order to demonstrate some practical applications of measure theory, I have provided several examples of how measure theory can be used in public health disciplines e.g., environmental sciences, nephrology, ophthalmology, and clinical trial methodology.

I have devoted a section of this treatise to some of the technical details of measure theory (e.g., the monotone convergence and Lebesgue dominated convergence theorems). These are provided to demonstrate the veracity of the theory. However, one can use measure theory without having proved these theorems for themselves just as one can compute and manipulate the normal distribution without themselves having proved the central limit theorem.

Thus, instead of the usual measure theoretic didactic (theory centric with limited applications),  I have constructed and connected multiple sections into tracks, blending measure theory into the discussions of probability. These are provided in the tract development section.

In addition, there are other smaller but important tools that we can use to further interleave measure theory with the teaching of probability. One such elaboration is to use the term "measuring tool" interchangeably with "probability mass function" and "probability density function".

A second is to redesignate the integral sign, $\int\limits_A$ into a symbol that simply announces our intention to measure the set *A*. We can use any tool we chose in this measure e.g., counting "measure", Poisson "measure", or uniform "measure". Once the tool is selected, we simply have to know how to use it to take the measure, e.g., summation, standard Riemann integration, and more complicated set integration using simple functions.

With this approach, measure can be introduced with the simplest of probability distributions. Bringing measure along with us as we compute more complicated probabilities allows it to be a constitutive tool of application, and not just a theoretical affectation and burden for students.

## History and probability

Sections on the history of probability and brief biographies of some of the well-recognized contributors to our field have been included as well. They are referenced in the sections that discuss their contributions through hyperlinks to inform about the person behind the creation or advance.

Antepenultimately, all of this treatise' examples are in health care. It is my goal to update these as more become available. Such is the advantage of electronic publishing.

## An admonition

One warning that I consistently have given my students over the years remains as true today as it was two generations ago.

In order to compute a probability, one must first be sure that they understand the event.

Calculating annihilation probabilities for lymphocyte NK cells is impossible without some understanding of immunology.  Computing the likely rate of spread of an infectious disease is impossible without an understanding of Koch's postulates and basic epidemiology.

If you are going to have to kill some of your own brain cells solving a probability problem, then kill them in the process of understanding the nature of the event whose probability you are working to solve. The solution is so much easier once you clearly view the problem. Without that perspective, the exercise is a waste of your time.

# Track Development

Since every reader can assemble their own sequence of sections from those provided here, this work offers a different experience for each reader. In addition, since both elementary as well as advanced topics in probability are covered in detail, this work can be used by the student with no background in probability as well as by advanced graduate students.

The only price we pay for this is that this treatise should not be printed. This book is constructed to be studied and critiqued only as the digital version.

A review of the table of contents reveals the extent of topic coverage. There are a plethora of sections in this text. Historical, mathematical background, measure theoretic, basic probability, advanced probability, asymptotics, and tail events. Consider these as building blocks. However the sequence of topics to be covered is wholly up to the reader. Examples follow.

## Non Calculus Introductory Track

For example, if one has no background in calculus or probability but wishes to obtain a probability overview, then they can take the following path.

Why Probability
From Whence Did This Come
Probability and the Renaissance
Elementary Set Theory
Definitions and Basic Rules

The Random Event
Sigma Notation
Factorials
Counting Events - Combinatorics
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
The Continuous Probability Function
Basics of Normal Measure

## More advanced tracks

With the exception of the non-calculus introductory track, measure theory suffuses these writings; however the approach taken here offers this theory at different levels, making measure theory available to all available levels of students.

This is accomplished by introducing measure theory early on in a simple, almost nonmathematical way, then gradually moving it into the basic language of probability. In this way, the reader's intuition grows naturally while moving forward from simpler to more complicated probability distributions.

## Measure introductory track

The following track best serves students with one or two semesters of calculus. It begins with background.

Why Probability
From Whence Did This Come
Probability and the Renaissance
The Random Event
Elementary Set Theory

It then includes the basics of measure theory

An Introduction to the Concept of Measure
Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Measurable Functions
Simple Functions in Public Health
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution

With this now as background, the reader can move on to more advanced treatment of standard probability topics

Basic Properties of Probability
Counting Events
Bayes Theorem
Bernoulli Distribution – In Depth Discussion
Moment and Probability Generating Functions
Advanced Binomial Distribution
Hypergeometric Measure
Geometric an Negative binomial measures
General Poisson Process
Immigration-Emigration Modeling
Emigration-Death Process
Immigration-Death Process
The Continuous Probability Function
Uniform and Beta Measure
Survival Measure: Exponential, Gamma, and Related Measures
Cauchy, Laplace, and Double Exponential
Compounding
Ordering Random Variables
Normal Measure
Asymptotics

Tail Event Measure

## Physician-epidemiology track

This track replaces some of the background history (which is always available to the reader) with the physician centric components plus advanced Poisson process development.

The Random Event
Elementary Set Theory
Basic Properties of Probability
Counting Events
Bayes Theorem
Physicians and Conditional Probability
Assessing Diagnostic Tests
Bernoulli Distribution – In Depth Discussion
Moment and Probability Generating Functions
Advanced Binomial Distribution
Probability, Incidence, and Prevalence
Hypergeometric Measure
Geometric an Negative binomial measures
General Poisson Process
Immigration-Emigration Modeling
Contagion
Death Process
The Emigration-Death Process
Emigration-Death Process
Immigration-Death Process
The Continuous Probability Function
Survival Measure: Exponential, Gamma, and Related Measures

## Measure-centric probability

Alternatively, a sequence to introduce the student to a measure theoretic treatment of different probability distributions would be

Why Probability
From Whence Did This Come
Probability and the Renaissance
The Random Event
Elementary Set Theory
An Introduction to the Concept of Measure
Functions in Measure Theory
Sequences of Sets
Sequences of Functions
Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution

These students then should review the Properties of Real Numbers before proceeding.

- Basic Properties of the Lebesgue-Stieltjes Integral
- Monotone Convergence Theorem
- Some Classic Measure Theory Results
- Asymptotics
- Tail Event Measure

Then measure based probability treatment of classic distributions

- Basic Properties of Probability
- Counting Events
- Bayes Theorem
- Bernoulli Distribution – In Depth Discussion
- Moment and Probability Generating Functions
- Advanced Binomial Distribution
- Geometric an Negative binomial measures
- General Poisson Process
- Survival Measure: Exponential, Gamma, and Related
- Ordering Random Variables
- Normal Measure
- Tail Event Measure
- Asymptotics

## Mathematics review sections

There are many review sections that are available, listed in both the table of contents and here.

- Sigma Notation
- Factorials Permutations and Combinations
- Binomial Theorem
- Vandermond's Inequality
- Pascal's Triangle
- Properties of Real Numbers
- The Concept of the Limit
- Convergent Series
- Cauchy Sequences
- Pointwise vs. Uniform Convergence
- Convergence and Limit Interchanges
- Passing Limits Through Functions
- Uniform Convergence and Continuity
- Uniform Convergence, Integrals and Derivatives
- Curve Slopes
- Exponential Functions
- Differential Equations
- The Mean Value Theorem
- Polar Coordinates
- Exponential Limit
- Integration of Exponential Families
- Integration by Parts
- Gamma Function
- Fubini's Theorem

These review components are listed as prerequisites for the relevant sections of this book. Thus before reading a section, e.g., the contagion process, the reader will have a list of hyperlinked sequences to review so that they can retain their orientation in the original sequence. Also, throughout the sequences, relevant probability links are provided, e.g., a review of how to compute the probability of the union of events.

In addition, embedded in the discussions are useful intermediate results that are germane to the issue at hand, e.g., the skewness of the binomial distribution, or the asymptotic relationship between the Poisson and the binomial distribution.

Finally, if the reader has no background in probability but has to master several distributions, they may begin with the easier discussion and go to the more difficult discussion. Such a track would look like.

Why Probability
From Whence Did This Come
Probability and the Renaissance
The Random Event
Elementary Set Theory
An Introduction to the Concept of Measure
Basic Properties of Probability
Conditional Probability
Basics of Bernoulli Trials.
Bernoulli Distribution – In Depth Discussion
Basics of the Binomial Distribution
Advanced Binomial Distribution
Hypergeometric Measure
Basics of the Poisson Distribution
General Poisson Process
Basics of Normal Measure
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

If the reader has ideas for other sequences, please let me know as probability@principalevidence.com, and I will include them in the next eBook version.

# Why Probability?

Prerequisite: None

## Do we really need probability?

Critical questions and the search for their solutions propel us through life.

From its distant beginnings tens of thousands of years ago up to the present, our species asks and answers questions interminably, its cadence generated by the developing sophistication of our minds.

They encompass the questions posed by a member of a primitive culture (“Where will our next meal come from? Will we be attacked? Will our children survive?” ), questions of agriculture (“Will pestilence destroy our crop this year?  Will there be enough rain? Will the fire in that other village come here?”), modern national security (“When will the terrorists next attack? Will COVID-19 affect my community? Are our children safe at school?”).

While societies  are buffeted by this sea of questions, there are no fewer personal questions. (“Will I graduate? Can I find a good job? Will I survive this car accident?”). The consuming capability of emotional questions (“Will he ever leave me? Does she really love me? Will my death be long and painful? Will my daughter survive?”) are no less demanding.  Our survival may or may not depend on the presence of questions, but questions are its constant companion, and the characters of our lives are shaped by your and my search for their answers.

From thousands of years ago to the present day, many believe the answers to questions are found wholly in the supernatural. In time, society learned to appreciate its own role in shaping its destiny.  “The answers lie not in the stars, Brutus, but in ourselves,” Shakespeare’s character Cassius reminded 16th century audiences in *Julius Caesar*.

Yet, whatever the source of the final truth, we are unwilling (and many times cannot afford) to wait until the answer is self-evident. We seek an early view to the future so that we can influence that future (e.g., companies which to abandon the development of a new drug that will not be very effective).

The goal is accurate predictions. Accurate predictions redirect our actions, thereby changing the future and perhaps ourselves. Whether it is a general trying to predict the movement of her enemy, or a gambler discerning the next card he’ll be dealt in blackjack, the goal is the same, to change the course of the future.

Ultimately, we desire to shape (and not be shaped by) events that could have been predicted.

## Probability and Determinism

Critical questions beg us to apply our best tools to their answers. While determinism can have clear religious overtones (e.g., Calvinists who believe that God predetermines each and every one of our actions, completely removing the roles of both randomness and our own free will in our lives), technology and the advancement of science have grown their own deterministic taproots.

Consider the number of patient arrivals to an emergency department. This process has been studied intently; a common probability model implemented for predictions is the Poisson distribution. This distribution is used to assess resource questions, e.g., how busy the emergency department will be (and do I need to increase staff?), or the impact of the opening of a rival emergency department across town. Helpful predictions have been made in these settings by assuming what the Poisson model requires about the random, independent arrivals of patients within a given time interval.

Yet no patient believes that their arrival is random. For each of us, our ED arrival was fixed by a collection of events and circumstances that we picked our way through. We don't see these as random, but as impacted by our own nonrandom decisions and actions.

The trap here is believing that a prediction problem must be deterministic or random.

However, this choice was never the agreement. Nature gives us problems to solve that are unlabeled. It is we who label them as residing in the deterministic domain, probability, or stochastic systems.* In fact, these problems have features of both. They are simply problems.

Sometimes the role of probability is an issue of education. Consider hurricanes. For hundreds of years, hurricanes were considered acts of God. When records began to be kept of them each year, it was quite natural to consider them as random events and to apply probability models to them. In fact, Poisson and negative binomial probability distributions (or measures) have been helpfully applied to predict the expected number of hurricanes for a region in a month or in a season.

However, we have also learned that physics plays a role in hurricane development and movement. We now understand that these complex systems are governed not so much by chance, but by air-fluid interface dynamics, changes in atmospheric waves, wind shear, air pressure gradients, and ocean temperature dynamics.

In the near future, hurricane birth and movement will be examined in even greater detail, allowing mathematical models to determine with more refined precision when a hurricane will occur, the location of its birth, what part of the ocean it will churn, and where it will die.

With each passing decade, there is less of probability and more of physical science in determining the life cycle of hurricanes.

## Probability in Health Care Research

We need probability because we operate in a universe in which the same experiment leads to different outcomes. Whether one is observing if mice die (or not) in three days, or whether a patient is hospitalized or not in six months, or whether an individual's LDL cholesterol level will produce a heart attack, each subject's outcome is uncertain.

Different mice, and different patients will have different outcomes from the same experiment. We have come to understand this as natural or biologic variability, but how do we manage the ensuing chaos? If one individual's mouse's outcomes from an experiment is uncertain, than how do we describe the uncertain outcomes of each of hundreds of the scampering rodents?

The answer lies in the knowledge that repeated experiments —if they have certain properties — produce not disarray but predictable regularity; they can be summarized with

* Stochastic means probabilitistic in time, e.g., "Will the Uber driver arrive in five minutes?"

precision. The chaotic subject-to-subject findings are not inchoate in the aggregate but have a regularity that is governed by the rules of probability.

However, accepting this regularity comes with a price. We give up the notion of knowing what will happen tomorrow in order to embrace the possibility of learning what events will occur and at what frequency in the universe of tomorrows. For example, we may be able to predict how many women in our town will have a stroke tomorrow, but can say little about the fate of the woman standing next to us.

Probability is based on the random experiment, i.e., a well-designed experiment with an uncertain outcome (e.g., observing which patients with New York Heart Association Class III patients are hospitalized over a six month period). We compute the relative frequency (which is the probability) of this occurrence. With this, and the number of NYHA Class III patients, we can compute for example (using simple rules of unions, intersections, and the property of independence), the likelihood that one of these patients, or at least one, or between three and five of these patients are hospitalized in six months.

Many complex events can be constructed from the simplest one, and we don't have to wait to observe the complicated event; we can compute the frequency of its relative occurrence directly. Probability provides the structure for this understanding of the long term behavior of complex systems that at their heart have random occurrences.

Thus, while we have different laws to govern how to compute probabilities of simple events (e.g., binomial laws, geometric models, Poisson laws, etc.), the underlying process is always the same. Identify a simple event whose probability can be easily identified. Then, construct a more useful event, matching the introduction of research complexity step by step with mathematical representations and manipulations of those events, until we have in the end both the complex clinical event and the mathematical formulation (however, complicated that might be) of that event.

And, once we have the probability of the complex event, we can see what value (e.g., an arrival rate, or a mean) it requires and on which it depends. We then estimate that parameter from the data (this is estimation theory that produces for example means, or incidence rates, or hazard ratios) and then learn if the research effort changes that parameter, and thereby changes the probability distribution of the event.[*]

## Are there other models?

If we are to be honest, the answer must be "Yes", even though we do not know what those models are.

The three competing models with which we have experience are religious determinism, science determinism, and random models. Based on their cultural values, societies oscillate between these models in predicting important events.

However, we limit ourselves if we think that these are the important critical explanations out there. While we know much, most we do not know. In science, the vastness of the unknown dwarfs the known, and we must always be prepared for new perspectives that first upset, then bypass our current reasoning paradigms.

It is wise to temper our enthusiasm for any of these models by the observation that we do not know all that we need to circumscribe the universe of all possible answers.

Next sections to select
From Whence it Came – An Early History of Probability
Probability and the Renaissance

[*] This is the current role of statistical hypothesis testing, which will not be the subject of this book.

Mathematics Review

Measure

Probability Foundations

Basic Probability Distributions

Advanced Probability

# From Whence Did This Come? An Early History

Prerequisite: None

It is hard to imagine life without the concept of random uncertainty. Its presence is the only core requirement for the concept of probability to take root, because, at its very essence, probability is simply the use of mathematics to manage randomness. And, as we will see, when the concept of randomness is squeezed out by a culture, then probability–like fire without oxygen–dies.

## Ancient Use of Chance

No one knows where or when the notion of chance first arose.[1] It may have begun thousands of years ago with the use of the heel bones of sheep and other animals, known as astragali. These astragali are common products of ancient world archaeological digs, appearing far more frequently than one would expect based on solely a reasonable need of the people. Some believe that these astragali were the primary mechanism through which chosen people obtained the opinions of their gods.

In fact, for thousands of years, people threw dice to determine their fate. Whether they believed that this mechanism put them in contact with the gods, or removed man as the determining factor in the prediction, cultures relied on dice to learn of the future.

In Asia Minor, oracles cast or rolled five astragali; each possible result or configuration was associated with the name of a god, and carried with it the sought-after advice. For example, the outcome (1,3,3,4,4) signified Zeus and was encouraging, while the dreaded occurrence (4,4,4,6,6) evoked the frightening child-eating god Cronos.*

Over time, astragali were surpassed by dice as event generators, and the Greeks, in their full flower, embraced them.† They as well as the following Romans loved to gamble, and while rules for their games have long been lost, many can be traced forward to the Middle Ages. Yet not all were western games, as evidenced by the Crusaders arriving home with knowledge of eastern games, e.g., one which was very much like the modern day “craps”.

## Randomness without probability

* We will see later that there are $6^5$ such possible combinations or 7,776 combinations, ample room for varied advice from the gods.

† Loaded dice have also been found from antiquity. Thus, while developing an understanding of the true nature of random events and their description would take centuries, cheating was more easily mastered.

Yet while gambling blossomed, probability, or the development of an organized logical thought process based on the concept of randomness to predict outcomes, did not flower. There is no record of early ruminations of the behavior of random occurrences, not even the most coarse attempt to provide structure for the types of events that may be produced from a simple game in the western world.[*]

One might have expected the Greeks with their love of philosophy to inaugurate this field given their fondness for gambling. However, while Greek philosophers commonly discussed randomness and were comfortable with its role in their lives, they had no cultural interest in attempting to quantify it in any useful fashion. Even the opportunity to learn about probability in hopes of "beating the odds" and winning their bets was not a sufficient inducement to push them into developing a probability calculus. In fact, Plato wrote that arguments derived from chance are "imposters and…apt to be deceptive."[2]

However, with the Greeks' downfall, there arose a culture that not only shunned attempts to measure randomness, it would banish the entire concept.

## Descent

"*We say that those causes that are said to be by chance are not nonexistent but are hidden, and we attribute them to the will of the true God.*"

*St Augustine*

The Romans were on the march, pushing aside philosophical discussions of random events and ending the high time of Greek mathematics.[†] After his rise to power in Egypt, Ptolemy VII banished all scholars and scientists who were not loyal to him, forcing many of these Alexandrians to flee to more remote areas. Syracuse fell in 212 B.C., followed by Carthage in 202 B.C., Greece in 146 B.C, and Mesopotamia in 64 B.C.

Upon the assassination of Julius Caesar in 44 B.C., his grandnephew Augustus rose to rule the Western Roman Empire. After the defeat of Mark Anthony, Caesar Augustus now became ruler of the Eastern empire as well, moving on to conquer Egypt upon the suicides of Anthony and Cleopatra. The final fall of Egypt heralded a calmer era, leading to a brief resurgence of mathematics by Diophantus and Pappas.

However, the Roman empire was soon beset by its own set of problems.

Christianity began as a sect within Palestinian Judaism. Initially tolerated by the Roman state, this fervent religion spread throughout the Roman world. By the second and third centuries A.D. unruly mobs of Christians, Jews, and Egyptians clashed, producing widespread bloodshed.

Themselves singled out as a principal cause for internal unrest, Christians endured withering persecutions at the hands of the Romans. However, the emperor Constantine, after divining a sign during battle, converted to the controversial doctrine himself, and under Emperor Theodosius's rule, Christianity became the official religion of the entire empire.

However, good as this was for Christians, their ascendency heralded the deceleration of mathematical development as scholars were compelled to turn from mathematics and academic pursuits to issues of theology.

Faith was the main topic of study now, as physical science and mathematics were ridiculed, the Bible now being the source of all knowledge. In behavior that would recur through the centuries, both learning and the science of scholarship were debased. Associated with paganism, libraries and temples were looted and their holdings destroyed.

Meanwhile, the empire itself, beset with incessant internal civil wars and external threats, dividing in 330 AD into an eastern and western half, the latter portion to be overrun by peoples from the north.

[*] The Chinese were perhaps the earliest people to formalize odds and chance 3,000 years ago.

[†] Developed from http://www.saintjoe.edu/~karend/m441/DeclineAndRevival.html

## Stagnation and the depravity

August 24, 410 AD. Rome falls.

After a two year siege by invaders on a desperate search for food, a Roman city elder opened the strong gate, removing the last obstacle that blocked the raging Visigoth army that now poured in.

Defenseless, the hub of the Roman empire was sacked. All fell into chaos as property was demolished, cultural valuables plundered, and citizens, weakened by chronic hunger and sickened by measles and smallpox, were slaughtered.

After a seventy-two hour search for food, the invading hordes moved on. But they had accomplished in three days what the empire's enemies had never dared attempt. By destroying Rome, the center of civilization, "the mother of the world" had been killed.

Its destruction ushered in an unprecedented period of depravity.

For generations, tribe after tribe stormed into the empire to gorge on its dying corpus. Sewer systems failed, and buildings crumbled from the removal of supporting stones by invaders building their own shacks. Access to education and health care disappeared. Urban living was no longer tenable and surviving citizens fled, leaving entire segments of the cities to packs of wild animals. Civilization broke apart, and lasting knowledge was obliterated.

Deserting the cities, former citizens, many suffering from chronic starvation and disease, now struggled in the country to eke out an existence. Forced to provide for the first time their own food supply, they spent long hours behind makeshift plows on barely arable land, desperate to harvest food before a long and biting winter drove them inside for months of inactivity.

Typically, twenty-five percent of a family's children died in childbirth, and another twenty-five percent died by the age of twelve. Those that survived into their second decade were likely to do so without mother, father, or both.

And they were never safe.

Always under the threat of new attack with loss of life and crops, with no reliable source of information beyond the short horizon, anarchy and upheaval reigned as time and again small political problems quickly became military ones. There were only two choices. Join the rampaging gangs, or join the church. Many flocked to the monasteries, not to be devout but to simply survive.

## Death of randomness

The only potential havens for knowledge were the monasteries, serving as a residual of an academic and intellectual climate in a savage and impulsive world. The church provided for their education in the monasteries, developing and supporting literate cleric.

Elements of Latin reading and writing along with biblical study were the primary focus in the monasteries, and a hint of the ancient pagan Greek authors persisted through their writings. Old Latin manuscripts were preserved and copied, preserving them from loss.

Left to their devices, the monasteries might have been a haven for development of a thought process and calculus that governed random events. In fact, arithmetic was available during this period, and arithmetic texts thrived because of their ideological neutrality.

However, there were two insurmountable issues raised by randomness. The first was its source – gambling. This pastime's only use in Roman civilization had been to make money, and those who survived the empire's collapse remembered that the process tended to gather the more unsavory aspects of urban living. Monasteries were to be a bulwark against self-indulgence and the material world; their keepers refused to let the gambling "wolf" enter under the sheep's clothing of "the study of randomness".

The second far more fundamental matter focused on the nature of randomness itself. Christians believed that God controlled all things. Events, no matter how big or small were the product of God's direct intervention. The text of Matthew 10:30 that "And even the very hairs of your head are all numbered" attests to the omniscience and all controlling presence of God.

If God managed and controlled the myriad details and innumerable events taking place across the universe, each second of each minute of each hour of each day since the beginning of time, then there was certainly no room for randomness. There would be uncertainty, because no one knew God's will. But He knew His will and acted on it. All was determined by God. Nothing was left to chance.*

In fact, the suggestion that randomness did play a role in determining events suggested that God was not omnipotent, a line of reasoning that set the budding probabilist up as a heretic. In this inimical climate, the notion of the random event had few champions and therefore played little role in the monks' intellectual pursuits. Randomness was akin to disorder and the devil.

Meanwhile, the eastern (Byzantine) Empire, remained independent and isolated. Arab scholars set up the House of Wisdom which acquired and translated Greek manuscripts, placing them in a library for their use. It was here that the Greek philosophy and knowledge remained comatose but alive, awaiting the opportunity for revival.

It would have to wait one thousand years.

Next Section
From Whence it Came – An Early History of Probability
Probability and the Renaissance

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Properties of Real Numbers
An Introduction to the Concept of Measure

Mathematics Review
Sigma Notation
Factorials Permutations, and Combinations
Binomial Theorem
Vandermond's Inequality
Pascal's Triangle

Basic Probability Distributions
Counting Events

Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Counting Events
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution

* This argument resonates with many through the 20th* and 21st century.

Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

Also…

Blasé Pascal
Pierre Fermat
Abraham de Moivre
Famous Correspondence between Pascal and Fermat
Simon Laplace

Bayes and Price
The Inversion Problem

References

1. Larsen RJ, Marx, ML. A Brief History of Probability Bill Abrams, (Coeditor Second Moment) adaptation of An Introduction to Mathematical Statistics and its Applications by located at http://www.secondmoment.org/articles/probability.php

2 . Zorich, JN.“A Prehistory of Probability,” Statistics Group of the Santa Clara Valley Chapter of ASQ, November 8, 2000.

# Probability and the Renaissance

Prerequisite: From Whence This Came

## Yersinia pestis

It had been an achingly slow process, but by the mid fourteenth century, Florence, Italy had seen something of a resurgence in trade with the kingdoms to the east. So it was no surprise in 1348 that newly arrived ships from Constantinople, overflowing with cargo, were seen in its ports. Nor was it a surprise that sailors on those ships were ill from a variety of maladies including scurvy and malaria. The presence of rats, some infested with fleas also was nothing new.

But this infestation was different. These rats moved restlessly, and on quiet nights one could hear them rustling and squealing as they died.

Beginning in the east, the bacteria *Yersinia pestis* had followed the slow moving western trail of commerce, accompanying humans and their ever present companions, rats. While the rats generally shunned healthy people, the fleas that they carried easily jumped to them, and through their bites injected volumes of bacteria into the healthy human subcutaneous tissue and blood stream. Multiplying by the hundreds of millions in this new fertile ground, the bacteria produced their toxins.

And killed.

Europe–indeed, mankind–had seen nothing like this. The contagion spread rapidly through the growing urban populations, injecting what pestilence always spreads through human communities ─ death and fear.

Within a few months, half the population of Florence, one of the largest Italian cities, was dead.

The plague moved north and west, killing a quarter of Europe's population, driving people once more out of the urban communities that were struggling after hundreds of years of decay to recover population and an urban dynamic.

## Emergence

However, unlike the consequences of the collapse of the Roman Empire at the hands of the Visigoths, recovery from *Yesinia* was not so prolonged.

When the plague passed (as all plagues do), the survivors faced not just a new landscape but opportunity. They were thankful they had been able to endure. And for the first time in centuries, the dominating and omnipresent church was not nearby. There was room not just for new life, but for new thought.

While many thanked God for being spared a horrible death, they had to come to grips on their own with the fact that many of their loved ones and relatives, neighbors and workmates died. This left them with not just a sadness, but a heretofore unrecognized new essence of life; being alive had its own sweet value, and was more than just a mere stepping stone to the afterlife.

A second difference was money.

With people having died by the hundreds of thousands, the need for large quantities of fresh food declined, and good food became both more plentiful and cheaper to obtain. In addition, for the first time, the value of the working class increased. There were fewer workers now, and for the first time in their lives, this class found itself in high demand. They were free to relocate as they frequently did for better pay and treatment.

And they were on the move.

Also, though the population of Florence was cut in half by the plague, civilization did not collapse. Civil government continued. There began a new interest in study, with calls going forth from Italy to the east to return the works of the Greeks. These tomes were returned, sometimes with their tireless Arab curators accompanying them.

New discourse between peoples of different cultures began, and the study of Greek and Roman culture, once banned as heretical by the Church, was now offered in the rapidly developing universities. Banking families such as the Medici were generous with loans that expanded education and business.

And for the first time in generations, the church, supreme for almost a thousand years, came under attack for fiscal and moral corruption. Legal fiats to limit the power of the Pope were attempted, and, although the Church fought many of these off successfully, and would remain a central fixture in life for centuries more, it was no longer seen as the omnipotent personification and perfection of God on earth.

The church was powerless before new technical innovations such as the compass that overturned ideas of navigation. The printing press revolutionized not just how one could communicate, but the meaning of communication, as now common workers had access to information in abundance and never before available. Almost overnight, people demanded to learn to read.

Then, in the 1490's, the announcement of the discovery of Christopher Columbus changed the world.

Once more, the pundits had been wrong, about something so fundamental as the earth's very shape. "What else had they been wrong about?" people murmured.

The church did react, sometimes viciously, and many free thinking heretics were martyred. But something fundamental had changed in the relationship of men and women and the Catholic Church.

For a thousand years, the church served as the shield of the innocent from the terrors of post Roman barbarism. Now, new and free thinking individuals wondered if its power extended beyond its usefulness. Common people saw that perhaps they too could have a relationship with God, could in fact see something of Him, without looking through the eyes of the Catholic Church.

There was the spark of inquiry and free intelligence. The thinking world could now, once again, be open to the concept of the random event.

In 1494, Fra Luca Paccioli wrote the first printed work on probability, *Summa de arithmetica, geometria, proportioni e proportionalita*[*]

[*] David 1962

## The Best and the Worse

After its long and forced absence during the middle ages, cards were introduced in the fourteenth century and immediately gave rise to a game known as Primero, an early form of poker. Board games such as backgammon also became popular. Gambling, banned for a thousand years, was back and thriving, and everyone wanted to win.

But how?

The main orchestrator of this effort was one of the most mercurial figures in the history of probability, Gerolamo Cardano.

By his own admission, Cardano personified the best and worst of the Renaissance man. He was born in Pavia in 1501, but facts about his personal life are difficult to verify. He wrote an autobiography, but his penchant for lying raises doubts regarding much of what he says.

Cardano, both a formally trained physician and a gambling addict[*] had rich experiences in both winning and losing. This background led him to postulated that there must be an underlying structure to the unpredictable outcomes of these games, and he began to look for an abstract perspective on the random event.

After much obsession, he settled on a definition that is now held as the classic definition of probability:

### *Cardano's definition of probability*

Cardano articulated that if the total number of possible outcomes, all equally likely, associated with some action is $n$, and if $m$ of those $n$ possible outcomes result in the occurrence of some given event, then the probability of that event is $m/n$.[†]

Put another way, suppose a fair die is rolled, and there are $n = 6$ possible outcomes. If the event of interest is the outcome "the result must be a face with less than 4 spots", then the events {1}, {2}, {3} enumerate the outcomes that describe this event. Since there are three of them and they are equally likely, then, the probability of the event is 3/6, or 1/2.

His was the first recorded instance of computing a theoretical, as opposed to an empirical, probability, and by first elucidating and then tapping into the most basic principle of probability, he propelled the field forward.

In 1550 Cardano, inspired by the Summa, wrote a book about games of chance *Liber de Ludo Aleae*[‡]. While calculations of probabilities became more noticeable during this time period, his advance by and large went unnoticed as the momentum was shifting from Italy to France.

But, the game was now afoot.

## Letters

Chevalier de Méré was a French noblemen who loved the games, and gambled frequently to increase his wealth. During one gambling spell, he bet that at least one 6 would appear during a total of four rolls (a problem that is now addressed through the use of the binomial distribution.[§] From past experience, he knew that he was more successful than not with this game of chance.

[*] One story says that, so consumed was he, that he sold all of his wife's possessions in order to get table stakes.

[†] This would come to be seen as the classic definition of relative frequency probability, distinguished from that of the modern day Bayesian.

[‡] *Liber de Ludo Aleae* means *A Book on Games of Chance*

[§] This is a problem from the binomial distribution. The probability of at least one six in four rolls of a fair die begins with our letting $K$ = number of 6's thrown. Then $\mathbf{P}[K \geq 1] = 1 - \mathbf{P}[K = 0]$. We get $\mathbf{P}[K = 0]$ as $\mathbf{P}[K=0] = \binom{4}{0}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^4 = (1)(1)(0.482) = 0.482.$ The solution is $\mathbf{P}[K \geq 1] = 1 - \mathbf{P}[K = 0] = 1 - 0.482 = 0.518$.

However, after tiring of this approach, he decided to change the game. He bet that he would get a total of 12 (a double 6), on twenty-four rolls of two dice.[*] Soon he realized that his old approach to the game resulted in more money. He asked several prominent mathematicians, including his friend Blasé Pascal why his new approach was not as profitable, and, in addition, what has become known as the problem of points.

> Two people, A and B, agree to play a series of fair games until one person has won six games. They each have wagered the same amount of money, the intention being that the winner will be awarded the entire pot. But suppose, for whatever reason, the series is prematurely terminated, at which point A has won five games and B three. How should the stakes be divided?[†]

Pascal was intrigued by de Mere's questions and shared his thoughts with Pierre Fermat, a Toulouse civil servant and one of the most brilliant mathematicians in Europe. Fermat graciously replied, and from the now famous Pascal-Fermat correspondence came not only the solution to the problem of points but the foundation for more general results.

More significantly, news of what Pascal and Fermat were working on spread quickly. Others were attracted, of whom the best known was the Dutch scientist and mathematician Christian Huygens. The delays and the indifference that plagued Cardano a century earlier were not going to happen again.

Best remembered for his work in optics and astronomy, Huygens, early in his career, was intrigued by the problem of points. In 1657 he published *De Ratiociniis in Aleae Ludo* (Calculations in Games of Chance), a very significant work, far more comprehensive than anything Pascal and Fermat had done. For almost 50 years it was the standard textbook in the theory of probability.

Almost all the mathematics of probability were still waiting to be discovered. But the foundation was there. The mathematics of probability was finally on firm footing.

Background
Why Probability
From Whence it Came – An Early History of Probability
Probability and the Renaissance

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Mathematics Review

[*] $\mathbf{P}\left[\,1\text{ set of }12's\right]=\binom{24}{1}\left(\frac{1}{12}\right)^{1}\left(\frac{11}{12}\right)^{23}=0.270$

[†] The correct answer is that A should receive seven-eights of the total amount wagered. (Hint: suppose the contest was resumed, what scenarios would lead to A being the first person to win six games?

Sigma Notation
Factorials Permutations, and Combinations
Binomial Theorem
Vandermond's Inequality
Pascal's Triangle

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Counting Events
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

Also…

Blaise Pascal
Pierre Fermat
Abraham de Moivre
Famous Correspondence between Pascal and Fermat
Simon Laplace

Bayes and Price
The Inversion Problem

# The Random Event

Prerequisite: None

*The unexpected touches each of our lives.*

## What makes an event random?

To some individuals, random events appear to have no definite purpose, aim or direction, the result of an incomprehensible combination of events. Following no logical order, their occurrence is not the result of a cohesive series of steps, but instead are part of an unintelligible pattern.

A tornado demolished a sequence of three houses on an urban block, leaves the fourth undamaged, and moves on to destroy the rest of the homes in the neighborhood. Cards appear randomly in a game of poker. A patient appears randomly at a physician's office. Radioactive particles strike a Geiger counter randomly. Fire spares some horses while incinerating others.

We are the observers watching what occurs and recording the result. We see the arrival of a mother and child to an emergency department, or we watch with fear and awe as a super hurricane moves, stalls, and moves again, nature controlling this event. At the end of the exercise or experiment, an outcome is observed whose precise occurrence was unknown and unpredictable, and we say that it was a random event.

But are they really random?

Sometimes use of the term random merely represents our ignorance of the underlying mechanism. For example, it one were given the sequence of digits 592653, some would say that this is a random pattern of digits, while others might recognize this as the sequence of digits in $\pi = 3.141\underline{592653}$….. The sequence of digits for pi is interminable, and does not repeat. Without a discernible pattern, the digit sequence is unpredictable, exhibiting some of the characteristics of a random sequence. However, they are generated by a discernible and reproducible mechanism, failing this test of randomness.[*]

There are other complications embedded in the concept of randomness. Consider the mother and child arriving to the emergency department (ED). To the observer in the ED who had no foreknowledge of what compelled the subjects to visit the ED, their arrival at 10:03AM, following no knowable pattern, appeared random. However, to this mother (who first noticed her child's illness, then consulted with family and an older, befriended neighbor, then made a conscious decision to go to the doctor) her activity was purposeful and determined.

These actions from the mother's perspective are not random at all, but the consequence of her conscious thought and deliberate action. The same could be said of the gentlemen arriving to

[*] Such a sequence is described with the sobriquet "pseudorandom".

the ED with the worst chest pain of his life, or the daughter bringing her father in because he suddenly lost control of the right side of his body. From the perspective of each of these subjects there was nothing uncertain about their arrivals at all.

So, is randomness, like beauty, in the eye of the beholder? Is randomness an inherent property, or simply a reflection of our inability to know all?

## Religion and randomness

Discordians, who believe that both order and disorder are illusions imposed on the universe by humans, have a strong belief in randomness and unpredictability. Alternatively, Hindu and Buddhist philosophies state that any event is the result of previous events (karma), and as such, there is no such thing as a random event or a first event.

The development of probability was delayed approximately one thousand  years in part because there was no longer room for the random event in modern culture, all events being determined either by man or God. Martin Luther, considered by many to be the founder of Protestantism, believed that, based on his interpretation of the Bible, not only were there no random events, but that there essentially was no free will either. If indeed purpose governs the universe, then randomness is impossible. [*]

However, not all Christians believe in the absence of free will. For example, C. S. Lewis, a 20th-century Christian philosopher wrote: "God willed the free will of men and angels in spite of His knowledge that it could lead in some cases to sin and thence to suffering: i.e., He thought freedom worth creating even at that price." He later went on to say that God "gave [humans] free will. He gave them free will because a world of mere automata could never love..."

## Random models in a purpose driven universe?

However, even if we grant for a moment that purpose governs the universe, that, as Einstein said “God does not throw dice with the world,” then random laws still have a role. Consider the emergency department example above. Each patient, each arrival, was determined purposeful, yet models based on random processes, e.g., the binomial model, the Poisson model, and the negative binomial model continue to function well in describing the overall system.

We learn about the overall process by studying these random models. We can understand measures of central tendency and dispersion (means and variances), and we can compute the likelihood of events. Even though each arrival is nonrandom, the characteristics of these arrivals in their ensemble resemble a random process; so much so that models based wholly on probability can describe them.

That is not to say the individual arrivals or results are random; in fact, none of them are. However, the entire system can adequately be characterized by a random process, even though the process itself would be seen as wholly deterministic if only our knowledge about it was infinite.

A fine example of this is managing the arrival of airplanes at an airport. Clearly no plane arrives “randomly”; each adheres to a predetermined flight plan that describes to the minute when the plane takes off, the route it will follow, and when and where it will land. Yet when there are many such nonrandom arrivals, the ensemble is as though the system was governed by random arrivals, and, using the Poisson process[1] we can learn about the behavior of the entire system.

In some sense randomization is the first structure that we can place on understanding a complicated system. For example, suppose we wanted to learn about the Bangladesh culture. We

[*] To some degree this sense is alive and well today in the ongoing debate about evolution, with those who advocate for God determined evolution or intelligent design contend with  evolution’s proponents who argue that evolution is based on random genetic variations, that are nonrandomly selected by environmental stresses.

know no one from Bangladesh, and cannot travel there. However, what we can do is hang a microphone over its largest city. Now this is certainly a woefully inefficient way to learn about the people of this country. We would not learn their dialects, neither would we learn of the cultural intricacies of its peoples.

However, we would learn some things. For example, we would discover that that there is more activity during daylight hours than during the late evening, and that some days have more activities then other days. This is useful information in a vacuum of ignorance. We would know nothing of the details of every conversation, but collecting information about the sum total of the spoken words teaches us something of value.

This is what the application of the random model does. We rely on the random model to prove some overall characteristics. The individual outcomes are not random, but the aggregate of events behave as though they were. The model provides illumination about some of the characteristics of the deterministic process, whose intricacies we do not and may never know.

So also for the roll of die. We compute probability rules that successfully govern and predict the outcome of a single roll of two die. Yet, in today's age, one can compute the exact outcome of any role. We need only the most modern, intricate, and fast computer, plus the weight, height, and balance of each die, the speed, direction, and torque of the roll, the ambient temperature and humidity, and the elasticity of the surface the die strike on the first and subsequent landings.

Given sufficient computing power and control of the immense equations one must understand to manage the physics of this, one could predict each outcome. It is deterministic. Yet, the resultant of all of these forces while not random, resembles a random process enough so that 17th century gamblers and mathematicians could deduce the patterns and probabilities of outcomes. It is a property of the resultant or ensemble of deterministic forces that permits probability based models to be accurate in predicting the system's behavior.

Another perspective is to simply say that in our universe, events can have different properties. We say that randomness or deterministic are two contradictory properties, but that may appear to be true simply from our limited perspective. Just as light has properties of particles and also those of a wave, so to can events have properties of randomness and determinism.

Why Probability

From Whence it Came – An Early History of Probability
Probability and the Renaissance

Mathematics Review
Sigma Notation
Factorials Permutations, and Combinations
Binomial Theorem
Vandermond's Inequality
Pascal's Triangle

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers

An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

Also…

Blaise Pascal
Pierre Fermat
Abraham de Moivre
Famous Correspondence between Pascal and Fermat
Simon Laplace

Bayes and Price
The Inversion Problem

[1] N. Bauerle, O. Engelhardt-Funke and M. Kolonko On the Waiting Time of Arriving Aircrafts and the Capacity of Airports with One or Two Runways March 30, 2006.

# Elementary Set Theory

## Preamble

The functions that we work with in algebra and calculus are functions that map one single number (e.g., $x$) to another number (i.e., $x^2$) as in the function $y = x^2$. The probability function is different since it does not map one number to another, but instead maps a set (which is just a collection of objects, or events) to a number. Measure (and its subfield, probability)[*] relies on the properties of sets and operations involving sets.

We are accustomed to the familiar operations used on numbers (addition, subtraction, multiplication, division); After defining sets, we need a collection of operations to manage them. In this section we develop some fundamental set definitions and then define a collection of elementary operations on sets.

We will end with the concept of a $\sigma$-algebra which we will see is a very special and rich collection of sets. With this understanding of sets, we then begin to discuss how to measure them.

Fortunately, the concept of sets is very easy; it is one of the simplest concepts in mathematics. Set operations are similarly straightforward. We will need to be careful in implementing then correctly, and sometimes the collection of operations applied to the sets can be a challenge to comprehend.

However, comprehensive is primarily a matter of time. It we take the time, we will understand.

## Prerequisite

None

## Definition of a set and its elements

A *set* is simply a collection of objects. These objects can be physical, or they can be numbers. The set is defined by its contents. For example, we create a set $A$ such that

$A$ = {penny, nickel, dime, quarter, half dollar, dollar piece, two dollar piece}.

[*] It was Andre Kolmogorov who demonstrated that the field of probability (with its 500 year old history) was not a field unto itself but instead was a subfield of the much larger area known as measure theory.

The set $A$ is defined by its contents. It is the set of US coin denominations.

Note the set is denoted by braces $\{\}$. Each distinct entry in the set is called an *element* in (or of) the set. We denote the element status by the symbol $\in$ which means "is a member of". Thus nickel $\in A$ is a true statement while a ruble $\in A$ is false. In this case we say a ruble is not an element of $A$, i.e., a ruble $\notin A$.

Another set with many more element would be the set $B$ which contains all of the whole (or natural) numbers i.e.,

$$B = \{0, 1, 2, 3, 4, 5, ....\}$$

Looking at set $B$, define the set $C = \{1, 3, 5, 7, 9, ...\}$ which is the set containing only the odd whole numbers. Since each element in $C$ is also in $B$ we say that $C$ is a subset of $B$. Another common way to note this is that the set $C \in B$, or that C is contained in $B$ i.e., $C \subset B$. However $B \not\subset C$ since $B$ contains members or elements that $C$ does not. We will also define the null set or the empty set as the set with no elements, written as $\varnothing$ or $\{\}$.

## Set operations

Numbers can be added or subtracted with ease. We need to develop the same construct for sets. The concepts that we will use are complements, unions, and intersections. These operations will allow us to combine sets with other sets in order to create new sets. We might think of these as set operations that map sets to other sets. Or we can think of these operations as set generators, in that using these operations, new sets are created from other, established sets.

### *Complements*

To begin to familiarize ourselves with these set operations, let's start with a set that we will call $\Omega$, where

$$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8 , 9, 10\}$$

Now, let's define a subset $A = \{1, 2, 3, 4\}$. We see at once that $A$ is a subset of $\Omega$.

We will now define the complement of $A$ as the elements of $\Omega$ that are not in $A$. Thus we can write $A^c = \{5, 6, 7, 8 , 9, 10\}$, an operation that is "like" $\Omega - A$, if the subtraction operation were legal in set theory (which it is not). Note also that $\left(A^c\right)^c = A$.

### *Unions*

We define the *union* of two sets as the set containing only elements that are in either one set, the other set, or both. Using $A = \{1, 2, 3, 4\}$, let's define the set $B$ as $B = \{3, 4, 9, 10\}$. Thus we can write

$$A \cup B = \{1, 2, 3, 4\} \cup \{3, 4, 9, 10\} = \{1, 2, 3, 4, 9, 10\}$$

We can also find

$$A \cup B^c = \{1, 2, 3, 4\} \cup \{1, 2, 5, 6, 7, 8\} = \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

Note that there is no double counting. Each element that is contained in both sets is counted once and only once. Thus $A \cup \Omega = \Omega$ since $A \subset \Omega$. Also, $A \cup A = A$. Note that $B \cup B^c = \Omega$. While $\Omega \cup B = \Omega \cup B^c = \Omega$.

### *Intersections*

Finally, the intersection of a collection of sets is simply the elements that they have in common. Thus,

$$A \cap B = \{1, 2, 3, 4\} \cap \{3, 4, 9, 10\} = \{3, 4\}$$

The intersection of two sets that have no common elements is the null set, $\{\ \} = \varnothing$. These sets are called *disjoint sets* and have important properties that will be of great use to us later.

## Venn diagrams

We can see that set operations can become complicated. In order to help with visualization, Venn diagrams are particularly useful in visualizing the impact of these operations (Figures 1 and 2).

For example, from Figure 1 we can see that $A \cup B$ contains the overlap between the sets. Also we can see that the non-overlapping sets, $A \cap B$ and $A \cap B^c$ are disjoint.

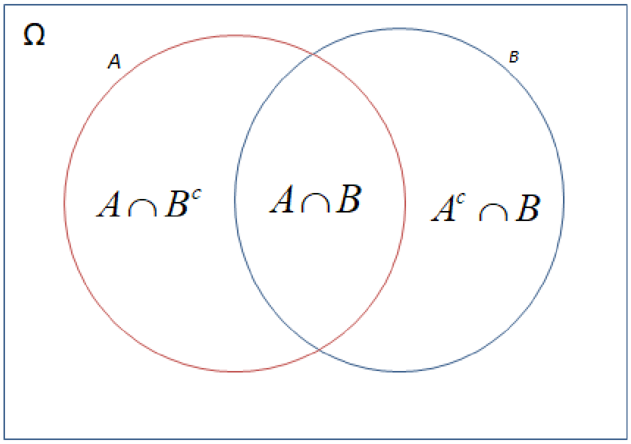


**Figure 1.** Demonstration of set relationships: unions, intersections and compliments involving two nondisjoint sets *A* and *B*.

The union operation can produce some interesting results. For example, when *B* is wholly contained in *A*, i.e., $B \subset A$, then $A \cup B = A$ and $A \cap B = B$. In the case where they are disjoint, then we can see (Figure 2) that $A \cup B = \{A, B\}; A \cap B = \varnothing$.

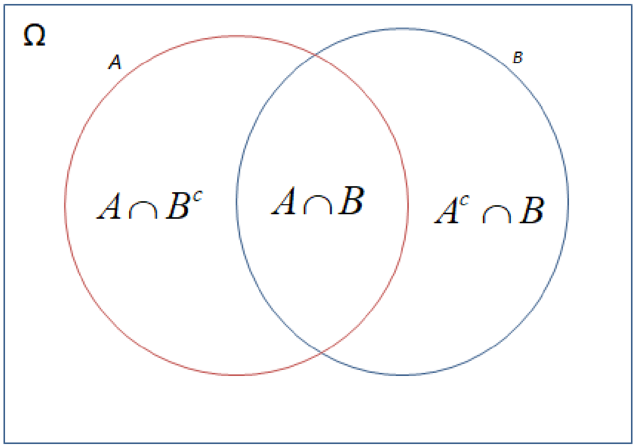


**Figure 1.** Demonstration of set relationships: unions, intersections and complments involving two nondisjoint sets *A* and *B*.

The distribution law shows us how to work with three sets.

**Distribution Law of Sets**

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

These are very useful rules. The first discusses the components in *A* or *B* that are common to set *C*. The second focuses on the components that are common to sets *A* and *B* and their relationship to a third set.

**DeMorgan's Law**

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

From DeMorgan's laws we learn that the complement of a union of sets is the intersection of those sets' complements, which is easy to see since a union's complement cannot contain any element in any of the sets.

Similarly, a complement's intersection cannot contained elements common to both.

Since, the sets *A*, *B*, and *C* can themselves be unions and intersections of other sets, the complement, union, and intersection operations can propagate entire new and larger collections of sets. Essentially, these operations put us in the set generation business.

## Example: Smart, "Live" Playlists

The common reaction to people exposed to set theory is that the topic is too abstract. They struggle with identifying a reason for using it in the real world.

Our smart devices provide a pertinent application that requires us to use set theory, perhaps without our knowing it.

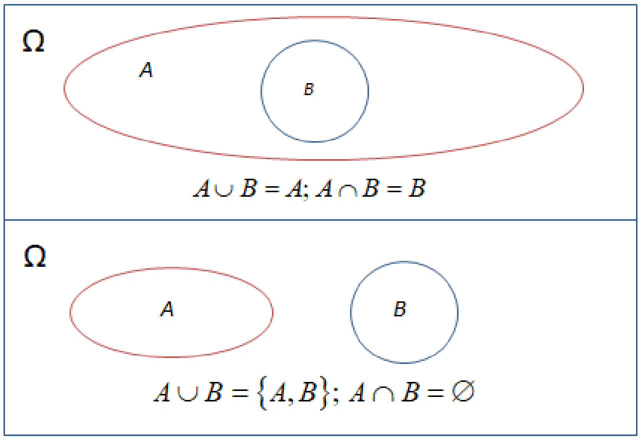

Ω
A
B
$A \cup B = A;\ A \cap B = B$
Ω
A
B
$A \cup B = \{A, B\};\ A \cap B = \varnothing$

**Figure 3.** The union set $A \cup B$ depends on the degree of overlap of sets $A$ and $B$, depicted by $A \cap B$.

Playlists, or a list of songs of interest, is a common way that music is organized on portable electronic devices. These lists of songs are simply a collection of songs that are selected from a larger universe of songs. The user goes through the universe of songs, and then manually selects tracks from the universe. This is simply creating a subset of tracks from the entire large collections of songs. If $\Omega$ is the entire music collection, and $P$ is the collection of tracks in the playlist, then $P \subset \Omega$.

More recently, some music players provide a semiautomatic way of creating playlists. Rather than manually selecting the subset of songs, one tags each song that resides in the universe. Examples of these tags are "Artist", "Title", "Genre", "Rating" (typically 1 to 5) and "Year", among others.*

In order to create the playlist, the user does not select the songs directly, but instead creates a "rule" based on the tags. For example, if the user wished to hear rhythm and blues music from the 1960's and 1970's with at least a three star rating they would select the rule

Year >1959
AND
Year <1980
AND
Genre contains "R&B"
AND
Rating > 2

The resultant playlist contains the required selection.

Examining the rule, we see that it is an example of set operations which in this case is a collection of intersections. The "And" operates as an $\cap$ from our set language.

Rules can be quite sophisticated, e.g., "Artist does not contain 'MaxJaw Willie'" and "Genre contains 'New Age'" and "Year >1980"

With the features of set theory, the owner can create very selective and focused playlists without having to go through and specifically select each track that is desired.

Other examples of set theory include database management and structured query language (SQL) which permit the creation of complex selection rules in patient databases.

* This information is commonly known as metadata.

## Set generation and σ-algebras

The operations of complement, union, and intersection permit us to combine sets in various ways. Through these combinations, we are actually generating new sets that are related to but different from the original sets, thereby spawning a relatively large number of sets. The number of sets that we can generate depends on Ω. For example, $\Omega = \{A, B\}$, then we can generate sets as follows.

$$\{A, B\}, A, B, A^c, B^c, A \cap B, A^c \cap B, A \cap B^c,$$
$$A^c \cap B^c, A \cup B, A^c \cup B, A \cup B^c, A^c \cup B^c, \varnothing$$

If $\Omega = \{A, B, C\}$, then we could generate many more sets through our familiar set operations.

This set generation feature is central to our use of sets in measure theory in general, and probability in particular. For us the result of this set generation is a particular family of subsets of a set on which we will rely and ultimately measure.

We will call the collection of sets that can be generated through this particular use of elementary set operations a *sigma algebra* or *σ-algebra*. A *σ*-algebra is nothing more than a collection of subsets of the set Ω (we will designate that collection of subsets as $\Sigma$ ) that follows certain rules of inclusion.

The precise definition of σ-algebra $\Sigma$ of subsets Ω is the following collection of subsets;

**a)** The null set is a member of $\Sigma$, $\varnothing \in \Sigma$
b) If the set $A \in \Sigma$, then $A^c \in \Sigma$.
**c)** If a countable number of sets $A_1$, $A_2$, $A_3$, … $A_n$,… are contained in Σ, then $\bigcup_{i=1}^{\infty} A_i \in \Sigma$.

A *σ*-algebra is a collection of sets, generated from the subsets of a set Ω. So, to create a σ-algebra, we start with a collection of sets, then generate from that collection the null set, and every possible combinations of unions, and complements.

However, this definition implies that intersections of sets are members of $\Sigma$ as well. Assume *A* and *B* are contained in $\Sigma$. Then $A^c$ and $B^c$, must be contained in $\Sigma$.. But their union $A^c \cup B^c$ must also be in $\Sigma$, as must their complement $\left(A^c \cup B^c\right)^c$, which by DeMorgan's law is $A \cap B$. So, defining a σ-algebra in terms of unions and complements also implies that this *σ*-algebra must contain their intersections as well.

***Example***: Consider a collection of the viral assay test results of five patients. Each have their unique test results $T_1$, $T_2$, $T_3$, $T_4$, $T_5$. The original set of them is simply $\{T_1, T_2, T_3, T_4, T_5\}$. We can construct the σ-algebra $\Sigma$ as

$$\varnothing, \{T_1, T_2, T_3, T_4, T_5\}, \{T_1\}, \{T_2\}, \{T_3\}, \{T_4\}, \{T_5\}, \{T_1^c\}, \{T_2^c\},$$
$$\{T_3^c\}, \{T_4^c\}, \{T_5^c\}, \{T_1 \cup T_2\}, \{T_1 \cup T_3\}, \{T_1 \cup T_4\}, \{T_1 \cup T_5\},$$
$$\{T_2 \cup T_3\}, \{T_2 \cup T_4\}, \{T_2 \cup T_5\}, \{T_1 \cup T_2 \cup T_3\}, \ldots$$

and on and on, continuing to build this collection of sets up through the unions, intersections and complements. From a set with five elements, by containing all unions, intersections, and complements of set elements, the resulting collection of subsets can be very rich. It all depends on the elements in the original set.

One useful way to consider the role of $\sigma$-algebras would be in painting. Suppose one had a gallon of red paint. Then the combinations of colors generated from it is very small; essentially no color (corresponding to the null set) or the color red.

Thus, the "$\sigma$-algebra" consists of only two elements. However, suppose you now add black, blue, and yellow gallons of paint.

The $\sigma$-algebra of all four colors is still all of the combinations of colors that can be generated by combining them, but because the original set is larger, the collection of subsets is very rich. The oranges, crimsons, purples, grays, teals, pinks, apricots, lavenders, boysenberries, etc. are all members of a huge mixture of new colors produced by combinations of the original set. Since the original set was richer, the $\sigma$-algebra has exploded.

This $\sigma$-algebra construct will be useful for us, because it will be the set of events from which we generate measure and probability. The greater and more diverse the original set of outcomes, the richer the algebra of events is on which we can construct probabilities.

The elementary track proceeds to
Sigma Notation
Factorials
The Probability of Unions of Events
Counting Events - Combinatorics
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
The Continuous Probability Function
Basics of Normal Measure

The more advanced track moves on to
Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties

# Sequences of Sets

Prerequisites
Elementary Set Theory
Properties of Real Numbers

For us to understand and work with measure theory and tail events, it would be best to appreciate the properties of not just sets, but sequences of sets. In general, this comprehension is only challenging if you do not have intuition about the nature of this new environment. However, obtaining this important experience is straightforward. We will begin simply, building from the ground up, letting the intuition develop as we work example after example.

A sequence of sets is an infinite collection of sets that is indexed by the integers, e.g., $A_1, A_2, A_3, \ldots A_n \ldots$ These sets can have not just different numbers of members (for example, set $A_1$ could contain seven members, and set $A_2$ can obtain 23 members) but also dramatically different members themselves. Our goal is to develop a collection of tools and concepts that can help us characterize their behavior.

## Convergence of a Set

What does convergence of a set mean? The convergence property of sets is related to the commonality of their elements throughout the infinite sequence. We have talked about sequences of real numbers converging using the "epsilon argument" of Cauchy as described previously, i.e., the sequence of real numbers $x_n$ converges to $\boldsymbol{x}$ if all but finitely many $x_n$ get as close as we would like to $\boldsymbol{x}$; we simply have to go far enough out in the sequence for this to be the case.

This approach suggests that we might consider a working definition of convergence of sets as all but finitely many sets contain the same elements; however we will see that this is only a portion of the complete solution.

Let's begin with a simple example. Consider the sequence of sets $A_n$

$$\{2\}, \{2,2\}, \{2,2,2\}, \{2,2,2,2\} \ldots$$

Does this continuing sequence of sets converge? Certainly, no two sets in the sequence are the same, but this poses no difficulty since the convergence of sets is about the convergence of their elements, denoted as $\omega$, irrespective of the number of elements. The element $\omega$=2 is common to all of these sets, so in a rather informal sense we might begin with saying that $\{2\}$ is the limit of $A_n$.

Now consider the sequence of sets $B_n$

$$\{2\},\{0,2\},\{0,2,0\},\{0,2,0,2\}...$$

In this sequence, there are two elements. Does it make sense for us to think about the sequence of sets $B_n$ converging when there is more than one common element, and that these two elements are all not common in all sets (for example, there is no element 0 in $B_1$)?

We need a way to consider the commonality of set elements that becomes apparent only after inspecting set elements sufficiently far enough out in the sequence. To help with this commonality concept, we can begin with

$$\bigcap_{n=1}^{\infty} X_n$$

where $X_n$ is a sequence of sets. Let's call this the "superintersection". It is a set whose members $\omega$ are in each and every one of the infinite number of sets. In our example, since $\bigcap_{n=1}^{\infty} A_n$ is the set of elements common to all $A_n$, we can write $\bigcap_{n=1}^{\infty} A_n = \{2\}$ since all sets have the element 2. Similarly, $\bigcap_{n=1}^{\infty} B_n = \{2\}$. as well (note that $\bigcap_{n=1}^{\infty} B_n \neq \{0,2\}$ since the first set $B_1$ did not contain the element 0).

We could also develop the superset

$$\bigcup_{n=1}^{\infty} X_n$$

which would provide for us the set that includes members of any of the sequence of sets $X_n$. This we can call the "superunion". We would therefore write $\bigcup_{n=1}^{\infty} A_n = 2$ and $\bigcup_{n=1}^{\infty} B_n = \{0, 2\}$. Thus, the superunion of the sequence of sets $A_n$ contains all elements of this infinite sequence of sets and the superunion of $B_n$ contains any elements that occur in the sequence $B_n$.

As another example, define the sequence of sets $C_n = \left\{\left(\frac{-1}{n}, \frac{1}{n}\right)\right\}$ i.e., the open interval from $-\frac{1}{n}$ to $\frac{1}{n}$. Thus the sequence begins $(-1, 1), \left(-\frac{1}{2}, \frac{1}{2}\right), \left(-\frac{1}{3}, \frac{1}{3}\right), \left(-\frac{1}{4}, \frac{1}{4}\right)$.... i.e., sets whose interval lengths decrease in size. In this case $\bigcup_{n=1}^{\infty} C_n = (-1,1)$, since there is at least one of the members of this set in the sequence $C_n$. Similarly $\bigcap_{n=1}^{\infty} C_n = \{0\}$ since only the element 0 is contained in each and every member of the sets.

One relationship we notice at once for any sequence $X_n$

$$\bigcap_{n=1}^{\infty} X_n \subseteq \bigcup_{n=1}^{\infty} X_n.$$

In general superintersections are smaller than superunions. This follows since elements in $\bigcap_{n=1}^{\infty} X_n$ must be in $\bigcup_{n=1}^{\infty} X_n$. However, elements that are in $\bigcup_{n=1}^{\infty} X_n$ need not be in $\bigcap_{n=1}^{\infty} X_n$ (such as the element 0 from set $B_1$). Thus, it follows that $\bigcap_{n=1}^{\infty} X_n = \bigcup_{n=1}^{\infty} X_n$ when all of the sets in the sequence have exactly the same elements.

Sometimes the superintersection is just the null set. Consider for example the sequence of sets $D_n = \{n\}$. Each set has one and only one element. In this case, the superunion of $D_n$ is all of the natural numbers while the superintersection is the null set since there are no common elements to the entire sequence. Thus $\bigcup_{n=1}^{\infty} D_n = \{1,2,3...n...\}$, while $\bigcap_{n=1}^{\infty} D_n = \varnothing$.

Finally, consider the sequence of sets $E_n = \left\{\left[2, 3-\frac{1}{n}\right)\right\}$. It is the sequence of intervals close on the left $\left[2, 2\frac{1}{2}\right), \left[2, 2\frac{2}{3}\right), \left[2, 2\frac{3}{4}\right), \left[2, 2\frac{4}{5}\right)$ each interval getting slightly longer. Here, the superunion $\bigcup_{n=1}^{\infty} E_n = [2, 3)$ while the superintersection $\bigcap_{n=1}^{\infty} E_n = \left[2, 2\frac{1}{2}\right)$.

## Liminfs and Limsups of sets

Now, let's take this one step further and combine the concept of unions of intersections, $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty}$. This is related to the "superunion of the superintersections", but notice that the index of the superintersections does not always begin with zero, but with $m > n$ as *n* increases. Can this concept help us?

Begin with our first set sequence $A_n$ and write $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n$. This superunion of superintersections can appear to be overwhelming but is easily considered when taken as a two-step process. The first is to create a sequence not of sets but of intersections of sets, and then take the union of these sets. Recall that $A_n$ equal to sets whose members are only twos i.e., $\{2\},\{2,2\},\{2,2,2\},\{2,2,2,2\}$... Then

$$m \geq 1: \ \bigcap_{m=1}^{\infty} A_n = 2$$

$$m \geq 2: \ \bigcap_{m=2}^{\infty} A_n = 2$$

$$m \geq 3: \ \bigcap_{m=3}^{\infty} A_n = 2$$

...

Thus these intersections are always 2. Taking the superunion of these intersections, we see that the intersections contain only one element, and we simply take the union of them to write $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n = \{2\}$.

Now, let's reverse the superunion and superintersection to write $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n$. Again, using $A_n$ and breaking this complicated procedure down into a two-step process, we compute

$$m \geq 1: \ \bigcup_{m=1}^{\infty} A_n = \{2\}$$
$$m \geq 2: \ \bigcup_{m=1}^{\infty} A_n = \{2\}$$
$$m \geq 3: \ \bigcup_{m=1}^{\infty} A_n = \{2\}$$
$$\ldots$$

The intersection of these sets or $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n = \{2\}$. In this case, the $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n = \{2\}$.

These two sequence of operations, $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n$ and $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n$ although the reverse of each other, each develops a sense of commonality of the members of the $A_n$ family of sets by looking in different directions. $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n$ starts by examining what all of the sets have in common (a relatively small collection of members most times) and builds that up by including all members of this sequence of intersections. The $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n$ looks at all members of all sets in the sequence and then "builds down" to find the common intersection. In the case of the set sequence $A_n$, the two approaches produce the same set.

Let's try this same process with the sequence of sets $B_n$ where we had $\{2\},\{0,2\},\{0,2,0\},\{0,2,0,2\}$... We compute $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} B_n$ and $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} B_n$ :

$$m \geq 1: \ \bigcap_{m=1}^{\infty} B_n = \{2\} \qquad \bigcup_{m=1}^{\infty} B_n = \{0,2\}$$
$$m \geq 2: \ \bigcap_{m=2}^{\infty} B_n = \{0,2\} \qquad \bigcup_{m=2}^{\infty} B_n = \{0,2\}$$
$$m \geq 3: \ \bigcap_{m=3}^{\infty} B_n = \{0,2\} \qquad \bigcup_{m=3}^{\infty} B_n = \{0,2\}$$
$$\ldots$$

The superunion of these superintersection is the set $\{0,2\}$ and we conclude that $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} B_n = \{0,2\}$.

The superintersection of the superunions is also $\{0,2\}$, and thus $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} B_n = \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} B_n = \{0,2\}$.

Notice that the fact that $B_1 = \{2\} \neq \{0,2\} = B_2$ did not disturb the equality.

For $C_n = \left\{\left(\frac{-1}{n},\frac{1}{n}\right)\right\}$, we write

$$m \ge 1: \quad \bigcap_{m=1}^{\infty} C_n = \{0\} \qquad \bigcup_{m=1}^{\infty} C_n = \left\{\left(-\frac{1}{2}, \frac{1}{2}\right)\right\}$$

$$m \ge 2: \quad \bigcap_{m=2}^{\infty} C_n = \{0\} \qquad \bigcup_{m=2}^{\infty} C_n = \left\{\left(-\frac{1}{3}, \frac{1}{3}\right)\right\}$$

$$m \ge 3: \quad \bigcap_{m=3}^{\infty} C_n = \{0\} \qquad \bigcup_{m=3}^{\infty} C_n = \left\{\left(-\frac{1}{4}, \frac{1}{4}\right)\right\}$$

...

The superunion of the superintersections is $\{0\}$ and the superintersections of these superunions is also $\{0\}$ so again we write $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} C_n = \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} C_n = \{0\}$.

Now if $D_n = \{n\}$, we can see that

$$m \ge 1: \quad \bigcap_{m=1}^{\infty} D_n = \varnothing \qquad \bigcup_{m=1}^{\infty} D_n = \{1, 2, 3, 4, 5, ...\}$$

$$m \ge 2: \quad \bigcap_{m=2}^{\infty} D_n = \varnothing \qquad \bigcup_{m=2}^{\infty} D_n = \{2, 3, 4, 5, ...\}$$

$$m \ge 3: \quad \bigcap_{m=3}^{\infty} D_n = \varnothing \qquad \bigcup_{m=3}^{\infty} D_n = \{3, 4, 5, ...\}$$

...

We see that $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} D_n = \varnothing$. However, what is the superintersection of these superunions of the set sequence $D_n$? There is no finite member of the set that is in all of the unions. Thus, $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} D_n = \infty$. Here, $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} D_n \ne \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} D_n$.

We should pause here to note that $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} D_n$ only picked up elements of the $D_n$ sequence that occurred infinitely often. Since no single positive integer met this criteria, its value became $\infty$.

Typically, the $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} X_n$ is called the limit infimum of liminf. It starts with a relatively small number of sets (the superintersection) and then builds them up. The $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} X_n$ is called the limit supremum or limsup (starting with a relatively large set, the superunion, and building down). As we saw with the $D_n$ sequence example, the limsup $X_n$ is the subset of $X_n$ that occur infinitely often.

We will say the limit of a sequence of sets exists when $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} X_n = \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} X_n$. We can write this more formally by writing that in this case

$$\lim_{n\to\infty}\{X_n\} = \lim_{n\to\infty}\inf_{m\ge n}\{X_n\} = \lim_{n\to\infty}\sup_{m>n}\{X_n\}.$$

The following reviews our findings for these four sets

$$A_n = \{2\},\{2,2\},\{2,2,2\},\{2,2,2,2\}... \qquad \lim_{n\to\infty}\{A_n\} = \{2\}$$

$$B_n = \{2\},\{0,2\},\{0,2,0\},\{0,2,0,2\}... \qquad \lim_{n\to\infty}\{B_n\} = \{2\}$$

$$C_n = \left\{\left(\frac{-1}{n},\frac{1}{n}\right)\right\} \qquad \lim_{n\to\infty}\{C_n\} = \{0\}$$

$$D_n = \{n\} \qquad \lim_{n\to\infty}\{D_n\} \text{ none}$$

This definition of set limit focuses on the commonality of the set members. The set sequence $D_n$ has no commonality at all, and therefore has no limit (we can say that the sequence of these sets diverges). However there is commonality (in fact, we could say, overwhelming commonality) of members in the sets $A_n, B_n, C_n$

What does "overwhelming" mean? Consider the sequence of sets $E_n = \{1_{n\text{ even}}\}$, where the single member of $E_n$ alternates between 0 and 1. Then, carrying our analysis we find

$$m \geq 1: \quad \bigcap_{m=1}^{\infty} E_n = \varnothing \qquad \bigcup_{m=1}^{\infty} E_n = \{0,1\}$$

$$m \geq 2: \quad \bigcap_{m=2}^{\infty} E_n = \varnothing \qquad \bigcup_{m=2}^{\infty} E_n = \{0,1\}$$

$$m \geq 3: \quad \bigcap_{m=3}^{\infty} E_n = \varnothing \qquad \bigcup_{m=3}^{\infty} E_n = \{0,1\}$$

...

Note here that $\liminf_{n\to\infty\; m\geq n}\{E_n\} \neq \limsup_{n\to\infty\; m>n}\{E_n\}$. However the set sequence $E_n$ contains both an infinite number of 0's and an infinite number of 1's. Thus, according to our definition, the occurrence of members infinitely often is not sufficient for the limit to exist.

Well, if an infinite occurrence of events is not what it takes, then what does matter? This is revealed by an examination of the sequence $F_n = \{1_{n>100}\}$. This set sequence consist of the set $\{0\}$ for $n \leq 100$, and the set $\{1\}$ for $n > 100$. We write

$$m \geq 1: \quad \bigcap_{m=1}^{\infty} F_n = \varnothing \qquad \bigcup_{m=1}^{\infty} F_n = \{0,1\}$$

$$m \geq 2: \quad \bigcap_{m=2}^{\infty} F_n = \varnothing \qquad \bigcup_{m=2}^{\infty} F_n = \{0,1\}$$

$$m \geq 101: \quad \bigcap_{m=101}^{\infty} F_n = 1 \qquad \bigcup_{m=101}^{\infty} F_n = \{1\}$$

$$m \geq 102: \quad \bigcap_{m=102}^{\infty} F_n = 1 \qquad \bigcup_{m=102}^{\infty} F_n = \{1\}$$

...

In this case both $\liminf_{n\to\infty\; m>n} F_n = \limsup_{n\to\infty\; m>n} F_n = \{1\} = \lim_{n\to\infty} F_n$. This was the result because the because the superintersection in the limsup remove those ω elements that appeared in the first 100 sets of $F_n$.

Some reflection reveals that it doesn't matter whether $F_n = \left\{1_{n>100}\right\}$, of $F_n = \left\{1_{n>1000}\right\}$, or $F_n = \left\{1_{n>10,000,000}\right\}$.

As long as there were only finitely many of the 0's, and all but finitely many of the 1's, the limsup would be one. Similarly for the liminf. The null set result for the intersection drops away after the set changes from 0's to 1's, and the final superunion produces only the set $\{1\}$. Just as a property of the limsup $X_n$ is it must occur infinitely often, a property of the liminf $X_n$ is that it occurs all but finitely many times. Thus, the $\lim\limits_{n\to\infty} X_n$ exists when the subset of the elements of $X_n$ sequence that occurs infinitely often is identical to the set of elements that occur all but finitely many times.

Consider the sequence of sets $G_n = \left\{\left(\frac{1}{n}, 1\right)1_{n\text{odd}} + \left(-1, -\frac{1}{n}\right)1_{n\text{even}}\right\}$. This sequence of sets is subject to two forces. The first is that the length of the interval increases as *n* increases. The second is that  the interval is either on the negative side of the reals or the positive ones, alternating as *n* is either odd or even. The analysis of the limiting behavior of this sequence follows:

$$m \ge 1: \quad \bigcap_{m=1}^{\infty} G_n = \varnothing \qquad \bigcup_{m=1}^{\infty} G_n = \{(-1, 0) \cup (0, 1)\}$$

$$m \ge 2: \quad \bigcap_{m=2}^{\infty} G_n = \varnothing \qquad \bigcup_{m=2}^{\infty} G_n = \{(-1, 0) \cup (0, 1)\}$$

$$m \ge 3: \quad \bigcap_{m=3}^{\infty} G_n = \varnothing \qquad \bigcup_{m=3}^{\infty} G_n = \{(-1, 0) \cup (0, 1)\}$$

$$\ldots$$

Continuing, we see that the liminf $= \bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} G_n = \varnothing$ while the limsup $= \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} G_n = \{(-1, 0) \cup (0, 1)\}$. Thus the limit does not exist. However, it is easy to convince ourselves that $\lim\limits_{n\to\infty}\left\{\left(\frac{1}{n}, 1\right)\right\} = \{(0,1)\}$, and $\lim\limits_{n\to\infty}\left\{\left(-1, -\frac{1}{n}\right)\right\} = \{(-1,0)\}$.

We can now say in words what it means for a sequence of sets to converge. By saying that the sequence of sets $\{A_n\}$ converges we have demonstrated that $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n = \lim\limits_{n\to\infty} A_n = A.$ In words, the limit exists if each member of A must be all but finitely many $A_n$, and that each member of set $A^c$ must be in only finitely many of the $A_n$.

## Example: Medical Text

Consider for example the text in all medical discharge diagnoses from 1950 to 2015. This constitutes a very large collection of medical records, and lets believe for a moment that this is an infinite collection of medical records. Let each record be a set of words.

Do these medical records have a limit, and what would that mean?  A first impression suggest that they do have a limit. For example, the articles "a" "an", and "the" are in all of them (or at least all but finitely many of them). How about the complement of these three "words." Again, an impression says that other words that describe particular diagnoses e.g., "esophagitis", "osteomyelitis", " myocardial infarction" may be in a great many of these medical records, but only in finitely many of them. So a set limit here represents common language.

[redacted]

Another example of set limits is the sequence of sets that is forever increasing, i.e., $H_1 \subset H_2 \subset H_3 \subset ...$ Such a set is easy to imagine (Figure 1.)

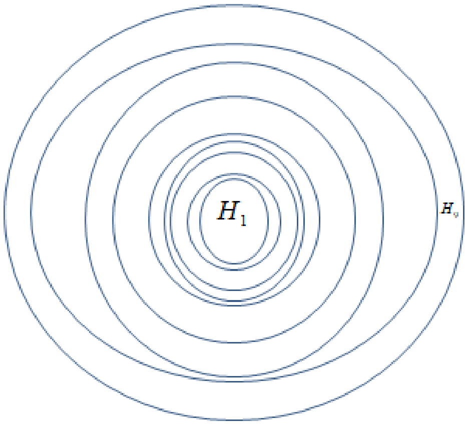


**Figure 1.** Depiction of a sequence of increasing sets.

Our computations follow:

$$
\begin{array}{lll}
m \geq 1: & \bigcap_{m=1}^{\infty} H_m = H_1 & \bigcup_{m=1}^{\infty} H_m = H_\infty \\
m \geq 2: & \bigcap_{m=2}^{\infty} H_m = H_2 & \bigcup_{m=2}^{\infty} H_m = H_\infty \\
m \geq 3: & \bigcap_{m=3}^{\infty} H_m = H_3 & \bigcup_{m=3}^{\infty} H_m = H_\infty \\
\ldots & &
\end{array}
$$

$$\bigcup_{n=1}^{\infty} \bigcap_{m>n}^{\infty} H_n = \bigcap_{n=1}^{\infty} \bigcup_{m>n}^{\infty} H_n = H_\infty = \lim_{n \to \infty} H_n$$

As another example, consider the sequence of sets defined as $J_n = \left[\frac{1}{n+1}, \frac{1}{n}\right)$. Does this set has a limit?

To answer this question, lets first look at this sequence to get a sense of its character. The sequence is $\left[\frac{1}{2},1\right),\left[\frac{1}{3},\frac{1}{2}\right),\left[\frac{1}{4},\frac{1}{3}\right),\left[\frac{1}{4},\frac{1}{5}\right)...$ The union of these sets is $[0,1)$. Since these sets are mutually exclusive, $\bigcap_{n=1}^{\infty}J_n=\varnothing$. Let's examine $\lim_{n\to\infty}\inf_{m>n}J_n$ and $\lim_{n\to\infty}\sup_{m>n}J_n$. Lets begin our analysis.

$$m\geq 1:\quad \bigcap_{m=1}^{\infty}J_m=\varnothing \qquad \bigcup_{m=1}^{\infty}J_m=[0,1)$$
$$m\geq 2:\quad \bigcap_{m=2}^{\infty}J_m=\varnothing \qquad \bigcup_{m=2}^{\infty}J_m=\left[0,\frac{1}{2}\right)$$
$$m\geq 3:\quad \bigcap_{m=3}^{\infty}H_m=\varnothing \qquad \bigcup_{m=3}^{\infty}J_m=\left[0,\frac{1}{3}\right)$$
...

Here $\lim_{n\to\infty}\inf_{m>n}J_n=\varnothing$, and $\lim_{n\to\infty}\sup_{m>n}J_n=[0]$. Thus the sequence $J_n$ has no limit.

Consider the sequence $K_n=\left[\frac{n}{n+1},1\right]$. What are the characteristics of this sequence? Does it have a limit?

The sequence is $\left[\frac{1}{2},1\right],\left[\frac{2}{3},1\right],\left[\frac{3}{4},1\right],\left[\frac{4}{5},1\right]$. The intervals gets smaller and smaller and always has an upper bound as 1. Does it make sense that $\lim_{n\to\infty}K_n=1$? Lets conduct the analysis

$$m\geq 1:\quad \bigcap_{m=1}^{\infty}K_m=[1] \qquad \bigcup_{m=1}^{\infty}K_m=\left[\frac{1}{2},1\right]$$
$$m\geq 2:\quad \bigcap_{m=2}^{\infty}K_m=[1] \qquad \bigcup_{m=2}^{\infty}K_m=\left[\frac{2}{3},1\right]$$
$$m\geq 3:\quad \bigcap_{m=3}^{\infty}K_m=[1] \qquad \bigcup_{m=3}^{\infty}K_m=\left[\frac{3}{4},1\right]$$
...

Here $\lim_{n\to\infty}\inf_{m>n}K_n=[1]$, and $\lim_{n\to\infty}\sup_{m>n}K_n=[1]$, and we discover that our intuition was correct.

As another example, consider the sequence of sets $L_n=\left[\frac{1}{n+1},\frac{n}{n+1}\right]$. Does this sequence of sets converge?

The sets begin as follows: $L_n=\left[\frac{1}{2},\frac{1}{2}\right],\left[\frac{1}{3},\frac{2}{3}\right],\left[\frac{1}{4},\frac{3}{4}\right],\left[\frac{1}{5},\frac{4}{5}\right]...$ The limit of the lower bound of the set is zero and the limit of the upper bound of the set is 1. What are the limit properties of $L_n$?

$$m\geq 1:\quad \bigcap_{m=1}^{\infty}L_m=\left[\frac{1}{2}\right] \qquad \bigcup_{m=1}^{\infty}L_m=[0,1]$$
$$m\geq 2:\quad \bigcap_{m=2}^{\infty}L_m=\left[\frac{1}{3},\frac{2}{3}\right] \qquad \bigcup_{m=2}^{\infty}L_m=[0,1]$$
$$m\geq 3:\quad \bigcap_{m=3}^{\infty}L_m=\left[\frac{1}{4},\frac{3}{4}\right] \qquad \bigcup_{m=3}^{\infty}L_m=[0,1]$$
...

Here $\lim_{n\to\infty}\inf_{m>n} L_n = [0,1]$, and $\lim_{n\to\infty}\sup_{m>n} L_n = [0,1]$ and we can write $\lim_{n\to\infty} L_n = [0,1]$.

Sequences of Functions
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Monotone Convergence Theorem
Some Classic Measure Theory Results
Asymptotics
Tail Event Measure

# Basic Properties of Probability

## Introduction

Here, we define probability as a set function that generates the relative frequency of the set, and provide our first simple examples of probability computations. Once we define probability, we will demonstrate the difference between the sample space and the much larger event space, or collection of sets on which the probability function will operate. We will also begin to see how operations on sets such as complements, unions and intersections translate to the mathematical combinations of probabilities.

## Prerequisite sections

The Random Event
Elementary Set Theory

## Probability as relative frequency

We define the probability of event $A$ as $\mathbf{P}[\ A\ ]$ as the relative frequency[*] of the event $A$

$$\mathbf{P}\left[A\right]=\frac{\text{number of outcomes in event } A}{\text{number of all outcomes}}$$

One simply computes the number of all possible outcomes, placing this number in the denominator. We then select the number of events[†] in the denominator that meet the criteria of the event whose probability we wish to find, placing this second number in the numerator, permitting us to compute the proportion. This serves nicely as a first working definition of probability. The concept of probability from its first modern inception has been one of assessing

[*] This is the source of the sobriquet "relativist" for practitioners of this classic use of probability.

[†] We could also say that we place the size of these events in the denominator and numerator and compute the proportion. Describing these collections by their size and not merely their number is the essence of measure theory's application to probability. In this case the measure of the events is simply the number of events. Measure most times is conceptually quite easy.

the relative frequency of an event[*] and is related to one of the most basic concepts in epidemiology – proportions.

## Example: Demography

An epidemiologist is interested in characterizing the demographics of  120 patients in her sample of subjects. She has and the breakdown of patients is available (Table 1).

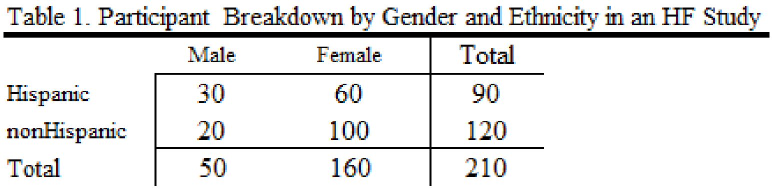


Table 1. Participant  Breakdown by Gender and Ethnicity in an HF Study

| | Male | Female | Total |
|---|---|---|---|
| Hispanic | 30 | 60 | 90 |
| nonHispanic | 20 | 100 | 120 |
| Total | 50 | 160 | 210 |

Table 1 tabulates the race, ethnicity and gender in the sample of 120 patients.  If we are to compute the probability of being Asian, then, using our definition of probability from the previous section , we simply count the total number of patients available in the study ($N = 120$). This is our denominator. The numerator contains the total number of Asians in the study = 15 + 10 = 25. The probability of Asian is simply $\mathbf{P}\left[\text{Asian}\right] = \frac{25}{120} = 0.208.$ Similarly, we can compute the probability of being a female in this sample as $\mathbf{P}\left[\text{Female}\right] = \frac{45}{120} = 0.375.$

■

Our intuition serves us nicely here.  From these computations we can observe that probability must be between 0 and 1. Secondly, there are many probabilities that we can compute from the rich structure of Table 1, for example, the probability of being  an African American, or of being an Hispanic female. The probabilities of all of the subsets of demography in the table are available to us directly; we simply count events in the numerator, and divide by 120.

We can also handle a complication of this approach easily as well. Suppose we wanted to know the probability of an individual being Caucasian if we already knew they were female. Here the denominator is not 120 because there are fewer females than this. If we wanted the probability of being Caucasian, we would compute $\frac{69}{120}$ However, if we want the probability of being Caucasian among only the females, we write $\frac{19}{45}$.[†]

## Notational structure of probability

Further examination of  Table 1 reveals some simplifications that should be available. For example, if we compute the probability of being female as $\frac{45}{120} = 0.375$ can we just compute the probability of being male as $1 - 0.375 = 0.625$? Must we always return to Table 1 to count up events and then compute the probability?

By adding some structure to our probability computations we will see that we can simplify our counting burden considerably. In fact, while we may already have some sense of

[*] This is the traditional definition of probability. There is an alternative branch of probability, termed subjective probability that has been developed by the Bayesian community. This will be a topic of later discussion.

[†] This second probability is known as a conditional probability because it is “conditioned” on being a female.

how to compute probabilities of events from Table 1, let's explore this concept of a general structure first.

Applying our new experience in set theory, we can first consider a universe or sample space of all available outcomes. For Table 1, we have 120 subjects. Let's also consider this an experiment where we select an individual from this population of 120 subjects, note that individual's race, ethnicity and gender, and then return that individual to the sample. The overall set **Ω**, which we will now call the sample space **Ω** contains 120 elements, each describing an individual.

From Table 1, we are interested in characterizing individuals by gender, race and ethnicity. This will generate subsets of individuals characterized by these three groups, using the operations of complements, unions, and intersections to generate a *σ*-algebra of subsets. It is on this *σ*-algebra[*] that we may build our probability function.

A straightforward examination demonstrates how rich this event space or *σ*-algebra is. For example, based on Table 1, we can derive the following events.

All subjects
All African-Americans
All Hispanics
All Males
All Caucasian females
Etc.

All Caucasians
All Asians
All Females
All Non-African American Hispanics
All Asian males

Clearly this is a fraction of the many events that may be enumerated. However, notice that there are events (e.g., the relative frequency of all subjects less than fifty years old) that we cannot compute, since there is no information in Table 1 about subjects' ages[†]. However, by confining ourselves to this rich *σ*-algebra we can "build" our probability function. This is the structure we need to compute probability. The role of elementary set theory in probability is to organize the sets into an event space or *σ*-algebra on which we can compute probabilities.

We can now return to Table 1 to compute some additional probabilities. For example, the probability that a patient is a Caucasian male or an Asian male, we might intuitively see as

$$\frac{50+15+10}{120}=\frac{75}{120}=0.625.$$

We may have intuitively known to compute 50 + 15 + 10 = 75 in the numerator. However, now we can see that being either a Caucasian or Asian male requires us to consider four categories, 1) Caucasian Hispanic male, 2) Caucasian non-Hispanic male, 3) Asian Hispanic male, and 4) Asian nonHispanic male. Using our established set operations, Males are unions of these sets. However, each set is disjoint, there are no non-null set intersections to consider, and we simply need add the set content by the properties of disjoint unions.

[*] There are commonly more than one σ-algebras from sets of observations. For example, suppose that Table 1did not categorize individuals by gender, race, and ethnicity, but by age (young vs. old), weight (thin, normal, or obese), and height (short, normal, tall). Then the 120 individuals would be divided not by subsets defined by gender, race, and ethnicity, but by these three new categorizations. This example reveals that, just as a σ-algebra can be a rich collection of subsets, there are multiple σ-algebra that can be produced from each sample size, **Ω** depending on the characteristics that are available to be inspected.

[†] We called this an unmeasurable function.

### *Complements*

One final useful feature in constructing events of interest is the use of complements. Any element that is not in the set $A$ is in the complement of $A$. As we have seen, the complement of a set is its opposite, or $A^c$, termed "$A$ complement" or "Not $A$."

We will now see that the use of complements, unions, and intersections of sets that generated the event space will permit us to more easily compute probabilities. Let's start with some obvious properties.

## Properties of Probability

1. $\mathbf{P}[\Omega]=1.$
2. $\mathbf{P}[\varnothing]=0.$
3. $\mathbf{P}[A]\geq 0 \text{ for any } A\subset\Omega.$

These properties come from the theory of measure, and are covered in our introduction to the properties of measure. However the implications are easy to appreciate. These properties together imply that probability maps sets to the $[0,1]$ interval. Property 3 follows from the simple observation that, if the null set has probability zero, then any non-null subset $A$ of $\Omega$ must also have probability at least as large as the null set. Our rule or measure is to compute probability as a relative frequency satisfies each of these three rules. In general, $\mathbf{P}[A^c]=1-\mathbf{P}[A].$

The next rule has profound implications for computing probabilities of more complex events

4. $\mathbf{P}[A\cup B]=\mathbf{P}[A]+\mathbf{P}[B] \text{ when } A\cap B=\varnothing.$

This last one requires some discussion.

### *Disjoint sets*

Two sets $A$ and $B$ that contain no common elements, i.e., $A\cap B=\varnothing$ are considered to be *disjoint;* the absence of any overlapping elements between the sets eases the computation of the probability of their union – we simply add probabilities.

Property 4 above is the basis of computing probabilities for nondisjoint sets.

When sets $A$ and $B$ are not disjoint we must include one additional term;

$$\mathbf{P}[A\cup B]=\mathbf{P}[A]+\mathbf{P}[B]-\mathbf{P}[A\cap B].$$

Note that this simplifies to Property 4 when A and B are *disjoint.* In fact this statement is a direct consequence of Property 4.

We can also show another useful feature of probability that is intuitive and easy to prove.

If $A$ and $B$ are sets in $\Omega$ and $A\subset B$, then
$\mathbf{P}[A]\leq\mathbf{P}[B].$*

## Other Relationships between Sets

Probability is a set function. How we compute the probability is based on the definition of the probability (relative frequency here, or Poisson measure, or counting measure) and the properties of sets, i.e., the events the sets represent.

Disjointedness is the property of events. Events are disjoint or not. Disjoint events permit us to sum probabilities directly. Non disjoint events ( i.e., $A \cap B \neq \varnothing$ ) require that we subtract the probability of a third, related event, namely the probability of the intersection. How we manage these nondisjoint intersections depends on the relationships between the sets.

### *Independence*

One of the most useful properties of events is the notion of independence. Events can be independent or dependent. The status of independence or dependence defines the relationship between sets.

Two events are independent if the occurrence of one event tells us nothing about the occurrence of the other. Dependent events are events where the occurrence of one event changes our assessment of the likelihood of the other, allowing us to adjust the probability of the second event in light of the observation that the first event occurred. As we might expect, while dependent relationships are the most informative, their probabilities are more complicated to compute.

The descriptors “independence” or “dependence” are properties not of events, but of event relationships. We don’t ask if the occurrence of osteoarthritis is “independent.” But we can ask if the occurrence of osteoarthritis is independent of the subject’s weight. The independence / dependence property is the descriptor of the event’s relationship.

The fundamental feature of independence events is that the occurrence of one does not affect the occurrence of the other. Specifically an observer who notes the occurrence of one event learns nothing about the occurrence of the second event.

Consider the thought process of a doctor who is examining a patient suffering from a bowel disorder that his physician believes may be ulcerative colitis. During the examination, the doctor may notice and record the patient’s height. However, the observation that the patient is six feet tall does not influence the likelihood that the patient is suffering from ulcerative colitis. Height simply does not inform the diagnostic process.

We say that the two events of ulcerative colitis and height are independent of each another [1]. Independent events are denoted by the “$\perp$”, and we denote the independence of events $A$ and $B$ as $A \perp B$.

Computing probabilities of independent events is straightforward. If $A \perp B.$then $\mathbf{P}\left[A \cap B\right] = \mathbf{P}\left[A\right]\mathbf{P}\left[B\right]$. For independent events, the probabilities multiply. [†]

[*] This follows since, if $A \subset B$, we can write $B = \left(A \cap B\right) \cup \left(A^c \cap B\right)$ Since $A \cap B$ and $A^c \cap B$ are disjoint, and $A = A \cap B$ (because $A \subset B$), then $\mathbf{P}\left[B\right] = \mathbf{P}\left[A\right] + \mathbf{P}\left[A^c \cap B\right]$, or $\mathbf{P}\left[A\right] = \mathbf{P}\left[B\right] - \mathbf{P}\left[A^c \cap B\right]$, producing $\mathbf{P}[A] \leq \mathbf{P}[B]$.

[†] Researchers take advantage of this property when they draw random samples of subjects from a larger population. Allowing each subject to have the same probability of being selected from the sample all but ensures that the sample subjects’ measurements are independent of each other. Thus, identifying probabilities of joint events reduces to simply multiplying probabilities of the individual events. This work is the foundation on which the formulas for commonly used statistical estimators (e.g., means, variances, and incidence rates) are formulated.

### *Dependence*

The computation of the probability of two events' joint occurrence is relatively simple if they are disjoint or independent. However, matters become much more interesting, and somewhat complex, when it comes to dependent events.

Suppose we wanted to compute from Table 1 the probability of being both Hispanic and female. If we made the assumption that these two events were independent, we would find the probability of being Hispanic, then the probability of being female, and simply multiply these two probabilities together.

Let's try it. The probability of being Hispanic is $\frac{32+15}{120}=\frac{47}{120}=0.392.$ The probability of being female as $\frac{45}{120}=0.375.$ Assuming these events are independent, we would compute the probability of Hispanic females as (0.392)(0.375) = 0.147.

However, looking at Table 1, we see another way to compute the relative frequency of Hispanic females = $\frac{15}{120}=0.125\neq 0.147!$

What has happened is that our assumption of independence that lead us to multiply probabilities was incorrect. The probabilities are not independent, but dependent.

The dependence property implies that knowledge of the occurrence of one trait informs us about the likelihood of the other's occurrence. However, in order to utilize this property, the nature of the relationship must be clarified. Specifically, the scientist must know exactly how to update her assessment of the occurrence of one event when another dependent event has been generated. This updated assessment is the *conditional probability*.

## Introduction to conditional probability

We may specifically denote the probability of an event *A* when the event *B* has occurred as $\mathbf{P}\left[A\,|\,B\right].$ This probability may be computed as

$$\boldsymbol{P}\left[A\,|\,B\right]=\frac{\boldsymbol{P}\left[A\cap B\right]}{\boldsymbol{P}\left[B\right]}$$

This formula may appear mysterious, but is actually quite intuitive when viewed graphically (Figure 1).

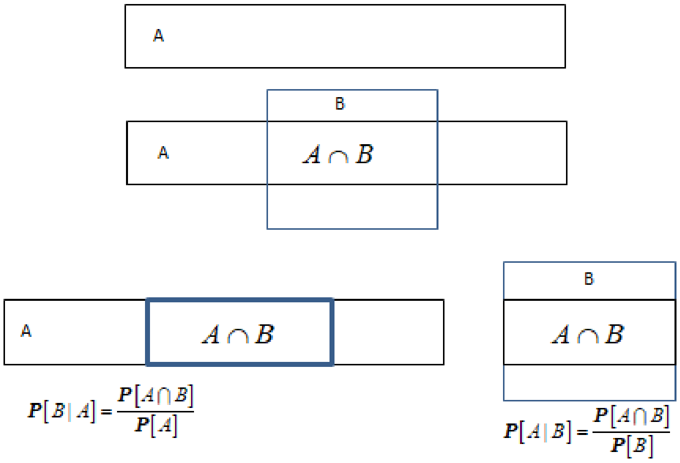

A
B
A
$A \cap B$
B
A
$A \cap B$
$A \cap B$
$P[B \mid A] = \frac{P[A \cap B]}{P[A]}$
$P[A \mid B] = \frac{P[A \cap B]}{P[B]}$

**Figure 1.** How conditional probability is constructed.

In Figure 1, we begin with an event A and then superimpose event B with an area of overlap.

We can understand the conditional probability formula by first recognizing that the event $A \cap B$ resides wholly in *B* (Panel C). We may think of the event *A* given *B* by beginning with the occurrence of event *B,* then measuring the relative size (i.e., the probability) of $A \cap B$ when compared to the size of *B,* asking what fraction of the time that event A also occurs, or $\frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}$. We can use the same reasoning to see that $\mathbf{P}[B \mid A] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[A]}$. Note that, since $\mathbf{P}[A \mid B]$ is the relative probability of the set $A \cap B$ to the set *B*, this conditional probability's value cannot be deduced from simply knowing the probability of *B* (sometimes referred to as the marginal probability of *B*) is large or is small. The $\mathbf{P}[A \mid B]$ can be large when the $\mathbf{P}[B]$ is small, and of course the reverse is true.

Consider the data from Table 1. The probability of being a Caucasian female is $\frac{19}{120} = 0.158$. However the probability that a subject is Caucasian given that they are a female is

$$\frac{19/120}{45/120} = \frac{19}{45} = 0.42$$

The conditional probability is large although the marginal probability of being Caucasian is relatively small.[*] Put another way, Caucasians are more common among females in this sample. It is commonly useful to write rewrite $\mathbf{P}[A \mid B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}$, that

$$\mathbf{P}[A \cap B] = \mathbf{P}[B]\mathbf{P}[A \mid B].$$

This is helpful when we already have the conditional and marginal probabilities, but need to compute the joint probability. While there is interest in $\mathbf{P}[A \cap B]$, our ultimate goal may be to compute yet another conditional probability, $\mathbf{P}[B \mid A]$, as the following example demonstrates

[*] A corollary of this observation is that one must always be careful when describing probability results to be clear what the sample space is, i.e., is the probability a marginal probability or a conditional one.

Of course, if events $A$ and $B$ are independent, then

$$\mathbf{P}[A|B]=\frac{\mathbf{P}[A\cap B]}{\mathbf{P}[B]}=\frac{\mathbf{P}[A]\mathbf{P}[B]}{\mathbf{P}[B]}=\mathbf{P}[A].$$

This concept of conditional probability will be is covered in greater depth shortly.

Introductory Track
Counting Events - Combinatorics
The Notion of Random Events
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Track
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

References

1. Moyé L. (2003). *Multiple Endpoints in Clinical Trials: Fundamentals for Investigators.* New York. Springer.

# Probability and the Measurement of Disease Occurrence

The definition of probability as relative frequency is well ingrained in the investigation of disease. However, in public health, this concept is not particularly useful unless it can be linked to the way in which public health scientists describe the occurrence of disease.  A brief introduction to the measurement of disease occurrence follows.

## Prerequisites

None

## The Dynamics of disease

Events that we have discussed thus far have been fixed characteristics in time. For example, the likelihood that a subject in a population is Asian or is a woman has been based on the assumption that the individual's property or characteristic remains the same. However, disease is different, and one important differentiation is its dynamism.

A population may have no cases of a given illness. When disease arrives it can be heralded by the arrival of a microorganism, or a change in the expression of a person's genotype, or a new environmental exposure.  However, although the disease is now resident in the population, it is not static. The arrival of new cases from the population, or continued exposure to a toxin, or the spread of contagion through the population from population members themselves can increase the number of cases. Advanced epidemiologic models quantitate this.

However, disease cases can also decline. Afflicted individuals can leave the population for another county, state, or country. They can also die. Others may recover on their own, or be successfully treated and cured.

These influences that increase or that depress cases commonly operate contemporaneously. Therefore one can only learn the direction of the disease's presence by being cognizant of the effects of all of these influences.

Probability is useful if we can construct the most helpful event spaces and σ-algebras that capture this dynamism. This is an ongoing challenge in epidemiology and biostatistics. Doing so in the presence of a dynamic disease is challenging.  We begin with the most basic concepts.

## Quotients, Proportions, Rates, and Ratios.

While raw changes in the numbers of patients with disease can be useful, the computations are most helpful if they are compared to or are relative to another quantity. For example comparing the number of deaths due to COVID-19 infection across countries does provide some information, but since country populations differ widely, perhaps it is a greater interest to divide the number of cases by the population to obtain a per subject or per capita assessment of deaths.

A quotient is the most globally descriptive of these concepts, although not all quotients are informative. For example, the number of COVID-19 deaths divided by new car sales is a quotient of little value.

### *Proportions*

A proportion is a relative frequency. Its numerator is subjects who have a specific characteristic among all other subjects. Thus the quotient of subjects with the characteristic divided by all subjects provides a proportion of subjects with the characteristic. We have already utilized this concept in our basic definition of probability as relative frequency.

The key to checking if a quotient is a proportion is to determine if every member of the quotient's numerator is also a member of its denominator. Should this test fail, the use of this quotient as a probability is very suspect.

### *Rates*

Rates in epidemiology are proportions that are anchored to a particular time interval. Is the number of deaths in a particular period of time (e.g., a day, or a week, or a month) divided by the number of individuals in the population during that period of time. The case fatality rate is the number of death over a particular period of time divided by the number of people who were diagnosed with the disease (cases).

Note that these two rates each deal with death, but the differences in the denominators changes the meaning of the rate. The monthly death rate provides the likelihood that the individual dies given they are in the population, while the monthly case fatality rate reveals the likelihood that a patient dies given, not that they are in the population, but that they actually have the disease.

For example, the monthly case fatality rate can be high, and the death rate low, if very few people in the population have the disease, but those who contract the disease are likely to die from it.

.

### *Ratios*

A ratio is a quotient whose numerator and denominator is each a rate. For example, the death rate for men can be compared to a death rate for women by computing a ratio. Prevalence ratios and incidence ratios are among the most common ratios in contemporary epidemiology.

## Prevalence, Background, and Incidence

There are three quantities that epidemiologists most commonly follow. The first is the incidence rate. This is the number of new cases per population size per time.

The second is the background rate, or the number of cases per population size that already reside in a community.

The third is the prevalence, which is simply the total of all cases in a community per population size regardless of whether they are new or have been in the community for some time.

There is commonly a profound difference between the incidence rate, and the prevalence rate.

Thus, it is important to distinguish between these rates when one is describing the probability of disease in a community. For example, for diabetes mellitus, the incidence rate may be 15 cases per 1000 per year, while the prevalence rate may be 19% (or 19 cases of diabetes for each 100 patients in the community.

Disease with high and early lethality (i.e., a high case fatality rate), can have incidence rates similar to prevalence rates, since the high mortality rate reduces the background rate to zero.

Counting Events

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Immigration-Emigration Modeling
Contagion
Death Process
The Emigration-Death Process
Immigration-Death Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Probabilities of Unions

We wish to demonstrate

$$\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B].$$

Our plan is to write this formula as the union of disjoint sets. Using our background in set theory, we begin by recognizing that another way to write the set $A$ is

$$A \cup B = A \cup \left(A^c \cap B\right).$$

which is seen from the Venn diagram of Figure 1. Since the sets $A$ and $A^c \cap B$ are disjoint, we may write

$$\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}\left[A^c \cap B\right]$$

## Helpful Construction

Now from the same diagram we see that the set $B$ can be written as

$$B = (A \cap B) \cup \left(A^c \cap B\right).$$

And since the sets $A \cap B$ and $A \cap B^c$ are disjoint, we may write

$$\mathbf{P}[B] = \mathbf{P}[A \cap B] + \mathbf{P}\left[A^c \cap B\right] \text{ or}$$
$$\mathbf{P}\left[A^c \cap B\right] = \mathbf{P}[B] - \mathbf{P}[A \cap B].$$

Now returning to our first equation we can now write

$$\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}\left[A^c \cap B\right]$$
$$= \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B].$$

Properties of Probability
Basic Properties of Probability
Counting Events
Measure and its Properties

# Conditional Probability Bayes Theorem

Prerequisite
Basic Properties of Probability

Recall from its earlier definition of dependent events that we may write the probability of an event $A$ when the event $B$ has occurred as $\mathbf{P}\left[A \mid B\right]$. This probability may be computed as

$$\mathbf{P}\left[A \mid B\right] = \frac{\mathbf{P}\left[A \cap B\right]}{\mathbf{P}\left[B\right]}$$

We have motivated this definition before through the use of figures. However, an axiomatic approach also has value. We may think of the event $B$ as the union of two separate events; $A \cap B$ and $A^c \cap B$. The concept of $A$ given $B$ must exclude $A^c$ since $A$ and $A^c$ are mutually exclusive. Thus we want to compare two probabilities, $\mathbf{P}\left[A \cap B\right]$ and $\mathbf{P}\left[B\right]$. Since we know that $\mathbf{P}\left[B\right] = \mathbf{P}\left[A \cap B\right] + \mathbf{P}\left[A^c \cap B\right]$,

$\mathbf{P}\left[A \cap B\right] \le \mathbf{P}\left[B\right]$. Thus, their ratio provides the relative size of the measures.

If for example $\mathbf{P}\left[A \mid B\right] = 1$, then $\mathbf{P}\left[A \cap B\right] = \mathbf{P}\left[B\right]$ and $\mathbf{P}\left[A^c \cap B\right] = 0$. Thus the event $A^c$ cannot occur when $B$ occurs. When $B$ occurs, the event $A$ must occur. A similar line of reasoning for $\mathbf{P}\left[A \mid B\right] = 0$ reveals that in the presence of $B$ only $A^c$ can occur.

## Example:

Atherosclerotic ischemic cardiovascular disease is very common in the United States and can lead to a heart attack (myocardial infarction), and subsequently, heart failure. Heart failure is a serious consequence; regardless of modern medical therapy, approximately 50% of patients with heart failure die within five years of the diagnosis.

From an observation in a cardiac clinic, we know that the probability that an individual has heart failure is 0.60, and the probability that they have had a heart attack is 0.33. The probability that, that they will have a heart attack, given that they have had heart failure in the past is 0.45. What we have been asked to compute is the probability the individual will have heart failure given they have suffered a heart attack.

**Example:**

Let *A* be the event that a patient has had a heart attack, and $HF$ be the probability that the patient has had heart failure. Then we must compute $\mathbf{P}\left[HF \mid A\right]$. We begin by writing

$$\mathbf{P}\left[HF \mid A\right] = \frac{\mathbf{P}\left[HF \cap A\right]}{\mathbf{P}\left[A\right]}.$$

We have been provided $\mathbf{P}\left[A\right] = 0.33$ but do not know the joint probability $\mathbf{P}\left[HF \cap A\right]$. However, we have been provided $\mathbf{P}\left[HF\right]$ and $\mathbf{P}\left[A \mid HF\right]$.

This permits us to write

$$\mathbf{P}\left[A \mid HF\right] = \frac{\mathbf{P}\left[HF \cap A\right]}{\mathbf{P}\left[HF\right]}$$

$$\mathbf{P}\left[HF \cap A\right] = \mathbf{P}\left[A \mid HF\right]\mathbf{P}\left[HF\right].$$

And

$$\begin{aligned}\mathbf{P}\left[HF \mid A\right] &= \frac{\mathbf{P}\left[HF \cap A\right]}{\mathbf{P}\left[A\right]} = \frac{\mathbf{P}\left[A \mid HF\right]\mathbf{P}\left[HF\right]}{\mathbf{P}\left[A\right]} \\ &= \frac{(0.45)(0.60)}{0.33} = 0.82.\end{aligned}$$

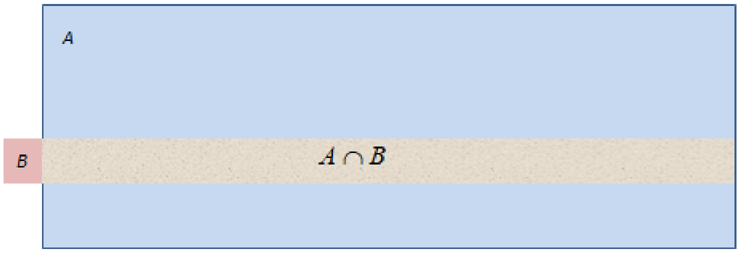
Thus, in this sample, while less than fifty percent of patients who have a heart failure have had a heart attack $\mathbf{P}\left[A \mid HF\right] = 0.45$, most patients with a heart attack progress to heart failure in this study, $\mathbf{P}\left[HF \mid A\right] = 0.82$. The two conditional probabilities $\mathbf{P}\left[A \mid HF\right]$, and $\mathbf{P}\left[HF \mid A\right]$ differ substantially.* The ability to move from one conditional probability to another, essentially reversing the condition, is known as the inversion process and has an interesting history.

How these conditional probabilities can be so different is worth examination (Figure 1) Health care providers rely on conditional probability – in fact, it is implicit in the differential diagnostic process.

∎

To understand, given two events A and B, the difference in magnitude between $\mathbf{P}\left[A \mid B\right]$ and $\mathbf{P}\left[B \mid A\right]$ we simply need to compare the magnitude of $\mathbf{P}\left[A \cap B\right]$ to each of the two marginal probabilities $\mathbf{P}\left[A\right]$ and $\mathbf{P}\left[B\right]$ (Figure 2). When most of the event B also includes A, $\mathbf{P}\left[A \mid B\right]$ will be large. However, that same figure (Panel 2) shows that when only a small fraction of the event *A* is also made up of event *B,* then $\mathbf{P}\left[B \mid A\right]$ is quite small.

* This occurred because the relatively large probability of patients with heart failure (0.60) increased the percentage of patients who had heart failure among those in the heart attack population.



$\mathbf{P}[B]$ and $\mathbf{P}[B \cap A]$ are approximately equal.

$\mathbf{P}[A \mid B] = \dfrac{\mathbf{P}[B \cap A]}{\mathbf{P}[B]}$ is close to one.

$\mathbf{P}[A]$ is much greater than $\mathbf{P}[B \cap A]$.

$\mathbf{P}[B \mid A] = \dfrac{\mathbf{P}[B \cap A]}{\mathbf{P}[A]}$ is small.

**Figure 1**. In the above panel most of event B is also contained in event A, so the $\mathbf{P}[A|B]$ is large. However, since very little of Event A is made up of Event B, $\mathbf{P}[B|A]$ is small..

In general, we have three circumstances in which we can compute the probability of the joint event, $A \cap B$ (Figure 2). We first determine if the events are either mutually exclusive, independent or dependent. If mutually exclusive, then $\mathbf{P}[A \cap B] = 0$. If the events are independent, then $\mathbf{P}[A \cap B] = \mathbf{P}[A]\mathbf{P}[B]$. Finally, if we find the events are dependent, and we can compute conditional probabilities, then $\mathbf{P}[A \cap B] = \mathbf{P}[A]\mathbf{P}[B \mid A] = \mathbf{P}[B]\mathbf{P}[A \mid B]$.

## Introduction to law of total probability

We have been free to categorize events as simple or as complex as we like. Returning to Table 1, we can consider events as simple as a subject being African-American, or as complex as nonHispanic Caucasian females. Sometimes it is of value to consider two types of events.

Let *A* be the event that a subject is African-American male and *H* be the event that a subject is Hispanic male. Then from the elaboration of all events on which we can assign probability (i.e., having established the $\sigma$-algebra of subsets over which we can assign probability), we know that subjects are either African American males or not. $A$ and $A^c$ are clearly mutually exclusive. Since there are only two possibilities for a subject we say that they "exhaust the space". Thus

$$\mathbf{P}\left[A \cup A^c\right] = \mathbf{P}[A] + \mathbf{P}\left[A^c\right] - \mathbf{P}\left[A \cap A^c\right] = \mathbf{P}[A] + \mathbf{P}\left[A^c\right] = 1.$$

We can write a similar equality for male subjects when classified as Hispanic or not.

In order to compute the $\mathbf{P}[A]$ from Table 1 we see that we do have a total for African American males and compute directly $\mathbf{P}[A] = \dfrac{10}{120} = 0.083$.

However, we might go about this computation another way. African-American males are either Hispanic or nonHispanic. Therefore, we can begin with the probabilities that a subject is an Hispanic African-American male, $\mathbf{P}[A \cap H]$, or a nonHispanic African-American male, $\mathbf{P}\left[A \cap H^c\right]$. Each of these are joint probabilities that are mutually exclusive. And since they exhaust all of the possibilities of being an African-American male, we may write from Table 1,

$$\mathbf{P}[A]=\mathbf{P}[A\cap H]+\mathbf{P}\left[A\cap H^{c}\right]=\frac{2}{120}+\frac{8}{120}=0.083.$$

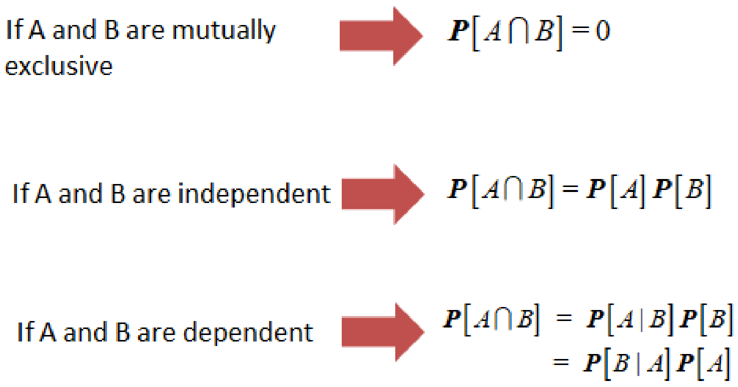
This calculation worked because the computation considered all Table 1-based ethnic considerations of African American males (Hispanic and nonHispanic).

We might think of this as "holding African American males constant" and summing over the two mutually exclusive possibilities for the Hispanic event. This is the essence of the Law of Total Probability.

Another way to write $\mathbf{P}[A\cap H]+\mathbf{P}\left[A\cap H^{c}\right]$ is

$$\mathbf{P}[A|H]\mathbf{P}[H]+\mathbf{P}\left[A|H^{c}\right]\mathbf{P}\left[H^{c}\right]$$

using the definition of conditional probability. Thus we have two ways of computing a marginal probability indirectly using the law of total probabilities. One way is through summing joint probabilities, the second is through summing a combination of conditional and marginal probabilities (Figure 2). We will use this law of total probability extensively in discussions of compounding.



If A and B are mutually exclusive → $P[A\cap B]=0$

If A and B are independent → $P[A\cap B]=P[A]\,P[B]$

If A and B are dependent → $P[A\cap B] = P[A\,|\,B]\,P[B] = P[B\,|\,A]\,P[A]$

**Figure 2.** Computing the probability of joint events. First determine if the events are mutually exclusive, independent, or dependent.

## *Example: Diagnostic Value of Cell Phenotypes*

A major area of investigation has been cell therapy, where a patient's own cells, when removed from one organ (commonly the bone marrow) and provided to another organ (e.g., the heart) can improve function in the destination organ. However, not all cell types are the same, and patients with a particular phenotype, CD34$^{+}$ have been of interest. A study shows that in patients who have had cell therapy and experience an improvement in cardiac function, the probability that they have a high level of CD34$^{+}$ is 0.85. The probability that a patient who has no improvement has a low level of CD34$^{+}$ is 0.90. The overall (i.e., marginal) probability of improvement in cardiac function is 0.18.

The health care provider is given a subject with a high level of CD34$^{+}$ What is the probability that the subject will experience an improvement in heart function?

Here we are asked to, given that a patient has a high level of CD34$^{+}$, to predict forward to what will happen to their heart function. We have the results of a study to guide us. However, that study did not start with knowledge of levels of CD34$^{+}$ and look forward to heart function, but instead, started with the identification of patients with heart function, identified those who improved and those who did not, and looked backwards to see how many of each had elevated $\mathrm{CD34}^{+}$. This is the classic inversion problem.

Let $I$ be the event of heart function improvement, and $C$ be the event that a patient has a high level of $\mathrm{CD34}^{+}$ cells. We are interested in $\mathbf{P}[I \mid C]$. However, we are given $\mathbf{P}[C \mid I]$, $\mathbf{P}\left[I^{c} \mid C^{c}\right]$, and $\mathbf{P}[I]$. We begin by writing

$$\mathbf{P}[I \mid C]=\frac{\mathbf{P}[I \cap C]}{\mathbf{P}[C]}.$$

Now write the numerator as a function of the inverted conditional probability, using $\mathbf{P}[C \mid I]=\dfrac{\mathbf{P}[I \cap C]}{\mathbf{P}[I]}$ or $\mathbf{P}[I \cap C]=\mathbf{P}[C \mid I]\mathbf{P}[I]$.

The denominator $\mathbf{P}[C]$ can be written using the law of total probability as $\mathbf{P}[C]=\mathbf{P}[C \cap I]+\mathbf{P}\left[C \cap I^{c}\right]$. We know $\mathbf{P}[C \cap I]=\mathbf{P}[C \mid I]\mathbf{P}[I]$. To find $\mathbf{P}\left[C \cap I^{c}\right]$ we only need write $\mathbf{P}\left[C \mid I^{c}\right]=\dfrac{\mathbf{P}\left[C \cap I^{c}\right]}{\mathbf{P}\left[I^{c}\right]}$ to see $\mathbf{P}\left[C \cap I^{c}\right]=\mathbf{P}\left[C \mid I^{c}\right]\mathbf{P}\left[I^{c}\right]$. So we may write

$$\begin{aligned}\mathbf{P}[I \mid C]&=\frac{\mathbf{P}[C \mid I]\mathbf{P}[I]}{\mathbf{P}[C \mid I]\mathbf{P}[I]+\mathbf{P}\left[C \mid I^{c}\right]\mathbf{P}\left[I^{c}\right]}\\&=\frac{\mathbf{P}[C \mid I]\mathbf{P}[I]}{\mathbf{P}[C \mid I]\mathbf{P}[I]+\left(1-\mathbf{P}\left[C^{c} \mid I^{c}\right]\right)\mathbf{P}\left[I^{c}\right]}\end{aligned}$$

We can now solve to find

$$\begin{aligned}\mathbf{P}[I \mid C]&=\frac{(0.85)(0.18)}{(0.85)(0.18)+(1-0.90)(0.82)}\\&=\frac{0.15}{0.15+0.08}=\frac{0.15}{0.23}=0.65.\end{aligned}$$

A answer which is somewhat lower than $\mathbf{P}[C \mid I]$.

In the previous example, we actually proved Bayes Theorem, attributed to Thomas Bayes and Richard Price, which states that given events $A$ and $B$, we can write

$$\mathbf{P}[A \mid B]=\frac{\mathbf{P}[B \mid A]\mathbf{P}[A]}{\mathbf{P}[B \mid A]\mathbf{P}[A]+\mathbf{P}\left[B \mid A^{c}\right]\mathbf{P}\left[A^{c}\right]}$$

It is a simple and elegant use of the law of total probablity, and is commonly used in assessing diagnostic tests

Conditional Probability
The Inversion Problem
Physicians and Conditional Probability
Assessing Diagnostic Tests

# The Inversion Problem

Prerequisite
Basic Properties of Probability
Conditional Probability

Conditional probability historically focused on the ability to deduce cause from effect mathematically.

By the mid-18th century, probability was an accepted, even respected branch of applied mathematics. Its users at the time were well acquainted with the binomial probability distribution, which computes from a sequence of *n* independent success-failure trials the probability that there are exactly *k* successes. The use of this elementary probability model generated some useful conclusions that sparked new interest in the relationship between cause and effect, in a new, contentious, and illuminating way.

## Probability and 18th Century Armies

Consider one of the most virulent scourge of the time - diarrhea. While the effects of diarrheal disease were disabling in urban life, the disease was devastating to an army.

At the time, organized meal preparation was an unknown concept in the armies of Europe. Each man was responsible for bringing his own utensils, carrying his own food, and cooking his own meal over a group campfire. However, careful observers quickly noticed that, while diarrhea did not have just one cause, it appeared to be related to how a soldier's meal was prepared.* Specifically, diarrhea was more prevalent in soldiers who prepared their meals with unboiled water. Racing through camps, it could bring huge segments of otherwise mobile units to a standstill, removing thousands of men from a battle at a critical moment.

This was a serious issue that involved a nation's readiness for war, and probability was used to help understand the problem. For example, it was easy to compute how many soldiers out of twenty would be sick if only 10% of them boiled their water. This was termed "reasoning from cause to effect," using knowledge of the frequency of boiling water to compute "forward" to the effect of that habit i.e., predicting the number of soldiers expected to be sick. This was a straightforward, correct, and commonly helpful probability application.

* Other causes were spoiled food, wounds, and sepsis.

However, suppose one reverses the logic. Now, the physician observes twenty soldiers, six of whom have diarrhea. How likely is it that unboiled cooking water is the cause of the diarrhea ? In this circumstance, the worker is compelled to reason "backward" from the effect (i.e., the sick soldiers) to the cause of their illness (unboiled water or some other cause). This reversal of the deduction process was mathematically known as inversion. Given events *A* and *B*, how could one fluidly move from knowledge of $\mathbf{P}\left[A \mid B\right]$ to the more interesting and useful $\mathbf{P}\left[B \mid A\right]$. The reversing of the condition was known as "inversion" or "reversing the given." *

The first semiformal method to solve this problem was provided posthumously by the Rev. Bayes, encapsulated in Bayes Theorem. In modern terms, it states that the probability of the hypothesis given the evidence can be computed from the probability of the evidence given the hypothesis, or

$$\mathbf{P}\left[\text{Hypothesis} \mid \text{Evidence}\right] = K\,\mathbf{P}\left[\text{Evidence} \mid \text{Hypothesis}\right]\mathbf{P}\left[\text{Hypothesis}\right].$$

where *K* is a proportionality constant. Further developed by Simon Laplace, this was the introduction to Bayesian statistics.

Physicians and Conditional Probability
Assessing Diagnostic Tests
Counting Events

* Knowledge of one conditional probability does not imply knowledge of the probability with the conditions inverted or reversed. In a modern day context, the probability that given the car is a Ferrari, a male is driving is high. However, the inverse probability, i.e., the probabilty that given the male is a driver, the car that he is driving is a Ferrari is low.

# Physicians and Conditional Probability

Prerequisites
Basic Properties of Probability
Conditional Probability
The Inversion Problem

## Conditional Probability and Diagnoses

Physicians commonly use conditional probability each day in their practices perhaps without being formally aware of it. Patients admitted to the hospital suspected of having a stroke are commonly administered tissue plasminogen activator factor (tPA) to limit the extension of the stroke. However, because of its tendency to produce intracerebral bleeding, it is best to give tPA within three hours of onset of symptoms, requiring 1) the rapid identification of the ill patient, 2) their rapid delivery to the hospital, and 3) the rapid administration of tPA.

A patient, their anxious family waiting just outside the exam room, undergoes a swift evaluation as the doctor rapidly works to identify the cause of the patient's symptoms. She may ask the family, "Does anyone in the family know if she has diabetes, or hypertension? Has she or anybody in her family had a stroke?"

The rapidly closing window of whether the patient can receive tPA requires the physician to ask the most informative questions. The answers to these questions alter and update his assessment of Other questions involve the patient herself. Is the patient conscious? Does she have new difficulty controlling her eye movements, and do her pupils react appropriately to light? Are there new facial asymmetries? Does she have sudden, new difficulty controlling the movement of her limbs?

the likelihood the patient has suffered a stroke.

Each of these questions has a well-deserved place in the evaluation of the patient, because each is believed to alter the probability that a patient has had a stroke through

Conditional probability is also useful because it is commonly difficult to specify the nature of the dependency persuasively and completely using other quantitative approaches.

Consider, for example, the relationship between health care access and cultural background. It has been well established that some cultures in the United States visit physicians and health care providers more commonly, receive prescriptions at a greater frequency, and are more likely to receive prenatal care than others.

However, the precise nature of the connection is unknown, and there is no equation that precisely depicts the relationship. Conditional probability allows us to formulate the relationship by computing different probabilities of health care access for different cultural backgrounds.

The differences in these probabilities are one of the best descriptors of the nature of the relationship between culture and health care access. They delineate the magnitude of the relationship without having to elucidate the dependency's nature.

Bayes Theorem
Assessing Diagnostic Tests

Counting Events

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Assessing Diagnostic Tests

Prerequisites
Basic Properties of Probability
Bayes Theorem
Assessing Diagnostic Tests
Conditional Probability
The Inversion Problem

## Goal of diagnostic testing

Diagnostic testing, represented by either an imaging procedure, laboratory testing for infectious disease, or a collection of sequential cancer screening procedures, to be helpful, must contribute information about the presence of a disease or a condition. It should provide a high level of confidence either making or ruling out the diagnosis of a particular disease. Ideally, we wish to have the procedure be helpful when it is positive, and also informative when it is negative.

Of course in health care, we desire certainty. If the test is positive, we want to ensure that the patient has the disease. This is called high positive predictive value. One such test would be high forced expiratory volume in one second (FEV1) tests for reactive airways disease.

We would also like for the test to have high negative predict value, i.e., if the test is negative, then the subject does not have the disease. An example of such a test would be PKU testing for newborn phenylketonuria.

Ideally, all diagnostic testing would have both high positive and negative predictive value. Unfortunately this is rarely the case, In the absence of certainty, how should we proceed?

## Developing a diagnostic test.

In assay development, testing is carried out on subjects who are known to have the disease $\left(\mathrm{D}^{+}\right)$ and also in subjects who are known to be disease free $\left(\mathrm{D}^{-}\right)$. Of course, individuals are identified who have positive $\left(\mathrm{T}^{+}\right)$ and negative $\left(\mathrm{T}^{-}\right)$ tests.

Since we know the number of individuals in both populations of disease and non-diseased subjects tested, we can compute several useful quantities with relative frequency. We can compute the sensitivity of the test. This is the probability the subject has the positive test, given that they have the disease, or $\mathbf{P}\left[\mathrm{T}^{+} \mid \mathrm{D}^{+}\right]$. Clearly we want this to be high as possible.

We also can compute the likelihood that the test is negative given that the patient does not have the disease, or $\mathbf{P}\left[\mathrm{T}^{-} \mid \mathrm{D}^{-}\right]$. This should be as high as possible as well. This is the specificity of the test. We desire tests with high specificity and sensitivity.

But this is not all that we need because of the impact of inversion. Recall that specificity and sensitivity are conditional probabilities based on knowledge of the disease. The “given” in the conditional probability is the disease state.

However, health care practitioners are not given the disease state. They wish to learn the disease state, given the test result. They know if the test is positive or not. They would like the probability of disease given the test result. These probabilities are known as “predictive value” since they are predicting the presence of the disease. The predictive values of interest are positive predictive value (PPV) which is $\mathbf{P}\left[\mathrm{D}^{+} \mid \mathrm{T}^{+}\right]$ and negative predictive value (NPV) or $\mathbf{P}\left[\mathrm{D}^{-} \mid \mathrm{T}^{-}\right]$ . These are related to, but separate from the test’s sensitivity and specificity.

We move from sensitivity and specificity to PPV and NPV using Bayes theorem. Specifically

$$PPV = \frac{\mathbf{P}\left[\mathrm{D}^{+} \cap \mathrm{T}^{+}\right]}{\mathbf{P}\left[\mathrm{T}^{+}\right]} = \frac{\mathbf{P}\left[\mathrm{T}^{+} \mid \mathrm{D}^{+}\right]\mathbf{P}\left[\mathrm{D}^{+}\right]}{\mathbf{P}\left[\mathrm{T}^{+} \mid \mathrm{D}^{+}\right]\mathbf{P}\left[\mathrm{D}^{+}\right] + \mathbf{P}\left[\mathrm{T}^{+} \mid \mathrm{D}^{-}\right]\mathbf{P}\left[\mathrm{D}^{-}\right]}$$

$$NPV = \frac{\mathbf{P}\left[\mathrm{D}^{-} \cap \mathrm{T}^{-}\right]}{\mathbf{P}\left[\mathrm{T}^{-}\right]} = \frac{\mathbf{P}\left[\mathrm{T}^{-} \mid \mathrm{D}^{-}\right]\mathbf{P}\left[\mathrm{D}^{-}\right]}{\mathbf{P}\left[\mathrm{T}^{-} \mid \mathrm{D}^{-}\right]\mathbf{P}\left[\mathrm{D}^{-}\right] + \mathbf{P}\left[\mathrm{T}^{-} \mid \mathrm{D}^{+}\right]\mathbf{P}\left[\mathrm{D}^{+}\right]}$$

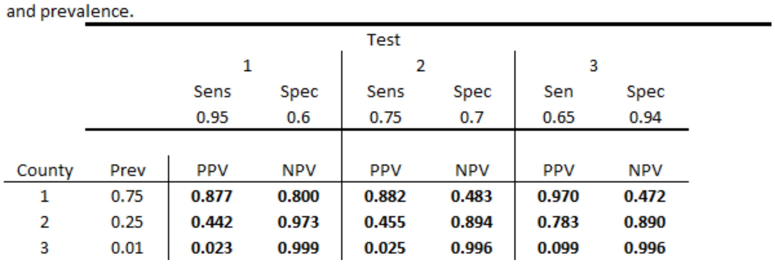
Note that these computations are based on not just functions of sensitivity and specificity, but on the prevalence of the disease $\mathbf{P}\left[\mathrm{D}^{+}\right]$ as well.

## Example COVID-19 testing

Several tests are available to test for the presence of COVID19 coronavirus. Assume that we have three such candidate tests. Test 1 has a sensitivity of 0.95 and a specificity of 0.60. Test 2 has a sensitivity and specificity of 0.75 and 070 respectively. The sensitivity of Test 3 is measured as 0.65 with a specificity of 0.94.

These tests are being considered by three different counties. County 1 has a COVID19 prevalence of 0.75. County 2’s prevalence if 0.25 and County 3’s prevalence is 0.01. Is there an optimal test for all three counties?

Table 1.Positive and negative predictive value as a function of sensitivity, specificity, and prevalence.



| | | Test | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | 3 | |
| | | Sens | Spec | Sens | Spec | Sen | Spec |
| | | 0.95 | 0.6 | 0.75 | 0.7 | 0.65 | 0.94 |
| County | Prev | PPV | NPV | PPV | NPV | PPV | NPV |
| 1 | 0.75 | **0.877** | **0.800** | **0.882** | **0.483** | **0.970** | **0.472** |
| 2 | 0.25 | **0.442** | **0.973** | **0.455** | **0.894** | **0.783** | **0.890** |
| 3 | 0.01 | **0.023** | **0.999** | **0.025** | **0.996** | **0.099** | **0.996** |

Table 1 provides the PPV and NPV for each of the three counties for each of the three tests. Note that each of the counties will need to make an individual choice of tests, based on that county's prevalence.

Physicians and Conditional Probability
Assessing Diagnostic Tests

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Counting Events

## Prerequisites

This section requires prior study of the following sections.
Elementary Set Theory
Basic Properties of Probability
Conditional Probability
Sigma Notation
Factorials Permutations, and Combinations

Thus far, our definition of probability as relative frequency has been intuitive and relatively easy to use, so long as we can count 1) all possible outcomes in the denominator, and then 2) those outcomes in the denominator that meet the requirement for the event of interest and therefore can also be entered into the numerator. This section will review important ways to enumerate events in a way that permits us to apply the relative frequency definition of probability to compute the probability of increasingly sophisticated events.

## Counting Repeated Events

The first occurrence of an event can sometimes be counted in a straightforward way. However, repeated events can be more complicated. Consider the following example.

### *Example: Physician Scheduling*

An administrator must ensure that doctors are assigned according to a schedule that adequately staffs the emergency department. She has ten doctors who are available for scheduling. What is the probability that any particular physician is selected to staff the clinic next Monday?

Of course there are ten individual physicians from which one can make the selection. The administrator can choose from among a variety of rules (for example, choose the doctor whose surname occurs earliest in the alphabet, or choose the oldest). However, let us assume that she chooses the physician randomly.

### *Random selection*

By random, we mean that each member of the population (in this case each of the ten physicians) has the same likelihood of being selected as any other physician in the sample. Furthermore, the selection mechanism is completely independent of any characteristic (e.g., height or personality) of the physician.

This is a very specific definition, and somewhat at odds with the general perception in the culture, which can perceive random events as unplanned, chaotic, and uncontrollable. However, we will see that in probability models, processes will be carefully plan, not to provide the same result, but to produce unpredictability on which we can capitalize.

The random selection mechanism in this model helps us to compute the probability of individual physician selections. Let $P_i$ denote the $i^{\text{th}}$ physician, $i = 1,\ldots,10.$ Since one physician must be selected we can write $\Omega = \{P_1, P_2, P_3, \ldots, P_{10}\}$, and create the σ-algebra of all of the subsets of these 10 entries.

Let $S_1$ be the event that a physician is the first selected, for example, $S_1 = P_7$.

We know that one physician must be selected. Thus $\mathbf{P}\left[\bigcup_{i=1}^{10}\{S_1 = P_i\}\right] = 1$. However, since we also know that these events are disjoint, (because only one physician can be selected first, and the selection of one physician excludes the possibility of selecting any of the others), we write $\mathbf{P}\left[\bigcup_{i=1}^{10}\{S_1 = P_i\}\right] = \sum_{i=1}^{n}\mathbf{P}[S_1 = P_i] = 1.$

In addition, the process of random selection here means that the selection of any physician is as equally likely as any other physician. Thus $\mathbf{P}[S_1 = P_i] = p$ for each of the ten events, and we can now write $\sum_{i=1}^{10}\mathbf{P}[S_1 = P_i] = \sum_{i=1}^{10} p = p + p + p + + \ldots = 10p.$ Thus $10p = 1$ or $p = \frac{1}{10}$.

This is the formal proof of the probability solution that $\mathbf{P}[S_1 = P_1] = \frac{1}{10}$. While perhaps this seems like the long way around to solve a probability problem whose answer might have been easily intuited, it is useful to show how the sequence of formal steps produces the correct solution.

Also, unlike from Table 1 in Definitions and Rules, we required no data to compute the probability. Instead, the probability was calculated from intimate knowledge of the problem. However we had to understand the experiment to derive it accurately.

■

Now, however, suppose a second physician is to be selected (represented by event $S_2$ ). What is the probability that any particular physician will be the second physician selected? Before we can compute the probability of this event, we must ask, “What happened to the first one?

## Sampling schemes

How we proceed depends on the circumstances of the problem. Suppose, for example, that the second selection is for staffing the clinic for a day that is a month from now. Who are the candidate physicians?

### *Sampling With Replacement*

In this case, given that a month has lapsed since the physician first selected has completed her assignment, it makes sense to place the first physician back into the pool of possible candidates for the second selection. This reinsertion of the physician back into the population of choices, permitting the possibility that they may be selected again, is called s*ampling with replacement*.[*]

Sampling with replacement eases the computation of probabilities. The probability that the same physician is selected the second time is the same probability as their being selected the

[*] This is somewhat of a misnomer since no one is being replaced, but instead the selected subject is a candidate for an additional selection. However, we will continue to use the historical language.

first time, or $\frac{1}{10}$. Thus the probability that any given physician is selected the first time and then again on the second time is the same.

This makes sense because for each selection the candidate physician pool is the same, and the selection mechanism is independent, i.e., knowledge of who was selected the first time does not inform us one way or the other as to who will be selected the second time.

We can confirm this by computing the probability that (for example) physician three is selected the second time given that this same physician was selected the first time

$$\mathbf{P}\left[S_2 = P_3 \mid S_1 = P_3\right] = \frac{1}{10} = \mathbf{P}\left[S_2 = P_3\right]^{*}$$

Sampling with replacement allows these conditional and marginal probabilities to be independent, easing the computational burden of some complicated events. For example, if we carry out this physician selection mechanism many times using this same mechanism of sampling with replacement, then the probability that physician 3 is selected to fill any of these schedule slots on any particular time is $\frac{1}{10}$.

### *Sampling Without Replacement*

However, let's change the paradigm. Assume now that a physician is selected on the first night, and that the second selection is for service on the next consecutive night. Here, the physician selected for the first night cannot be selected again. How do we manage this?

Clearly, the event space changes, and the probabilities change. It is impossible to select the same physician. If we focus on, for example, physician three, then while $\mathbf{P}\left[S_1 = P_3\right] = \frac{1}{10}$. then $\mathbf{P}\left[S_2 = P_3 \mid S_1 = P_3\right] = 0.$ However, for the other nine physicians indexed by $j$, $j$ = 1,2,4,5,…9, then $\mathbf{P}\left[S_2 = P_j \mid S_1 = P_3\right] = \frac{1}{9}$. Sampling without replacement reduces the number of candidate physicians by one, and since the selection is equally likely among the remaining nine physicians, the probability of any of the nine candidate physicians being selected is $\frac{1}{9}$.

This is clearly a case of dependency between the first and second selections. Sampling without replacement induces a complication in our experiment that required a more complex probability computation.

To demonstrate this dependence another way, let's show that the marginal probability $\mathbf{P}\left[S_2 = P_3\right]$ is equal to neither $\mathbf{P}\left[S_2 = P_3 \mid S_1 \neq P_3\right] = \frac{1}{9}$ nor $\mathbf{P}\left[S_2 = P_3 \mid S_1 = P_3\right] = 0.$ To find $\mathbf{P}\left[S_2 = P_3\right]$ we use the law of total probability to write

* We can see this from the conditional probability of the second physician given the first was selected is simply $\mathbf{P}\left[S_2 = P_3 \mid S_1 = P_3\right] = \frac{\mathbf{P}\left[S_1 = P_3 \cap S_2 = P_3\right]}{\mathbf{P}\left[S_1 = P_3\right]} = \frac{\mathbf{P}\left[S_1 = P_3\right]\mathbf{P}\left[S_2 = P_3\right]}{\mathbf{P}\left[S_1 = P_3\right]} = \mathbf{P}\left[S_2 = P_3\right] = \frac{1}{10}.$

$$\begin{aligned}\mathbf{P}\left[S_2 = P_3\right] &= \mathbf{P}\left[S_2 = P_3 \mid S_1 \neq P_3\right]\mathbf{P}\left[S_1 \neq P_3\right] \\ &\quad + \mathbf{P}\left[S_2 = P_3 \mid S_1 = P_3\right]\mathbf{P}\left[S_1 = P_3\right] \\ &= \left(\frac{1}{9}\right)\left(\frac{9}{10}\right) + 0\frac{1}{10} = \frac{1}{10}.\end{aligned}$$

Thus, we have confirmed the dependency.

Now, suppose we wanted the joint probability that the first selection was not the third physician and the second selection was, expressed as $\mathbf{P}\left[S_1 \neq P_3 \cap S_2 = P_3\right]$. Using our definition of conditional probability, we write,

$$\mathbf{P}\left[S_1 \neq P_3 \mid S_2 = P_3\right] = \frac{\mathbf{P}\left[S_1 \neq P_3 \cap S_2 = P_3\right]}{\mathbf{P}\left[S_2 = P_3\right]}$$

$$\begin{aligned}\mathbf{P}\left[S_1 \neq P_3 \cap S_2 = P_3\right] &= \mathbf{P}\left[S_1 \neq P_3 \mid S_2 = P_3\right]\mathbf{P}\left[S_2 = P_3\right] \\ &= \left(\frac{1}{9}\right)\left(\frac{9}{10}\right) = \frac{1}{10}.\end{aligned}$$

However, this is just the numerator of the solution. To finish, we find

$$\mathbf{P}\left[S_1 \neq P_3 \mid S_2 = P_3\right] = \frac{\mathbf{P}\left[S_1 \neq P_3 \cap S_2 = P_3\right]}{\mathbf{P}\left[S_2 = P_3\right]} = \frac{1/10}{1/10} = 1.$$

This makes sense to us since the first selection could not have been the third physician given that the second selection was.

## Enumeration

Another style of computing probabilities is called enumeration, or event counting. In the previous example, we counted the number of ways one can select physicians randomly from a collection. Returning to our relative frequency argument, in order to compute, $\mathbf{P}\left[S_1 \neq P_2 \cap S_2 = P_3\right]$, we need to consider for the denominator the total number of ways to select two physicians, and for the numerator, the number of ways to select physicians such that the first selection is not physician 3 but the second one is.

Using enumeration and first focusing on the denominator, we ask how many possible choices are there for the two physicians. There are 10 possible choices for the first selection, and then 9 possible choices for the second. The total number of sequences of two physicians is (10)(9) = 90. We can use what we know about factorials to go further.

For the numerator, there are 9 possible choices for the first physician and only 1 for the second giving us (9)(1) = 9 possible sequences. Thus we can now write

$$\mathbf{P}\left[S_1 \neq P_2 \cap S_2 = P_3\right] = \frac{(9)(1)}{(10)(9)} = \frac{9}{90} = \frac{1}{10},$$

confirming what we previously found. Thus counting events produced the same solution as manipulating these probabilities.

## Permutations

In the direct computation above, we computed the number of different possible sequences of events. For example, the denominator reflects that there were ninety sequences of physicians taken two at a time. They are

$P_1P_2$, $P_1P_3$, $P_1P_4$, $P_1P_5$, $P_1P_6$, $P_1P_7$, $P_1P_8$, $P_1P_9$, $P_1P_{10}$,

$P_2P_1$, $P_2P_3$, $P_2P_4$, $P_2P_5$, $P_2P_6$, $P_2P_7$, $P_2P_8$, $P_2P_9$, $P_2P_{10}$,
$P_3P_1$, $P_3P_2$, $P_3P_4$, $P_3P_5$, $P_3P_6$, $P_3P_7$, $P_3P_8$, $P_3P_9$, $P_3P_{10}$,
$P_4P_1$, $P_4P_2$, $P_4P_3$, $P_4P_5$, $P_5P_6$, $P_4P_7$, $P_4P_8$, $P_4P_9$, $P_4P_{10}$,
$P_5P_1$, $P_5P_2$, $P_5P_3$, $P_5P_4$, $P_5P_6$, $P_5P_7$, $P_5P_8$, $P_5P_9$, $P_5P_{10}$,
$P_6P_1$, $P_6P_2$, $P_6P_3$, $P_6P_4$, $P_6P_5$, $P_6P_7$, $P_6$P8, $P_6P_9$, $P_6P_{10}$,
$P_7P_1$, $P_7P_2$, $P_7P_3$, $P_7P_4$, $P_7P_5$, $P_7P_6$, $P_7P_8$, $P_7P_9$, $P_7P_{10}$,
$P_8P_1$, $P_8P_2$, $P_8P_3$, $P_8P_4$, $P_8P_5$, $P_8P_6$, $P_8P_7$, $P_8P_9$, $P_8P_{10}$,
$P_9P_1$, $P_9P_2$, $P_9P_3$, $P_9P_4$, $P_9P_5$, $P_9P_6$, $P_9P_7$, $P_9P_8$, $P_9P_{10}$,
$P_{10}P_1$, $P_{10}P_2$, $P_{10}P_3$, $P_{10}P_4$, $P_{10}P_5$, $P_{10}P_6$, $P_{10}P_7$, $P_{10}P_8$, $P_{10}P_9$

Essentially what we do is permute or rotate the positions of the physicians systematically, in order be sure that we incorporate them all. This rotation of ten physicians through two slots gave us 10 for the first slot and 9 for the second. A succinct way to write this uses factorial notation and can be written as

$$\frac{10!}{(10-2)!}=\frac{10!}{8!}=(10)(9)=90.$$

The denominator removes the counts of physicians filling the remaining eight slots.

Suppose we have a selection of 100 physicians from whom we wanted to select 4 at a time. With replacement, there would be $100^4$ or a one hundred million possibilities. However without replacement we have 100 for first position, 99 for the second, 98 for the third, and 97 for the fourth or (100)(99)(98)(97) = 94,109,400, possibilities. This is exactly

$$\frac{100!}{(100-4)!}=\frac{100!}{96!}.$$

In general, if we are permuting $n$ candidates or objects through $k$ possible slots without replacement, then the number of sequences is

$$\frac{n!}{(n-k)!}.$$

As an example, suppose we want to directly compute the probability that out of ten physicians we select three without replacement producing physician 3 for the first slot, and physician 9 for the third slot. To compute this specific probability for the second slot invoking the formal use of the law of total probability is calculable but complicated.

Returning to our definition of relative frequency, we must enumerate first all possibilities, then all possibilities related to the event of interest. The denominator is simply a permutation of 10 physicians rotated through 3 slots. The numerator requires one and only one selection for slot 1, 8 possible for slot 2 and 1 for slot 3. We therefore can write

$$\mathbf{P}\left[\{S_1=P_3\}\cap\{S_2\neq P_3,P_9\}\cap\{S_3=P_9\}\right]=\frac{(1)(8)(1)}{\left(\frac{10!}{7!}\right)}=\frac{8}{720}=\frac{1}{90}.$$

## Example: Antibody generation

One of the important functions of the immune system's B-cells is the generation of antibodies. These are short sequences of amino acids that, because of their unique chemical structure, attain a specific three dimensional configuration that allows it to "fit" on foreign bodies and begin to destroy them.

However, there are a seemingly innumerable number of different species of viruses, bacteria, rickettsia, fungi, and protozoa as well as other foreign organisms, each with its own chemical signature. Furthermore, viruses, as well as these other organisms, can mutate into strains of the same species. Each separate species/strain requires its own tailor-made antibody. How can one person's immune system make enough distinct antibodies to keep up with them all?

Let's assume that an antibody consists of a chain of 75 amino acids in a specific sequence. For our first consideration, let's assume that for each of the first ten positions, there are only three possible amino acid candidates with replacement. For the next fifty positions in the antibody there are eight possible amino acids for each slot, and for the last fifteen, there is only one possible candidate for each position. We recognize that, with these restriction, we are working in a sampling with replacement environment, since the filling of one position does not change the number of possible amino acids that can fill another. How many possible antibody configurations are there?

We can compute that there are $3^{10}8^{50}1^{15} = 8.4 \text{ x } 10^{49}$ possibilities, a huge number. Even with these restrictions on the antibody selection, the number of possible antibodies is immense.

To loosen this restriction, let's assume that there are twenty possible amino acids for each of the 75 positions, with replacement. In this circumstance, there are $20^{75}$ different possible antibodies, a number that is over $3.75 \text{ x } 10^{97}$, which is less than a thousand short of a google.*

The immune system has all of the flexibility it needs to cover the diversity of foreign invaders.

## Combinations

Returning to the simpler problem of permuting or rotating ten physicians through two slots, we see some interesting inclusions (noted in matching colors) below.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **$P_1P_2$,** | **$P_1P_3$,** | **$P_1P_4$,** | **$P_1P_5$,** | **$P_1P_6$,** | **$P_1P_7$,** | **$P_1P_8$,** | **$P_1P_9$,** | **$P_1P_{10}$,** |
| **$P_2P_1$**, | $P_2P_3$, | $P_2P_4$, | $P_2P_5$, | $P_2P_6$, | $P_2P_7$, | $P_2P_8$, | $P_2P_9$, | $P_2P_{10}$, |
| **$P_3P_1$**, | $P_3P_2$, | $P_3P_4$, | $P_3P_5$, | $P_3P_6$, | $P_3P_7$, | $P_3P_8$, | $P_3P_9$, | $P_3P_{10}$, |
| **$P_4P_1$**, | $P_4P_2$, | $P_4P_3$, | $P_4P_5$, | $P_5P_6$, | $P_4P_7$, | $P_4P_8$, | $P_4P_9$, | $P_4P_{10}$, |
| **$P_5P_1$**, | $P_5P_2$, | $P_5P_3$, | $P_5P_4$, | $P_5P_6$, | $P_5P_7$, | $P_5P_8$, | $P_5P_9$, | $P_5P_{10}$, |
| **$P_6P_1$**, | $P_6P_2$, | $P_6P_3$, | $P_6P_4$, | $P_6P_5$, | $P_6P_6$, | $P_6P_7$, | $P_6P_8$, | $P_6P_9$, |
| **$P_7P_1$**, | $P_7P_2$, | $P_7P_3$, | $P_7P_4$, | $P_7P_5$, | $P_7P_6$, | $P_7P_8$, | $P_7P_9$, | $P_7P_{10}$, |
| **$P_8P_1$**, | $P_8P_2$, | $P_8P_3$, | $P_8P_4$, | $P_8P_5$, | $P_8P_6$, | $P_8P_7$, | $P_8P_9$, | $P_8P_{10}$, |
| **$P_9P_1$**, | $P_9P_2$, | $P_9P_4$, | $P_9P_5$, | $P_9P_6$, | $P_9P_7$, | $P_9P_8$, | $P_9P_9$, | $P_9P_{10}$, |
| **$P_{10}P_1$**, | $P_{10}P_2$, | $P_{10}P_3$, | $P_{10}P_4$, | $P_{10}P_5$, | $P_{10}P_6$, | $P_{10}P_7$, | $P_{10}P_8$, | $P_{10}P_9$ |

Observe that in our ninety permutations, there are quite a few sequences (e.g., $P_1P_9$ and $P_9P_1$) that have the same elements, but just in a different order. This permuting is entirely appropriate in many circumstances. For example if the administrator is choosing physicians for a sequence of shifts, then $P_1P_9$ and $P_9P_1$ reflect different events.

However, suppose that the administrator is interested in choosing two physicians to staff the same shift. Then clearly $P_1P_9$ and $P_9P_1$ reflect the same assignment (i.e., both assigned to the same shift), and the sequence of assignments does not matter. Probabilists say, in this matter, that *order does not count.* Clearly, the number of possible sequences must be reduced, but by how much?

A quick way to see what the adjustment must be follows. Once we have selected a sequence of physicians, how can we identify the number of "duplicates". For the selection of a sequence of two physicians, there are two possible choices for the first slot, and once chosen, there is only one possible selection for the second producing (2)(1) duplicates. Thus to remove the duplicates, we simply divide the number of permutations by the number of duplicates, in this case reducing $\frac{10!}{8!}$ to $\frac{10!}{8!2!} = 45$.

* To help with perspective, there are $2 \text{ x } 10^{23}$ stars in the known universe, the sun weighs $2 \text{ x } 10^{27}$ tons, and there are $3 \text{ x } 10^{13}$ cells in the human body.

While there were ninety permutations, there are only 45 distinct sequences when order does not count. This final computation is called a *combination*, and we say the number of distinct sequences of $n$ objects when taken $k$ at a time is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The $k$! in the denominator is the correction necessary to reduce the duplication in the permutation since order does not count.

## Example: Ambulance arrival

Three companies each have ambulances. Company A has seven ambulances, Company B has twelve ambulances, and Company C has 9 ambulances. Assume only one ambulance is necessary to manage a single emergency, and that the ambulance manages only one accident at a time.

All ambulances are required one morning, responding to 28 accidents. A helicopter surveys ten accident sites randomly. What is the probability that of these ten accidents identified by the helicopter, three are covered by Company A, two by Company B, and five by Company C?

We will use enumeration to solve this problem. Let's term this event $A_3B_2C_5$. Then, using our relative frequency definition, we must compute how many possible ways there are to select ten accidents from 28, then compute the number of ways that $A_3B_2C_5$ can occur.

This is clearly sampling without replacement, since an ambulance, once tasked for an accident is removed from consideration for another accident response that morning. Also, since we are not concerned about sequences in which $A_3B_2C_5$ occur in different orders, we want to count distinct sequences in which order does not count.

The denominator of the probability is simply the combinatoric of 28 taken ten at a time, or $\binom{28}{10}$. We may proceed with the numerator, taking the events in small subcollections. We know that of the seven ambulances from Company A, 3 were "selected" as an accident responder. The number of ways this can happen is $\binom{7}{3}$. Similarly, we compute $\binom{12}{2}$ for Company B and $\binom{9}{5}$ for Company C. We can now conclude with

$$\mathbf{P}[A_3B_2C_5] = \frac{\binom{7}{3}\binom{12}{2}\binom{9}{5}}{\binom{28}{10}} = \frac{(35)(66)(126)}{13{,}123{,}110} = 0.023.$$

We can compute the probability that five accidents are covered by Company A and the remaining five by Company C as

$$\mathbf{P}\left[\mathrm{A}_5\mathrm{B}_0C_5\right] = \frac{\binom{7}{5}\binom{12}{0}\binom{9}{5}}{\binom{28}{10}} = \frac{(21)(1)(126)}{13{,}123{,}110} = 0.0002.$$

Suppose we wanted the probability that all ten accidents were covered by Company A and Company C. This is related to event $A_5B_0C_5$, but is much broader. We might write the desired event *AC*, and, set notation, as

$$AC = \bigcup_{i=0}^{10} A_i B_0 C_{10-i}.$$

Since these events are mutually exclusive, we can now write

$$\mathbf{P}\left[AC\right] = \sum_{i=0}^{10} \mathbf{P}\left[A_i B_0 C_{10-i}\right]$$

$$= \sum_{i=0}^{10} \frac{\binom{7}{i}\binom{12}{0}\binom{9}{10-i}}{\binom{28}{10}} = \frac{\sum_{i=0}^{10}\binom{7}{i}\binom{9}{10-i}}{\binom{28}{10}}.$$

## Sampling schemes in detail

The concepts of sampling with and without replacement are very straightforward and distinct from each other. However, there are circumstances where they are quite similar.

Consider the quality control operation for competing companies that make their own catheter based cell injector systems. Each of them conducts quality control operation on their units which are massed produced.

The first company assembles their catheters in batches of 25. From each batch, they select ten catheters randomly, subjecting each of the ten to a rigorous test. If they find one defective unit in the testing, then they discard the entire batch as defective. Assume that there are four defective catheters in the batch of 25. What is the probability that they will find one defective catheter?

This is sampling without replacement. In addition, since a collection of catheters with defectives has the same number of defective catheters regardless of their order, this is a problem of sampling where order does not matter. For the denominator of the probability, we need to know how many ways there are to collect ten catheters from 25, or simply $\binom{25}{10}$. The numerator consists of the number of sequences in which one defective catheter can be selected from or

$\binom{4}{1}$, and the number of sequences of four in which no defectives exists or $\binom{21}{9}$. The probability that we seek is $\frac{\binom{4}{1}\binom{21}{9}}{\binom{25}{10}} = \frac{(4)(293,930)}{3,268,760} = 0.36$

Company B is a much larger competitor, making batch shipments of 500. They select 10 at random and will reject the lot if 3 are defective. If there are 15 defects in the batch of 500, then the probability that the entire batch will be rejected is

$$\frac{\binom{15}{3}\binom{485}{7}}{\binom{500}{10}} = \frac{(455)(1.2 \text{ x } 10^{15})}{(2.46 \text{ x } 10^{20})} = 0.002.$$

The computation for this larger company is somewhat awkward because of the large quantities in the numerator and denominator of this probability calculation. However, for a very large batch size like this, while the assumption of sampling without replacement is true to the spirit of the experiment, the actual computation can be approximated by sampling with replacement.

Why is this?

The larger the population size, the less likely an individual will be selected twice just through chance. Thus sampling with replacement begins to resemble sampling without replacement (were the probably of reselection is zero).

Thus for this batch size of five hundred, we could try to approach this problem from a sampling with replacement perspective.

Sampling with replacement, we compute the probability that a defective is chosen is $\frac{15}{500} = 0.03$. The probability that one particular sequence of ten, has three defectives in specified positions is $(0.03)^3(0.97)^7 = 0.0000218$. But there are several different sequences, each with the same probability that will yield three defectives in ten. The number of these sequences is simply the number of ways to choose three objects from ten or $\binom{10}{3} = 120$. Thus the probability of having three defects in ten where the defects can occur in any position is $(120)(0.0000218) = 0.0026.$ * This is very close to the exact solution of 0.002. Thus as the batch size gets larger, the sampling without replacement model can be approximated by the sampling with replacement paradigm, a finding that can be formally demonstrated.

## The birthday problem

One of the most intriguing and counterintuitive examples in probability is the birthday problem. The question is how many people (selected randomly) must be present in a room to have a 50%

* This is actually a demonstration of the development of the binomial distribution.

chance that two of them will have the same birthday. Our intuition is some help, but commonly doesn't provide an answer that even approximates the actual solution.

Assume that individual birthdays are independent of each other, and that any particular day is just as likely to be a birthday as any other day. Let's also exclude consideration of leap year (although as we will see momentarily, it is easy to solve this for leap year as well.

Our intuition tells us that if we have 366 people in a room, we are guaranteed to have at least two with the same birthday. If one birthday event is equally likely as another, we might think that half of 366 or 183 individuals would be required. That is typically as far as we can get.

We will solve this problem by computing the number the probability of the complement, i.e., the probability that there are no common birthdays.

The denominator for this probability is simply the number of possible birthdays two people can have. The first can "choose" from 365, as can the second. Thus, the denominator is $365^2$. This is sampling with replacement paradigm.

The numerator requires a different computation. The first subject can have 365 possible birthdays. Given the first subject has "selected" their birthday, the second subject has from among 364 birthdays. Thus, the probability that two individuals do not have the same birthday is

$$1-\frac{(365)(364)}{365^2}= 1-0.997=0.003.$$

This solution makes sense to us, because we expect the likelihood that to are chosen at random is small. If we were to generalize, we would find that the probability *n* subjects selected at random have the same birthday is

$$1 \ - \ \frac{\frac{365!}{(365-n)!}}{365^n}.$$

Note, that here order does count, so we have a permutation and not a combination in the numerator of the main quotient. We can now graph this probability as a function of *n* the number of individuals in the room.

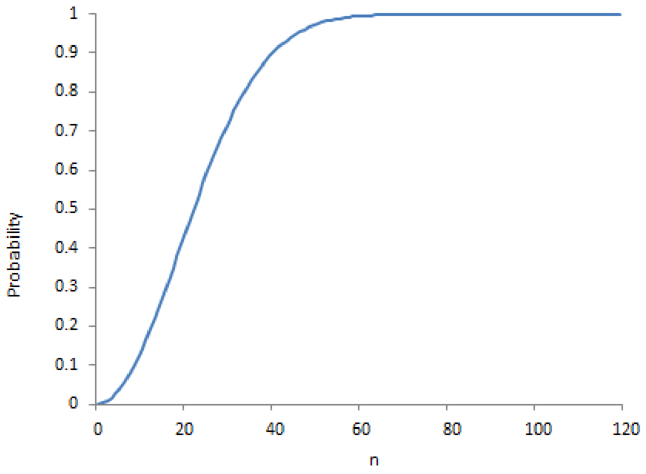


**Figure 1.** Probability of *n* people in a room having the same birthday

Most people, when exposed to this question would not guess that it only takes 23 people in a room to have a 50% chance that they have the same birthday, or that having only 120 people in a room all but ensures that two will have the same birthday. This is because the numerator of the probability reflecting the number of ways people do not have the same birthday decreases rapidly.

A related problem is the probability that two people have the same birthday, or a day apart. Following the previous example we can compute the probability that two individuals do not meet this criteria as

$$1-\frac{(365)(362)}{365^2}=1-0.992=0.008.$$

Note the second individual has only 362 possibilities for their birthday since the day of, the day before, and the day after the first person's birthday must be eliminated. As we would expect, far fewer individuals are required in a room to have a birthday within a day of each other (Figure 2).

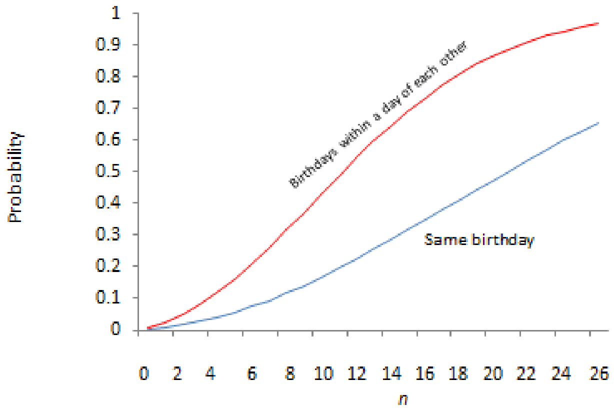


**Figure 2**. Probability of $n$ people in a room having the same birthday compared to number of people with a birthday within a day of each other

Additional variants of this problem are available. For example, suppose that we want the probability that out of $n$ subjects chosen at random, three individuals having the same birthday.[*] We will compute this directly, and not find the probability of the complement.

We take this problem in carefully sequenced steps. The denominator encompasses all possibilities and is simply $365^n$. For the numerator we first must choose three people from the collection of $n$ or $\binom{n}{3}$. Once we have these three individuals, we know that each has the same birthday. While there are 365 possibilities for this birthday, once selected, there is only one choice for a birth date for the second and third members of this collection. This leaves $(365)(1)(1)$. possibilities.

How we managed the remaining $n-3$ individuals opens up several possibilities. If we assume that they have unique birthdays, then the computation is straightforward. Assume, for

[*] This is from Judy Bettencourt and Rachel Vojvodic.

example that $n = 7$. Then three have the same birthday, leaving 4 to have different birthdays. There are $(364)(363)(362)(361) = \frac{364!}{(364-4)!}$ possible ways for this to happen. In general, if there are *n* subjects, then the number of possibilities is $\frac{364!}{(364-(n-3))!}$, and the final solution is

$$\mathbf{P}\left[3 \text{ common birthdays among } n \text{ subjects}\right] = \frac{\binom{n}{3}(365)\frac{364!}{(364-(n-3))!}}{365^n}$$

We can use this approach to compute the probabilities of a variety of other scenarios. For example the probability that there is one triple (that is three subjects have the same birthday) and two separate doubles in *n* subjects is

$$\frac{\binom{n}{3}(365)\binom{n-2}{2}(364)\binom{n-4}{2}(363)\frac{362!}{(362-(n-7))!}}{365^n}$$

Next sections
Relationship between Sampling with and without replacement

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics

Tail Event Measure

# Relationship Between Sampling With and Without Replacement

We have seen that in a sampling scheme without replacement, in certain circumstances, particularly when the size of the population becomes large enough, it makes sense to think about sampling without replacement as sampling with replacement. This does not suggest that the underlying design mechanisms are ever the same, only that one probability formula yields a solution that is close enough to the other.

Since formulas based on sampling with replacement are easier to use than those for sampling without replacement, it is helpful that there are times when we can use the simple sampling with replacement computations in the sampling without replacement paradigm.

## Prerequisites

Limits and Continuous Functions
Counting Events

To demonstrate this relationship, consider the following situation. From a population of *N* subjects, we draw a smaller sample of size *n.* We know that in the larger population, there are $N_0$ Hispanics. What is the probability that the sample of size *n* contains *x* Hispanics.

In this case, we are sampling without replacement, and we must write the exact probability as

$$\frac{\binom{N_0}{x}\binom{N-N_0}{n-x}}{\binom{N}{n}} = \frac{\binom{N_0}{x}\binom{N_1}{n-x}}{\binom{N}{n}}$$

which as we have seen, is computed simply from counting the number of ways to select *x* Hispanics from $N_0$ Hispanics, and selecting *n* - *x* nonHispanics from the remaining $N_1 = N - N_0$ nonHispanics in the population, divided by the number of ways one can choose *n* individuals from a population of *N*.

We can rewrite the previous formulation as

$$\frac{\binom{N_0}{x}\binom{N_1}{n-x}}{\binom{N}{n}} = \frac{\dfrac{N_0!}{x!(N_0-x)!}\dfrac{N_1!}{\left(N_1-(n-x)!\right)(n-x)!}}{\dfrac{N!}{n!(N-n)!}}$$

$$= \frac{N_0!}{x!(N_0-x)!}\frac{N_1!}{\left(N_1-(n-x)!\right)(n-x)!}\frac{n!(N-n)!}{N!}$$

A simple rearrangement of terms reveals that

$$\frac{\binom{N_0}{x}\binom{N_1}{n-x}}{\binom{N}{n}} = \frac{n!}{x!(n-x)!}\frac{N_0!}{(N_0-x)!}\frac{N_1!}{\left(N_1-(n-x)!\right)!}\frac{(N-n)!}{N!}$$

Expanding the terms $\frac{N_0!}{(N_0-x)!}$ and $\frac{(N-n)!}{N!}$ term by term, we may rewrite the expression above as

$$\frac{n!}{x!(n-x)!}Q$$

where

$$Q = \frac{N_0(N_0-1)(N_0-2)(N_0-3)...(N_0-x+1)}{N(N-1)(N-2)(N-3)...(N-n+1)}$$
$$\bullet N_1(N_1-1)(N_1-2)(N_1-3)...(N_1-n+x+1)$$

Since $0 \le x \le n,$ we can write $n = x+n-x,$ write the denominator as

$N(N-1)(N-2)(N-3)...(N-x+1-n+x)$.

This allows us to match each of the terms in the numerator with a term from the denominator. Thus we can write $Q$ as

$$\frac{N_0(N_0-1)(N_0-2)(N_0-3)...(N_0-x+1)\bullet N_1(N_1-1)(N_1-2)(N_1-3)...(N_1-n+x+1)}{N(N-1)(N-2)(N-3)...(N-x+1-n+x)}$$

Which can be expressed as

$$=\left[\frac{N_0\left(N_0-1\right)\left(N_0-2\right)\left(N_0-3\right)...\left(N_0-x+1\right)}{N\left(N-1\right)\left(N-2\right)\left(N-3\right)...\left(N-x+1\right)}\right]$$
$$\bullet\left[\frac{N_1\left(N_1-1\right)\left(N_1-2\right)\left(N_1-3\right)...\left(N_1-n+x+1\right)}{\left(N-x\right)\left(N-x-1\right)\left(N-x-2\right)\left(N-x-2.\right)..\left(N-n+x+1\right)}\right].$$

Which we can express as $Q_1 \bullet Q_2$

Examining $Q_1$ reveals

$$Q_1=\frac{N_0\left(N_0-1\right)\left(N_0-2\right)\left(N_0-3\right)...\left(N_0-x+1\right)}{N\left(N-1\right)\left(N-2\right)\left(N-3\right)...\left(N-x+1\right)}$$
$$=\left(\frac{N_0}{N}\right)\left(\frac{N_0-1}{N-1}\right)\left(\frac{N_0-1}{N-2}\right)\left(\frac{N_0-1}{N-3}\right)...\left(\frac{N_0-x+1}{N-x+1}\right).$$

The last line contains $x$ quotients. If we allow both $N_0$ and $N$ to get large (although $N_0$ must always be smaller than $N$) we will see that each of these quotients is dominated by $\frac{N_0}{N}$. Since we have $x$ such terms we may approximate $Q_1$ by $\left(\frac{N_0}{N}\right)^x$. We proceed analogously for $Q_2$ to write

$$Q_2=\frac{N_1\left(N_1-1\right)\left(N_1-2\right)\left(N_1-3\right)...\left(N_1-n+x+1\right)}{\left(N-x\right)\left(N-x-1\right)\left(N-x-2\right)\left(N-x-3\right)..\left(N-n+x+1\right)}$$
$$=\left(\frac{N_1}{N-x}\right)\left(\frac{N_1-1}{N-x-1}\right)\left(\frac{N_1-2}{N-x-2}\right)\left(\frac{N_1-3}{N-x-3}\right)...\left(\frac{N_1-n+x+1}{N-n+x+1}\right)$$

Here, there are n-x terms and again, allowing $N_1$ and $N$ to get large, such that $N_1 < N$, we find we can approximate each of these terms by $\frac{N_1}{N}$, we can approximate $Q_2$ by $\left(\frac{N_1}{N}\right)^{n-x}$. We can now write $\mathbf{P}[x]$ as

$$\mathbf{P}[x]\approx\binom{n}{x}\left(\frac{N_0}{N}\right)^x\left(\frac{N_1}{N}\right)^{n-x} = \binom{n}{x}\left(\frac{N_0}{N}\right)^x\left(\frac{N-N_0}{N}\right)^{n-x}$$
$$= \binom{n}{x}\left(\frac{N_0}{N}\right)^x\left(1-\frac{N_0}{N}\right)^{n-x}$$
$$=\binom{n}{x}p^x\left(1-p\right)^{n-x}.$$

This formula computes the probability of $x$ Bernoulli events in $n$ trials multiplied by the number of distinct (i.e., order does not count) ways to distribute x events in n trials. This is a sampling with replacement paradigm. So for large $N$ and $N_0$, we can compute the probability of an event that is based on sampling without replacement as though it is sampling with replacement. The final expression is the probability formula for the binomial distribution.

Counting Events

Bernoulli Distribution – In Depth Discussion

Advanced Binomial Distribution

Multinomial Distribution

Hypergeometric Measure

Geometric and Negative binomial measures

General Poisson Process

Survival Measure: Exponential, Gamma, and Related

Cauchy, Laplace, and Double Exponential

Continuous Probability Measure

Moment and Probability Generating Functions

Variable Transformations

Uniform and Beta Measure

Normal Measure

Compounding

F and T Measure

Ordering Random Variables

Asymptotics

Tail Event Measure

# An Introduction to the Concept of Measure[*]

Prerequisite sections
Elementary Set Theory

## Brief Background

Measure theory has historically been one of the most challenging topics in mathematical analysis. First developed by Henri Lebesgue and Andrey Kolmogorov in the early 20th century, it builds on set theory to produce some of the most profound findings in mathematics. Measure theory has provided a deeper understanding of probability, and has also fueled the development of the field of stochastic processes.[†]

Unfortunately, measure theoretic probability has classically been taught at such a high level, and with such a heavy mathematical preamble[‡] that its application to public health can be obscured.

Fortunately, measure theory can be demystified to demonstrate its applicability to probability, permitting us to smoothly incorporating it into our ongoing discussions concerning probability and public health.

## What is measure theory

Measure theory at its heart is the process of accumulating the value of sets. What we are "measuring" in measure theory is sets. We place "value" or "measure" on sets, and accumulate this measure.

Sometimes the rules of accumulation are complicated. Other times (e.g., simple counting) this aggregation or collection of value across sets is easy to follow. In any event, the guiding ideas are straightforward: 1)The particular "measure" or "value" must be well defined and follow a clear set of rules (this is the development and justification of the measure) and 2) the rules of set theory and set operations are used to guide the accumulation of measure across a wide variety of sets.

[*] This chapter is adapted from Moyé, L Weighing the Evidence: Duality, Set, and Measure Theory in Clinical Research. Trafford Inc. 2020.

[†] Stochastic processes are events whose occurrence is governed by probability and time e.g., the number of arrivals at an emergency department in the next hour.

[‡] Commonly two years of calculus and at least a semester of formal mathematical analysis.

Measure theory is a large field; probability is one of its subfields. Developed by the Russian probabilist Kolmogorov, it is based on the concept of set functions. We have already seen how helpful it is to observe that computing a probability is the same as computing the value of a function that operates on a set. Ultimately, we will see that the use of measure theory will expand our probability computing ability to rich collections of events that we might have first thought were too complicated to manipulate and measure.

The principal advantages of measure theory to the practicing probabilist is the ability to accumulate probability over complicated and sometimes dissimilar events. There are theoretical advantages as well, but to the applied probabilist, it is the ability to work with new types of "measurable" functions on new and heretofore unanticipated events that makes measure theory worth learning. In fact without it, we lose the ability to apply novel functions to nature's problems that are beyond the ability of standard probability theory to manage.

## Accumulation

Measure theory focuses on the process of accumulation (or valuing) of sets This valuation involves the use of set functions.

Recall Table's 1 demographic description. If we want to compute the probability that a subject is a women, we can "accumulate" probability over both non-Hispanic women and Hispanic women. This accumulation process is the heart of measure theory. In the example of Table 1, it is straightforward.

In order to carry this aggregation of measure out, we must have the right collection of events, or σ-algebra. We must also have a measure that follows specific rules.

Thus, in order to measure a set $A$ in the σ-algebra, $\Sigma$, we may have to manipulate sets in $\Sigma$ through our standard set operations (unions, intersections, and complements) to create $A$. We then parallel the generation of $A$ through these set operations with a process (typically no more than addition or subtraction) that accumulates the measure of $A$. As we build up $A$ from other sets in $\Sigma$, we accumulate the measure of $A$ from the measure of these same other sets in $\Sigma$. Let's start with an easy example.

## Example: Music tracks

Many people now manage their songs (or tracks) digitally. Suppose an individual with several thousand tracks wishes to get a sense of the "value" of them. How could they do this?

One way to assess value is simply to count the number of tracks, increasing the count by one for each distinct track. This is simply and naively, "counting measure". The value of a particular track is simply its presence.

### *Intuitive Rules of Accumulation*

Does "counting measure" make sense as a value to be accumulated? From a mathematical point of view, the answer is yes. It also follows our intuition about accumulation. For example, no set can have negative measure. In addition, if you have no tracks, you have a count or "measure" of zero. These two properties seem self-evident, but are a requirement of a measure or value system.

In addition, if you have two subsets of tracks and the first subset is completely contained in the second, then the "counting measure" of the first is less than that of the second. This is also a required property.

If you have five different playlists, then the measure of the five playlists is equal to the total number of tracks in the playlists minus the duplicate tracks, i.e., minus the measure of the intersections of the tracks across the five playlists. Accumulation makes sense to us with this "counting" measure.

So how do we "accumulate measure" in this music track example?

We use set operations to guide us. Define $\Omega$ as the set of all of our music tracks $\{T_1, T_2, T_3, \ldots\}$. Our σ-algebra, $\Sigma$, is the set of all subsets of these tracks (including the subsets representing our playlists.) Then we manipulate these sets and subsets using set operations to construct the final set, accumulating counting measure as we go along. We both "build up" our final set and also build up, or accumulate our measure.

For example the measure of the set of tracks $\{T_3, T_7\}$ is 2. We say $\mu(\{T_3, T_7\}) = 2$. It is also true that $\mu(\{T_7, T_{115}\}) = 2$ as well. Suppose that we want $\mu(S)$ where $S = \{T_3, T_7\} \cup \{T_7, T_{115}\}$ However we notice that there is overlap in the identities of the tracks. But we can also see that $S = \{T_3, T_7\} \cup \{T_7, T_{115}\} = \{T_3\} \cup \{T_{115}\} \cup \{T_7\}$ from our standard set operations. [*] Thus we see that $\mu(S)$ is three.

While this may seem the "long way around" to get the measure of three tracks, this formal process consists of the precise steps we take to isolate and identify these tracks, so that that we can measure the precise ones of interest. We start from other sets to "build up" $S$, building up $\mu(S)$ in the process.

### *Additional Music track "Measures"*

A great flexibility of measure theory is that there is a great variety of measures that are and can be developed. For example one can consider defining a music track measure as the play count, i.e., the number of times the track has been played. If it meets our criteria of measure described as rules of accumulation, then it is an admissible measure. Its use will produce a different measure of the music collection.

A third "measure" would be duration of the track in time. Here one simply accumulates or sums the length of each track in the end coming to a time ( e.g., 17.7 months).

Each of these measures (total tracks, number of plays, and total time), is legitimate. However, each measure is different, because it emphasizes a different characteristic of the track. Also, the characteristic of the track must be available to be measured. And of course each "measure" has to consistent with the rules of accumulation

Thus, from this perspective, "measuring sets" is not new. It, in fact, is all around us. Our task is to apply this process directly to sets in public health.

### *Example: Clinical Research Reimbursement*

As another example, suppose you are in charge of making payments in a clinical study that will follow subjects over a period of time.[†] The clinical centers that recruit these subjects will of course incur substantial cost as they see and examine each patient, draw blood work, and obtain modern (and expensive) imaging over the course of the study. Assume that each study patient will be seen six times over the course of the research. How should the coordinating center reimburse the centers for their costs?

One idea (Plan A) reimburses the centers directly in accordance with the way that costs were incurred; in this case making equal payments of 16.7% of the total cost on each of the entire six months so that by the conclusion of the study, the clinics have received 100% of the payments.

[*] Recall that for sets $A$ and $B$ that $A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$.

[†] This is based on example provided by Rachel W.Vojvodic, M.P.H.

However, Plan B assigns dollars differently. It provides 60% of the cost divided equally over the first two visits, then 10% during the remaining four visits. This front loading of cost permits the clinical center to expand their research team early in the study to provide more accurate and timely patient throughput and data transmission.

Alternatively, Plan C backloads costs, paying 10% of the total cost for each of the first 5 visits, then 50% for the last visit. This adds an important financial incentive to the scientific motivation of clinical centers to follow study subjects to the end of the research.

Each of these plans provides total cost disbursement at the conclusion of the study; however the distribution of costs is different (Figure 1).

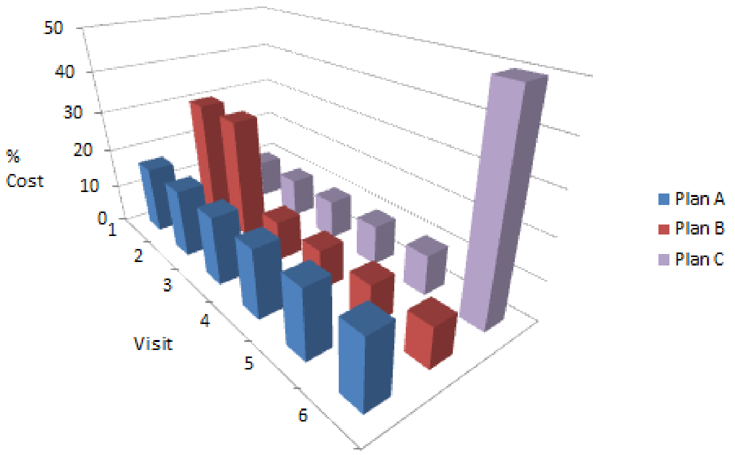


Figure 1. Three different plans for cost reimbursements over time in a clinical study.

Suppose that we want to compute the cost reimbursement for the first three visits of each plan. Plan A reimburses approximately 50% of the total patient care cost during this period. Plan B reimburses 60% during that period of time, while Plan C reimburses 30%. Now, define the cost for a visit as the measure of that visit. The costs or "measure" of each of these plans during the first three visits is different. The total "measure" over the six visits is the same or 100%.

If we characterize the visits as $V_1, V_2, V_3, V_4, V_5, V_6$, then we can go even farther and define measure $\mu$ as the reimbursed cost for each visit. So the cost for visit $V_1$ as $\mu(V_1)$, and the cost or measure of visit 1 under plan A is $\mu_A(V_1) = 16.7$. Then the system of cost or measure of both $V_1$ and $V_2$ is $\mu_A(V_1 \cup V_2) = \mu_A(V_1) + \mu_A(V_2) = 16.7 + 16.7 = 33.4$. We can also see that $\mu_B(V_1) > \mu_A(V_1) > \mu_C(V_1)$, and $\mu_C(V_6) > \mu_A(V_6) > \mu_B(V_6)$. In fact there are many types of relationships between these measures that are induced by the system of payments. However, each comports with our intuitive rules of accumulation, and is admissible as a measure.

Developing these systems (which appears to be quite like operating with sets) is at the center of our use of measure theory.

## Biomarkers and phenotypes

As a final example, consider the work of an investigator working on a biomarker that will hopefully predict the occurrence of heart failure (HF) in a patient for their next year. The levels of this biomarker are reported as being between zero and one. The risk of heart failure is related to the interval in which the biomarker is located.

However, its predictive value is strange. Specifically, the risk is related to each decile of heart failure, where the height of the risk alternates with the decile. The risk of heart failure is low for patients with biomarker levels in (0.40, 0.50] but high for patients in the level [0.05, 0.10] etc. In addition, if there are three contiguous intervals, then the risk is low in at least one of them and high in at least one of them. Finally, these relationships are only true for patients who have a particular phenotype. Patients with other phenotypes have an entirely different set of relationships between HF and biomarkers.

The use of such a tool (for example, to compute the risk to a population of subjects with a distribution of phenotypes) would likely lead to frustration because in health care we can easily manage monotonic risk or U/J shaped risk.

Risk which oscillates are outside of our experience and therefore confounds us.

However, all that this biomarker is really doing is assigning different heart attack risks to different regions of the $[0,1]$ interval, which is what set functions and measure theory are tailored to do. We can create a measure that measures, or integrates or accumulates risk over these intervals.*

## Symbols

In order to help us, we will need more notation. Typically, the symbology used in measure theory is $\mu(\;)$. Here, $\mu$ refers to the measure, (e.g., counting measure from the music track example) and the argument refers to the set being measured. An example is $\mu(A_1 \cap A_2 \cap A_3)$ denotes the measure of the collection of sets or objects common to the sets $A_1$, $A_2$, and $A_3$. It says nothing about how we actually take the measure, but instead only signals our intent to carry out the measuring procedure.

The integral sign serves the same purpose. In fact, we will use the integral sign and the measure denotation $\mu(\;)$ interchangeably. Just like the $\mu(\;)$notation, $\int du$ will announce our intent to measure a collection of sets with respect to the measure $\mu$ . These collections may be discrete objects, intervals on the real line, volumes of space, or combinations of these different metrics.

Again, *how* we measure them is not conveyed by either notation. *That* we will measure them is.

This can be a little disconcerting to an enthusiast of integral calculus with its collection of formulas denoting how to "integrate", such as $\int \cos(x)dx = \sin(x)$, or $\int_t^\infty \lambda e^{-\lambda t} dt = e^{-\lambda t}$.

However, it is useful at this point to take a step back and see just what this integration is doing.

The classic way for us to view these standard integration rules is that we are accumulating "area under the curve" and of course many times that is not a wrong perspective. However, another approach is to say that we are taking the "measure" of intervals of the real line. From this perspective, each of these formulas provides a different "measure" of the same interval. For example, consider an interval $(a,b)$ on the positive real line. Then we know

* This also begs the question, are we missing useful biomarkers because they don't conform with the standard monotonic or U/J risk predictive behavior to which we are accustomed.

$$\int_a^b dx = b - a$$

$$\int_a^b \cos(x)dx \; = \sin(b) - \sin(a)$$

$$\int_a^b \lambda e^{-\lambda x} dx \; = e^{-\lambda a} - e^{-\lambda b}$$

Each of these three integrals does something different with the same interval $(a,b)$, i.e., each "measures" the $(a,b)$ interval but uses a different measuring tool. For example $\int_a^b dx = b - a$ denotes that the measure of an interval as simply its length. This is known most famously as Lebesgue* measure.

However, the other two definite intervals demonstrate that there are additional ways to measure the same interval, each providing a different answer. In fact there are uncountably many measuring tools (many of which you already know) that provide the means to measure intervals of real numbers.

Thus, when we are taking a definite integral, we are measuring the interval, and the integrand is the measuring tool.

From this perspective, it is easy to see that $\int_0^\infty \binom{n}{k} p^k (1-p)^{n-k} 1_{k=0,1,2,3...n} = 1 = \int_0^\infty \lambda e^{-\lambda t} dt.$ This is a measure theoretic way to convey that binomial measure and negative exponential measure have the same measure of the entire real line. Again, the naked $\int$ sign does not tell us how to take the measure; this is conveyed by the integrand or measuring tool.

From a measure theoretic perspective there is no theoretical difference between measuring the real line by counting a subset of whole numbers on the one hand and completing a computation involving the length of the interval as the other. From the measure theory perspective, the only difference is the measuring tool.

### Goal of our approach

Measure theory as it is typically taught spends considerable time on the development of the underlying thought process, with very little time spent on the mechanics of actually how one uses it. It one wants to jump ahead into the details, a fair elaboration is provided here.

While it is useful to cover both, we will work not to get stuck on the deep mathematical elaborations, but will instead work to understand enough about the theory to see how it is applied. Some calculus will be involved from time to time and this will require a review of the indicated sections.

Our goal will be to incorporate measure theoretic concepts in our probability thought process so deeply that the measure theoretic tools become second nature, and we use them fluidly, almost without thinking (like counting music tracks).

Sequences of Sets
Sequences of Functions

* Pronounced LaBāg

Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Monotone Convergence Theorem
Some Classic Measure Theory Results
Asymptotics
Tail Event Measure

# Sequences of Functions

Prerequisites  
Indicator functions  
Pointwise and Uniform Convergence

In order to appreciate sum of the underpinnings of measure theory, we need to have an understanding of the concept of a sequence of functions, and from their what the limit of a sequence of functions can look like. This development will parallel our work on sequences of sets.

A sequence of functions, like a sequences of sets, is a concept that we must master if we are to actually wield measure theoretic constructs with some facility. In measure theory we build the support for functions that are measurable, or integrable by using sequences of simple functions. Therefore in order to understand the support for a type of measure or integral, we need to understand the behavior of these function sequences.

As we did with sets, we will approach these concepts beginning with simple examples, building up our intuition, and then summarize our findings.

## Building a function sequence

We are quite comfortable with functions e.g., $f(x)=3x^2-7$, and are not intimidated by this not being constant, but instead being a function of the variable $x$. However, how do we manage a function of "two variables" e.g., $f_n(x)=\frac{x^2}{n}$ for $n$ = 1,2,3… ? Here the domain of the function with regard to $x$ is the real numbers. However, $n$ (indexed by the non-negative integers) sets up an infinite collection of functions. For any $x$, the function sequence is

$$x^2, \frac{x^2}{3}, \frac{x^2}{4}, \frac{x^2}{5}, \frac{x^2}{6}, .....$$

These types of functions are commonly used in statistics. For example, the sequence of functions is $\bar{X}_1, \bar{X}_2, \bar{X}_3, ...\bar{X}_n, ...$ is the collection of means of independent samples from the same population, each time increasing the number of observations in the sample by one.

Another example of the use of sequences of functions is the study of the property of convergence, In that case we have not just a sequence of numbers, but a sequence of functions of $x$ that converge. This use of a function permitted us to differentiate pointwise versus uniform convergence.

The reason for this differentiation you will recall is that for pointwise convergence, the rate of convergence was related to the value of $x$; therefore we cannot be guaranteed that for all values of $n > N$ the function will be within some small distance $\varepsilon$ of the limiting value for all $\boldsymbol{x}$. It is uniform convergence that provided us this value that did not involve $x$.

Another example of a sequence of functions would be $f_n(x) = \frac{x^2}{n}1_{x=1}$. This combination of a function and an indicator function provides important flexibility as we construct sequences of functions that meet our needs. For example, the value of $f_n(x)$ produces the sequence $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, ....$ which clearly converges to zero. We can also convert this same function into $f_n(x) = \frac{x^2}{n}1_{x\in A}$. This is a sequence of functions that takes on the value $\frac{x^2}{n}$ for every $x \subset A$ and vanishes everywhere else. An example would be $f_n(x) = \frac{x^2}{n}1_{x\in\left[\frac{1}{n+1},\frac{1}{n}\right)}$, producing the functions

$x^2 1_{x\in\left[\frac{1}{2},1\right)}, \frac{x^2}{2}1_{x\in\left[\frac{1}{3},\frac{1}{2}\right)}, \frac{x^2}{3}1_{x\in\left[\frac{1}{4},\frac{1}{3}\right)}, \frac{x^2}{4}1_{x\in\left[\frac{1}{5},\frac{1}{4}\right)}, \dots$ .

We certainly have tremendous flexibility building these sequences of functions. The purpose of this flexibility is to be able to build up a from a sequence of simpler functions a more complicated function in which we have a central interest. While the final function may at first blush be difficult to work with, the simpler functions are not, and properties of sequences of these simpler functions can under some circumstances be absorbed into the final more complicated function.

## Step functions

One of the most important constructs for sequences of functions is the step function or indicator function $\alpha 1_{x\subset A}$. It is easy to think of this function residing on the real number line, such as the function $f(x) = \frac{1}{2}1_{x\in[0,1]}$ that takes on the value $\frac{1}{2}$ on the $[0,1]$.

However, the set A may not be an interval. For example, consider the function that is the value 1 for every natural number, $\frac{1}{2}$ for every $x \in \left(n, n+\frac{1}{2}\right)$, and $\frac{1}{3}$ for every $x \in \left(n+\frac{1}{2}, n+1\right)$. We write this function as

$$f_n(x) = 1_{x\subset\aleph} + \frac{1}{2}1_{x\subset\left(n,n+\frac{1}{2}\right)} + \frac{1}{3}1_{x\subset\left(n+\frac{1}{2},n+1\right)}.$$

Here the process of aggregating sets which take on the same value for $f(x)$ is more complicated.

## Example: Community viral testing

A consequence of this approach is that, although one cannot directly write the final limiting form of a function, one might be able to write a simpler function whose limit is the final form.

For example, suppose we wanted to assess the extent of a virus' spread through a community. There are parametric epidemiologic mathematics available to study this phenomenon e.g., the contagion model; however these models require the estimates of parameters that may not be available, or, if available, may not be accurate.

Consider an alternative approach. Suppose that there is a set of individuals in the community who are positive for the virus at time $t$. Let that set be $A_t$. There is also a set of individuals $B_t$ who at time $t$, while not positive for the virus now, are positive for antibodies, reflecting a prior infection. Finally, consider the set $C_t$. who are the individuals positive for both antibodies and virus. If $\omega_i(t)$ reflects the status of the $i^{\text{th}}$ individual in the community, then we create the function

$$C_t(n)=\sum_{i=1}^{n}1_{\omega_i(t)\subset A_t}+1_{\omega_i(t)\subset B_t}+1_{\omega_i(t)\subset C_t}$$

$C_t(n)$ is the cumulative exposure in the sample at time $t$. The $\lim_{n\to\infty}C_t(n)$ is the cumulative exposure in the population at time $t$.

While this function can be computed for each time point *t,* a plot of $C_t(n)$ overt reveals not just the increase in $C_t(n)$ over time, but plots of each of its three components demonstrate to what degree past and present infections are driving the cumulative counts over time, without the assumptions required to formulate a parametric model. We will see later that these simple functions can be integrated to provide a wealth of new information.

In this case, we have created a path to capture a complex process, beginning with a simple function. We will see that this permits us to not just build one, but many simple functions, using them as a starting point for the implementation of Lebesgue integration theory . We will see that commonly, the sequence of these functions is all that is required, and not the mathematical formulation of their limit.

■

## Infs and sups with function sequences

Now consider two function, $f(x)=x^2 1_{x\in[-1,2]}$ and $g(x)=x^3 1_{x\in[-1,2]}$. These familiar functions are easy to visualize (Figure 1). Now, based on these two functions, let's examine what the minimum of these functions looks like as a function of $x$. .

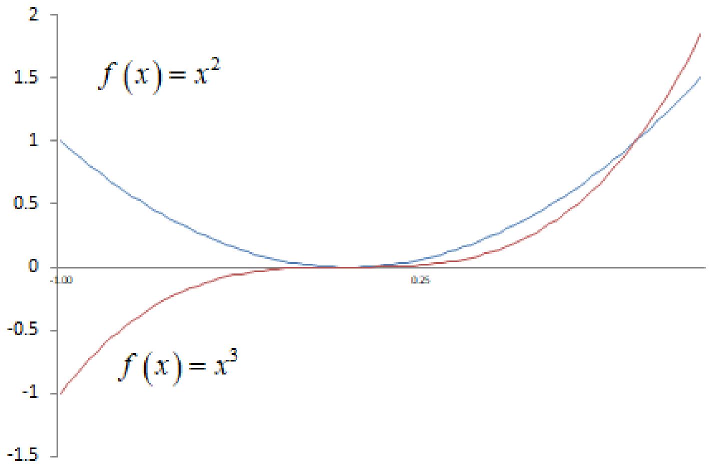


Figure 1. Comparison of two functions for the comparison of the infimum

And examination of these two common functions reveals that the minimum of the two functions changes over the region. However, we also notice that the minimum value is delivered by different functions over the entire $\left[-1, 2\right]$ range (Figure 2).

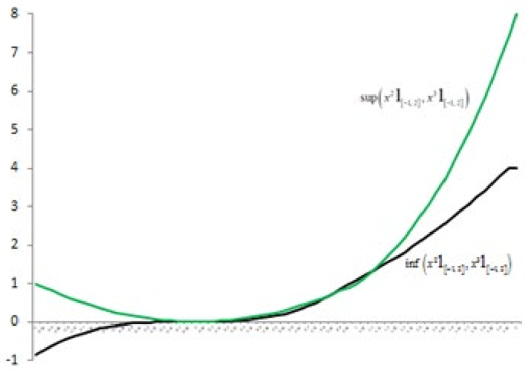


Figure 2. Examination of $\inf\left(x^2\mathbf{1}_{[-1,2]}, x^3\mathbf{1}_{[-1,2]}\right)$ and $\sup\left(x^2\mathbf{1}_{[-1,2]}, x^3\mathbf{1}_{[-1,2]}\right)$

For the smaller values of $x$ within this interval, $g\left(x\right)$ provides the minimum value of the two functions. This is followed by a region where the two values are essentially the same value, followed by a region farther to the right where the minimum value is provided not by $g\left(x\right)$ but by $f\left(x\right)$.

Thus, we can think of the minimum or infimum of these two functions as itself a function of $x$. A similar development reveals that the $\sup\left(f(x), g(x)\right)$ is also a function of $x$. In order to make this observation explicit we will write $I(x) = \inf\left(f(x), g(x)\right)$ and $S(x) = \sup\left(f(x), g(x)\right)$. Sometimes $I(x) = \inf\left(f(x), g(x)\right)$ is written as $I(x) = f(x) \cap g(x)$. Here the intersection sign which we have seen with sets means the minimum, or more precisely and helpfully, the greatest lower bound. The union sign mean maximum or smallest upper bound.

Having already established the concept of an infinite sequence of functions $f_n(x),\ n = 1, 2, 3, \ldots,$ we can now write the infimum and supremum functions as

$$I(x) = \bigcap_{n=1}^{\infty} f_n(x), \text{ and } S(x) = \bigcup_{n=1}^{\infty} f_n(x).$$

With this we can write the greatest lower bound of the sequence of functions $f_n(x)$

$$\liminf_{n\to\infty} f_n(x) = \bigcup_{n=1}^{\infty} \bigcap_{m>n}^{\infty} f_m(x),$$

and

$$\limsup_{n\to\infty} f_n(x) = \bigcap_{n=1}^{\infty} \bigcup_{m>n}^{\infty} f_m(x).$$

Again, note that these are functions of $x$. Thus the liminf is the least upper bound of all of the greatest lower bounds (the maximum of all of the minimum values for all intent and purposes) for that value of $x$. Similarly for the limsup which is the minimum values of all of the maximums for that value of $x$.

Finally, if $\liminf_{n\to\infty} f_n(x) = \limsup_{n\to\infty} f_n(x),$ then $\lim_{n\to\infty} f_n\left(x\right)$ exists and is equal to the liminf and limsup.

As an example, let's define $f_n(x) = a$ for all $n$. We start this trivial example by acknowledging that with regard to functions, the simple $\cap$ sign just means to take minimums, and $\cup$ simply means to take maximums. Then in this case $\bigcap_{n=1}^{\infty} f_n(x)$ denotes the minimum value of each member of the sequence of functions, and we see that $\bigcap_{n=1}^{\infty} f_n(x) = a.$ Thus, we have that $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} f_n(x) = a$ as well and conclude that $\liminf_{n\to\infty} f_n(x) = a.$ Similarly, $\bigcup_{n=1}^{\infty} f_n(x) = a,$ $\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} f_n(x) = a,$ and $\limsup_{n\to\infty} f_n(x) = a.$

Consider another sequence of functions, this time $f_n(x) = x^n$ at the point $x = 0.$ Our intuition tells us that at $x = 0, f_n(0) = 0$ for all $n$. Consider the following analysis:

$$\bigcap_{m=2}^{\infty} x^n 1_{x=0} = \min\left(x^2 1_{x=0}, x^3 1_{x=0}, x^4 1_{x=0}, x^5 1_{x=0}, \ldots\right) = 0$$

$$\bigcap_{m=3}^{\infty} x^n 1_{x=0} = \min\left( x^3 1_{x=0}, x^4 1_{x=0}, x^5 1_{x=0}, \ldots\right) = 0$$

$$\bigcap_{m=4}^{\infty} x^n 1_{x=0} = \min\left( x^5 1_{x=0}, \ldots\right) = 0$$

The smallest upper bound of all of these zeros is of course zero. Thus

$$\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} f_n(x) = \bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} x^n 1_{x=0} = 0.$$

Now consider the function $f_n(x) = x^n 1_{x=0.1}$. We proceed to find

$$\bigcap_{m=2}^{\infty} x^m 1_{x=0} = \inf\left(x^2 1_{x=0.1}, x^3 1_{x=0.1}\,, x^4 1_{x=0.1}, \ldots\right) = \inf\left(0.01, 0.001, 0.0001,\ldots\right) = 0$$

$$\bigcap_{m=3}^{\infty} x^m 1_{x=0} = \inf\left( x^3 1_{x=0.1}\,, x^4 1_{x=0.1}, x^5 1_{x=0.1}, \ldots\right) = \inf\left( 0.001, 0.0001,\ldots\right) = 0$$

$$\bigcap_{m=4}^{\infty} x^m 1_{x=0} = \inf\left( x^4 1_{x=0.1}, x^5 1_{x=0.1}, \ldots\right) = \inf\left( 0.0001, 0.00001, \ldots\right) = 0$$

Remember here that the $\cap$ sign means not just the minimum, but the greatest lower bound. This is most helpful, because, while $x^n$ never actually reaches zero, its greatest lower bound is zero.

The $\cup$ or maximum value of all of these zeros is zero, so we have

$$\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} f_n(x) = \bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} x^n 1_{x=0.1} = 0.$$

Now, what is the limsup? We write

$$\bigcup_{m=2}^{\infty} x^m 1_{x=0} = \max\left(x^2 1_{x=0.1}, x^3 1_{x=0.1}, x^4 1_{x=0.1}, x^5 1_{x=0.1}, \ldots\right)$$
$$= \max\left(0.01, 0.001, 0.0001, \ldots\right) = 0.01$$
$$\bigcup_{m=3}^{\infty} x^m 1_{x=0} = \max\left(x^3 1_{x=0.1}, x^4 1_{x=0.1}, x^5 1_{x=0.1}, x^6 1_{x=0.1}, \ldots\right)$$
$$= \max\left(0.001, 0.0001, \ldots\right) = 0.001$$
$$\bigcup_{m=4}^{\infty} x^m 1_{x=0} = \max\left(x^4 1_{x=0.1}, x^5 1_{x=0.1}, x^6 1_{x=0.1}, \ldots\right)$$
$$= \max\left(0.0001, 0.00001, \ldots\right) = 0.0001$$

Now we compute the minimum or greatest lower bound for this infinite set as we find $\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} f_n(x) = \bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} x^n 1_{x=0.1} = 0.$ With this preliminary work we can now write $\liminf_{n\to\infty} f_n(x) = \limsup_{n\to\infty} f_n(x) = \lim_{n\to\infty} f_n(x) = 0.$

All of this effort is a necessary preamble to understanding in what sense does $\lim_{n\to\infty}\int f_n(x) = \int \lim_{n\to\infty} f_n(x) = \int f(x)$? The ability to pass limits through integral signs is quite useful in advanced probability in particular and measure theory in general. When is this possible?

Supremums are also useful because we examine all simple functions that take on a value less than or equal to $f(x)$ then take its supremum we will have the integral of $f(x)$. This is covered in the discussion of the monotone convergence theorem.

Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

# Set Functions in Measure Theory

Our ultimate goal will be to develop what we will soon call measurable functions. Here we begin by describing their relatively uncomplicated building blocks; set, indicator, and simple functions. Ultimately, we will discuss their possible use in describing complex assessments in public health.

Prerequisite sections
An Introduction to the Concept of Measure
Elementary Set Theory
Sequences of Functions

So far, we have described measure in very basic terms as the process of accumulation, and the provided examples have demonstrated that sometimes different tools are used depending on what is being accumulated. In order to understand how to use this theory to good effect, we need to generate some new notation and tools.

We will assume that we have a sample space/σ–algebra complex denoted as $(\Omega, \Sigma)$ and that $\omega_i \subset \Sigma$. We begin with the concept of an indicator function $1_{A \subset B}$. This function takes the value one if $A \subset B$ and zero otherwise. Thus, an indicator function maps the result of a "set test" (i.e., either $A \subset B$ or not) to either the value 0 or 1.

We have seen indicator functions before, but there, their argument was part of the real line, which of course are wholly appropriate sets. Here we will work primarily with set functions whose domains are not of necessity subsets of the real line. We write $f(A) = 1_{A \subset B}$.

## Defining an elementary set function

The simplest set function operates on a single and specific member of the set, testing that particular set member against a specific condition and returning the value 0 or 1. We will denote this as one of the simplest set function as $e(\omega)$. Its argument or domain is a singleton element, $\omega_i$, of a set, and it maps this single set element to either 0 or 1; We write

$$e(\omega_i) = 1_{\omega_i \subset A}$$

which we interpret as

$$\begin{aligned} e(\omega_i) &= 1 \text{ for } \omega_i \subset A \\ &= 0 \text{ for } \omega_i \not\subset A. \end{aligned}$$

## Example: Demographic set function

As an example of the development of this uncomplicated set function using actual public health information, consider data that provides demographic information for subjects entered into a heart failure study (Table 1).

<<Table 1>>

Table 1 provides data by gender and ethnicity only. We can compute a large number of probabilities from data such as these. However, let's use this table to create an elementary function.

Define $\Omega$ as the space of all 210 individuals, and $\Sigma$, its σ-algebra. Let $\omega_i$ be the $i^{\text{th}}$ person of these 210 individuals in Table 1, (thus $\omega_i \subset \Sigma$ ) and define $X(\omega_i) = 1_{\omega_i \subset \text{Hispanic}}$. Here $X(\omega_i)$ is our function taking on the value of 0 or 1 depending on whether the $\omega_i^{th}$ individual is a member of the set of individuals with Hispanic ethnicity. Thus, every individual in Table 1 has a value connected to them, namely $X(\omega_i)$ depending on their ethnicity.

Now, define four mutually exclusive sets of **Ω**.

$A_{HM}$ the collection of Hispanic males
$A_{HF}$ the collection of Hispanic females
$A_{nHM}$ the collection of are non-Hispanic males
$A_{nHF}$ the collection of are non-Hispanic females

Clearly $A_{HM} \bigcup A_{HF} \cup A_{nHM} \bigcup A_{nHF} = \Omega$. Lets now let the set $B \subset \Sigma$ be an arbitrary subset of individuals. Can we map the function $U(B) = 1_{[B \subset A_{HM}]}$?

The answer is "Yes", but it may not provide the result of interest. The condition $B \subset A_{HM}$ is met if every member of our arbitrary set $B$ can be found in $A_{HM}$, that is every member of $B$ is an Hispanic male. The function $U(B)$ does not count the number of Hispanic males in $B$; It simply determines if all members are Hispanic males or not. It is mappable but something of a coarse tool.

If we wanted to use set functions to count the number of Hispanic males in set $B$, we would create the function $R(\omega_i) = 1_{[\omega_i \subset A_{AH}]}$ which would return the value of 0 or 1 based on the status of a single individual. We could then $\sum_{\omega_i \subset B} R(\omega_i)$ as returning the accumulated number of individuals in set $B$ who are Hispanic males.

So we have considerable freedom in setting up our set function depending on our goals.

## Combining Indicator Functions

As we saw in the previous section, we can sum indicator- set functions to compute for example, the total number subjects who are Hispanic males in set in an arbitrary set $B$. We can expand this concept to consider linear combinations of set functions.

For example, suppose we wanted an approximation of the impact of diabetes mellitus in a community, and we recognized that this morbidity impact varied by the ethnicity and gender. We could proceed as follows. Define four set functions.

$$R_{HM}(\omega_i) = 1_{[\omega_i \subset HM]} : R_{HF}(\omega_i) = 1_{[\omega_i \subset HF]}$$

$$R_{nHM}(\omega_i) = 1_{[\omega_i \subset nHM]} : R_{nHF}(\omega_i) = 1_{[\omega_i \subset nHF]}$$

These are four set functions that map an individual based on their sex and ethnicity. Now if $m_{HM}$ is the morbidity burden for Hispanic males, and we had the analogous quantities for the remaining three ethnic gender combination, then the quantity

$$M = \sum_{\omega_i} m_{HM} R_{HM}\left(\omega_i\right) + m_{HF} R_{HF}\left(\omega_i\right) + m_{nHM} R_{nHM}\left(\omega_i\right) + m_{nHF} R_{nHF}\left(\omega_i\right)$$

In this case, each member of the sample has their gender and ethnicity assessed, and then the appropriate morbidity weight is applied.
Such a linear combination of set function is called a simple function.

## Other uses of simple functions

Simple functions historically make important foundational contributions to measure theory because we can approximate measurable functions by them.

Consider a simple polynomial function e.g., $y = x^2$ on the $[0,1]$ real number line. This is a parametric, smooth, and well behaved function that is easy to work with. It is possible for us to approximate this function by creating a collection of indicator functions.

We would begin with a collection of disjoint intervals on $[0,1]$, $\{A_i\} = (a_i, b_i)$ for $i = 1,..,n.$ Let's also assume that these are adjacent, non-overlapping intervals of equal lengths. We might have a first approximation as $f\left(A_i\right) = \left(\frac{b_i + a_i}{2}\right)^2 1_{[a_i \le x \le b_i]}$ and define our approximation function as $\tilde{y}(x) = \sum_{i=1}^{n} f\left(A_i\right).$ The larger the value of $n,$ the smaller the distance between $a_i$ and $b_i$, the approximation of $y(x)$ by $\tilde{y}(x)$.

However, the value of simple functions in public health is not to approximate functions like $y = x^2$, but to approximate functions whose final form is not known. For example, there is no generally accepted parametric function that maps diabetes type II morbidity to ethnicity and gender. However, simple functions can be built up to approximate such a function, even though the final functional form is unknown and likely unknowable. This is the practical contribution of simple functions to public health.

## Simple functions in public health

Simple functions that are based on subintervals of the real number line and useful in understanding the mechanics of measure theory, but the public health applications will require us to broaden these simple functions to assess characteristics of the environment, or renal function.

Examples that we have discussed before would be function built to assess viral susceptibility Another example follows.

## Example: GI track homogeneity

As an example, consider the heterogeneity of the bacterial species that inhabit the gastrointestinal (GI) track. There are several thousand bacterial species that colonize the GI track. How could we construct an adequate representation and summary measure of the GI track's heterogeneity?

Divide the GI track into contiguous, non-overlapping square mm of surface area. Survey each square mm for the dominant bacterial species. Let $A_j$ be the total area of GI tissue where the dominant bacterial species is *j*. Define $\omega_i$ as the condition of the *i*th square mm of the GI track. Then $e(\omega_i) = 1_{\omega_i \subset A_j}$. This function is 1, if the dominant bacterial species in the $i^{\text{th}}$ square mm of the GI track is $j$ . We can proceed by defining $B_t = \sum_i W_t(\omega_i) = \sum_i \sum_j \alpha_j 1_{\omega_i(t) \subset A_j}$ which is the overall virulence status of the bacterial state of the GI track at time $t$ where $\alpha_j$ is an assessment of the threat of bacterial species $j$ to the host.

We can see that the manipulation of these combinations of set functions into simple functions gives us the ability to generate functions with substantial flexibility. We will see in later examples that we can determine the shape of the function that we want by using the resilience that comes from creating combinations of different set functions.

# Measurable Functions

## Set functions versus measurable functions

Measurable functions are simply set functions that have two additional traits.

The first and the easiest is that the function must map to a non-negative number. For example the function $h(A)$ equal to the number of males in set $A$ minus the number of females in set $A$ is a perfectly fine set function, but since it might be negatively valued, it is not a measurable function.

The second requirement of a measurable function is a concept that, while somewhat awkwardly described mathematically, is easy to understand.

It has to do with a property of the domain of the function.

Consider Table 1, from which we know the demography of the 210 individuals. We will assume that we have a sample space/σ–algebra complex denoted as $(\Omega, \Sigma)$ and that $\omega_i \subset \Sigma$, where $\Sigma$ contains all of the subsets of subjects in Table 1. Let's take a set of subjects $M$ from Table 1 and create the function $j(M)$ which is the mean creatinine value for the individuals in the set.

This is a fine set function, i.e., its maps a set to a number. It also meets the measurable function criterion of positivity. However, the trait that it measures (creatinine) is not available from Table 1, which provides only the gender and ethnicity of the participants. Therefore, we say that the function is non-measurable with respect to the contents of Table 1, since the function requires knowledge of a characteristic of these patients that is not available.

In measure theory parlance, we say that the function $j$ is not measurable since there are creatinines $c$, for which we cannot find preimages of $j$ (i.e., individuals in Table 1), $j^{-1}(c)$). One cannot determine creatinine values (only ethnicity and gender) from the elements of $\Omega$. Thus, $j^{-1}(c) \not\subset \Sigma$.

### *Measurability and inspection testing*

Another way to think of this second trait of measurable functions is that a function, to be measurable, must pass an inspection test. This inspection assesses whether the trait that the function requires is available.

Obviously, again using Table 1, functions that are based on ethnicity and gender are measurable. Other functions requiring other traits are not (and are nonmeasurable) since an inspection reveals that the individual characteristics on which they are based are not obtainable from Table 1.

Elementary functions can be measurable functions. For example, the function from Table 1, $X(\omega_i) = 1_{\omega_i \subset \text{Hispanic}}$ assigns a nonnegative value to a set (which happens to be a single participant from Table 1). Also, ethnicity can be identified so that the inspection (preimage) requirement is met (i.e., we can identify individuals in Table 1 whose ethnicity is known). In fact, even if there were no participants in Table 1 who were Hispanic, the function is still measurable – it just does not ever return the value one.

We can also easily see that simple functions constructed from measurable, elementary functions are measurable. As an example of why this is true, consider the example where $\Omega$ is a collection of individuals on whom demographics including age and race are available, and $\Sigma$ is the σ-algebra of these individuals. Now consider the function $k(\omega_i) = 1_{Age(\omega_i) \leq 45} + 1_{\omega_i \subset Caucasian}$. For any value of $k(\omega_i)$ one can find the age and race of individuals so that the function value can be assigned. The key to this is noting that since the elementary functions are measurable, so too is the simple function on which it is constructed.

The set function is the simplest type of measurable function. We will rely on it as heavily as Bernhard Riemann relied on the rectangle for the classic Riemann integral.

## Measurable spaces and functions

We are now ready to talk in a little more detail about the structure of sets on which measurable functions are based.

We begin with the collection of all possible sets of interest. This is the sample space, commonly signified as Ω. It is the collection of all possible events or outcomes, each denoted by ω. It can be all of the subjects that comprise Table 1, taking each subject one at a time as one example. In another example, it can be the set of real numbers on $[0,1]$.

From the sample space, we construct the σ-algebra, known as Σ. This is the expanse of sets generated from Ω using the set operations union, intersection, and complement. Both the null set and $\Omega$ themselves are members of Σ. The combination of $\Omega$ and $\Sigma$ – denoted by (Ω, Σ) – defines the measurable space.

With this structure, we now define a set function on that measurable space. This function will map sets in (Ω, Σ) to a number. This function $f$ is a measurable function if it meets the prior two conditions of non-negativity and mappable preimages, on (Ω, Σ) (i.e., the inspection criterion).

There is an uncountably many number of measurable functions on (Ω, Σ).[*] Although most are not helpful, we can be encouraged by the observation that we typically have tremendous freedom in defining the measurable function of interest $f$ to suit our interest.

Let's now explore some simple functions.

Next Sections
Simple Functions in Public Health
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Monotone Convergence Theorem
Some Classic Measure Theory Results
Asymptotics
Tail Event Measure

[*] For example, the set of indicator functions $\{f_r(\omega_i)\}$ indexed by $r$ where $f_r(\omega_i) = r1_{[\omega_i \subset A]}$ where $A \subset \Sigma$ and $r$ is a unique real number is a set containing uncountably many measurable elementary functions.

Advanced Probability

Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Simple Functions in Public Health

Prerequisites
Elementary Set Theory
Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Measurable Functions

## Constructing applied simple functions

It is easy to construct simple functions in general mathematics, and in fact that is what is commonly done in classical real analysis. However, how can we construct simple functions that would be helpful in public health? The following are examples of the building process for measurable functions in biology and ecology.

## Index of soil pollution

Soil pollution is typically caused by industrial activity, agricultural chemicals, improper disposal of waste, and/or radiation. Here, we are interested in building set functions that would permit us to estimate contaminants in a cubic mm of soil.*

We will begin by classifying soil contaminants into four categories 1) industrial, 2) agricultural, 3) waste (human and animal waste product), and 4) radioactive. Each of these sets we will signify as *I*, *A*, *W*, and *R*. The set *I* is the set of all soil that contains industrial contaminants. We will define the sets *A*, *W*, and *R* similarly. Let $\omega_i$ be the *i*th square millimeter of soil selected for chemical evaluations. Let's now define the indicator function $e_I(\omega_i) = 1_{\omega_i \subset I}$ that denotes membership in the soil samples containing industrial pollutants. Note that this is a measurable function. Define analogous functions for the other three classes of ground pollutants. Then define the spectrum of the *i*th soil sample, is the short subsequence

$$e_I(\omega_i), e_A(\omega_i), e_W(\omega_i), e_N(\omega_i)$$

i.e., a collection of zeros or ones, depending on the soil's contaminant components. For example 0,0,0,0 denotes no soil contamination, while the sequence 0,1,0,1, denotes animal and nuclear contamination. We can define a more refined elementary function where there are multiple contaminants e.g. $e_{IA}(\omega_i)$. Now define the simple function

$$f(\omega_i) = \alpha_I e_I(\omega_i) + \alpha_A e_A(\omega_i) + \alpha_W e_W(\omega_i) + \alpha_N e_N(\omega_i)$$

where the constants $\alpha_I, \alpha_A, \alpha_W, \alpha_N$ are the weighting factors necessary to permit the relative danger posed by the pollutants to be compared to one another. Then $f(\omega_i)$ reflects the pollution burden of that square millimeter of soil.

* This is certainly very small, but it permits us to reasonably rely on that volume having one and only type of polution

If we can further refine the class of pollutants (for example, industrial can be broken down to benzene compounds, nitrous oxides, sulfur oxides, etc., while nuclear waste can be further classified by element, e.g., uranium, thorium, etc. ) we can more explicitly define the burden function. Thus we might think of $f_n(\omega_i)$ as $f_n(\omega_i)=\sum_{j=1}^{n}\alpha_j 1_{\omega_i\subset A_j}$ that reflects a more refined elaboration of the burden of waste in the cubic millimeter of soil. Of course $f_n(\omega_i)$ becomes increasingly refined as *n* continues to increase. In a time of increasingly precise and detailed refinement, *n* can become quite large.

In addition, we have the freedom to build up sets $B_i$ from the $\omega_i$ elements. For example, we can build up $B_i$ to be a square foot of soil (comprised of 90,000 individual $\omega_i$ 's.) We could also define $B_i$ to the be the set of land comprising the banks of a meandering stream. In this circumstances, we write $f_n(B)=\sum_{\omega_i\subset B}\sum_{j=1}^{n}\alpha_j 1_{\omega_i\subset A_j}$.What provides flexibility here is the many different ways we can assemble the sets $B_i$ and the degree to which we can partition the pollutants all using simple, measurable functions.

## Building a disease burden function

We can also take a similar approach for a more general set function. Consider a patient at risk for a cardiovascular disease. Let $A_i(t)$ be the set of clinical events that the patient has experienced at this point in time *t*. Then define a collection of sets that contain classes of these events. For example let $H_t$ be the number of heart attacks at time *t*. Similarly, let $S_t, U_t,$ and $D_t$ be the sets of numbers of strokes, cardiovascular surgeries, or deaths on or before time *t* respectively.[*] Now define a simple function

$$f_t(\omega_i)=a_t 1_{\omega_i\subset H_t}+b_t 1_{\omega_{ii}\subset S_t}+c_t 1_{\omega_{ii}\subset U_t}+d_t 1_{\omega_{ii}\subset M_t}.$$

Where $a_t, b_t, c_t, d_t$ are non-negative constants. This function $f_t(w_i)$ comprises a weighted sum of the events that have occurred for the $i^{\text{th}}$ subject at time *t*. However, note that although each of the four components of the indicator function is a set function, it is a set function at a particular point in time. We can sum this over time to compute

$$g_T(\omega_i)=\sum_{t=0}^{T} f_t(\omega_i)=\sum_{t=0}^{T} a_t 1_{\omega_i\subset H_t}+\sum_{t=0}^{T} b_t 1_{\omega_i\subset S_t}$$
$$+\sum_{t=0}^{T} c_t 1_{\omega_i\subset U_t}+\sum_{t=0}^{T} d_t 1_{\omega_i\subset M_t}.$$

The sum is a collection of the accumulated burden of events up to time *T*. Also note that as in the previous example, this is a function of $w_i$. This is a complicated function and not parametric. However, by starting with set functions, we have built up a simple function with many components that through its flexibility (e.g., different function weights that are themselves a

[*] The elements of each of $H_t$, $S_t, U_t$, and $D_t$ are simply natural numbers reflecting the number of possible events that could have occurred at time *t*. For example $H_t=\{1,2,3,...\}$ simply reflecting the number of possible heart attacks that the patient may have sustained at time *t*. An individual experiences only one death of course.

function of time) permit us to coalesce the function's complexity. A useful question is what does this function look like when the increments of *t* are quite small (e.g., seconds) and *T* becomes very large. This begins to look like $\lim_{T \to \infty} \sum_{t=0}^{T} f_t(A_i)$, representing the lifetime burden of these events.

Consider that any attempt to compute this disease burden function parametrically is a challenge. The use of indicator functions, converting them to the limit of simple functions was much easier because we did not have to make assumptions about relationships between complicated variables; essentially, we just counted. We will later see that although we may not be able to identify this function parametrically, we will be able to integrate it.*

## Urine production

The continued normal function of the body relies on the intake of water, nutrients, and calories, and the removal of waste products from the body. The major organ systems involved in removing wastes are the skin and kidneys.†

The bulk of the kidney is made up of nephrons, which, like the alveolus in the lung is not a cell, but a functioning physiologic unit that itself is made up of different cell types. It is principally composed of a thick mesh of capillaries and tubules. The capillaries are specially adapted to identify and remove metabolic waste products and toxins. This filtrate is then collected into the tubules, which when they join with other tubules of the kidney produce the ureter which guides the flow of urine to the bladder. Normal kidneys have between 800,000 and 1,500,000 nephrons apiece.

Now let's define an $(\Omega, \Sigma)$ sample space, σ-algebra pair representing all of the nephrons in the kidney. We select one nephron $\omega_i \subset \Sigma$ and ask how well that nephron is functioning over time. Let's sample this nephron once per minute *t* and define $f_t(\omega_i)$ as the function that summarizes the performance of the $i^{\text{th}}$ nephron at time *t*. It would be difficult to characterize the function of this complicated unit parametrically since it must handle all waste products and toxins. What would be the parametric form? Log linear? Quadratic? Hyperbolic? Even trigonometric? Some combination of these?

However, what we can do is create a large collection of indicator functions that capture this physiologic function. Define a collection of sets $S_j$ one set for each of the known waste products and poisons. Set $S_{j,t}$ is the set of all nephrons that appropriately filter toxin *j* at time *t*. Then let $e_{j,t}(\omega_i) = 1_{\omega_i \subset S_{j,t}}$ characterize the ability of the $i^{\text{th}}$ nephron to filter the $j^{th}$ waste product of poison at time *t*. Each nephron's function can then be characterized at time point *t* by a spectrum of activity

$$e_{1,t}(\omega_i), e_{2,t}(\omega_i), e_{3,t}(\omega_i), e_{4,t}(\omega_i) \ldots\ldots e_{j,t}(\omega_i) \ldots\ldots ,$$

which like in the preceding environmental toxin example is a sequence of zeros and ones defining a spectrum of filtering functions. Now we define a simple function at time *t*, $f_{n,t}(\omega_i) = \sum_{j=1}^{n} \alpha_j e_{t,j}(\omega_i)$ where the $\alpha_j$ are scaling constants that permits the function $f_{n,t}(\omega_i)$ to be unitless. Then this function reflects the urine producing functionality of the nephron, defined not by the simple measure of urine output, but by the nephron's ability to filter each of the

* Drawing on the monotone convergence theorem.

† While liver heptocytes play a major role in the detoxification of harmful substances, these detoxified compounds are not directly excreted from the body by the liver but are transported to the kidney for final removal.

known molecules that it was designed to remove. We need only compute $f_T(\omega_i) = \sum_{t=0}^{T}\sum_{j=1}^{n}\alpha_j e_{t,j}(\omega_i)$ to capture the filtering experience of the single nephron $\omega_i$ in the $[0,T]$ time interval.

With this result in hand, we can compute the filtering experience of any collection of nephrons. Let $A_j \subset \Sigma$ be a set of nephrons. We have complete freedom in determining $A_j$. For example, $A_j$ could be all of the nephrons in the left kidney, or all cortical nephrons, or all nephrons that are in an area of disease. Then $f_T(A_j) = \sum_{\omega_i \subset A_j}\sum_{t=0}^{T}\sum_{j=1}^{n}\alpha_j e_{t,j}(\omega_i)$ would represent the functioning of that unit. In fact, we can use the measurability of that function to define all nephrons for which the function is less than some value, beginning to identify a measure of the precursor of kidney disease, here define $A_D = \{\omega_i \,/\, f_T(\omega_i) < C\}$.

## The eye and color vision

We can develop a similar model for the reception of light by the retina. The retina is a cell-dense organ at the back of the eye. It is composed of two different active types of light sensitive cells, cones and rods. The rods are sensitive to low light circumstances and transmit principally black and while images. The cones principally transmit color.

In the fovea (the small area of the retina where the visual acuity is the greatest), the density of cones is approximately 150,000 per mm$^2$. This density suggests that we might use indicator functions to develop a metric for governing the sensitivity of the retina to light. Let's begin with identifying a $(\Omega, \Sigma)$ pair reflecting the sample space and σ-algebra of cones in the retina. We let $\omega_i$ be a particular cone.

Consider a collection of sets $\{A_k\}$ each representing a range of mutually exclusive sets of contiguous light wavelengths.

Each individual set $A_k$ is a length in nanometers and can be quite narrow. Now let's also define an indicator function $e_k(\omega_i) = 1_{\omega_i \subset A_k}$ This indicator function maps 1 to a cone if that cone's sensitivity to light is in the set $A_k$. Then following the previous examples, then the sequence

$$e_1(\omega_i), e_2(\omega_i), e_3(\omega_i), e_4(\omega_i) \ldots e_k(\omega_i) \ldots e_n(\omega_i)$$

is the spectrum of the cone's sensitivity.

We can also summarize that cone's sensitivity to light by computing $f_m(\omega_i) = \sum_{k=1}^{m}\alpha_k 1_{\omega_i \subset A_k}$.

We can also define the set's sensitivity to light in a region of cones denoted by $R$ by $f_m(R) = \sum_{\omega_i \subset R}\sum_{k=1}^{m}\alpha_k 1_{\omega_i \subset A_k}$. Should the set $R$ be composed of millions of cones, then we have a very detailed and intricate definition of that retina region's sensitivity to light.

## Clinical trial analyses

The statistical analyses of a clinical trial follows a collection of rules that are both efficient but also heuristic. Typically, analysis types are divided into prospective (protocol driven) or exploratory evaluations. Prospective analyses are then divided into primary evaluations where the type I error rate is controlled for multiple testing, or secondary, where type I error testing occurs at a nominal (typically 0.05) level. Consider a design in which the statistical information (the inverse of the variance of an estimate) is also formally considered. We will use a simple function to assess the contribution an estimator makes.

Here our $\left(\Omega_q, \Sigma_q\right)$ will be the space of all statistical analyses that can be conducted from a clinical trial's data set that addresses a particular question. Let $\omega_i$ be a single analysis from this effort, i.e., $\omega_i \subset \Sigma$. Let us now create a collection of analysis sets $A_k$. For example, one such set might be the class of exploratory analyses, another would be prospectively declared secondary analyses, etc. Then we define the indicator function $e_k\left(\omega_i\right) = 1_{\omega_i \subset A_k}$. This simply classifies the $\omega_i^{th}$ analysis. In this case, the collection of sets $\left\{A_k\right\}$ need not be mutually exclusive. For example an analysis might be an exploratory subgroup analysis, where there is a member of $\left\{A_k\right\}$ $A_{i*}$ that denotes exploratory and another $A_j$ that denotes an analysis that is a member of a subgroup. In this case the sequence of indicator functions

$$e_1\left(\omega_i\right), e_2\left(\omega_i\right), e_3\left(\omega_i\right), e_4\left(\omega_i\right)\ldots e_k\left(\omega_i\right)\ldots$$

represents the classification, or the spectrum of properties of the statistical analysis $\omega_i$. We now create the simple function $f_n\left(\omega_i\right) = \sum_{k=1}^{n} \alpha_k 1_{\omega_i \subset A_k}$ where the constants $\alpha_k$ reflect the weights of the analyses. Then we might consider $f_n\left(\omega_i\right)$ the contribution of the $\omega_i^{th}$ statistical analysis to the conclusion of the trial. We then define the set $R$ as a collection of analyses $\left\{\omega_1, \omega_2, \omega_3 \ldots\right\}$ of interest. Then $f_m\left(R\right) = \sum_{\omega_i \subset R} \sum_{k=1}^{m} \alpha_k 1_{\omega_i \subset A_k}$ is the contribution of the set of analyses *R* to the clinical trial results.

## Summary of simple function creation

Through these examples we have developed a process that we can use to build set functions in public health. Specifically, after creating our $\left(\Omega, \Sigma\right)$, we focus on an element $\omega_i \subset \Sigma$ and construct an indicator function for a relevant measurable characteristic of $\omega_i$.

Recognizing that this particular characteristic is just one of many possible facets of the characteristics of interest, we create first the spectrum of characteristics of $\omega_i$, $e_1\left(\omega_i\right), e_2\left(\omega_i\right), e_3\left(\omega_i\right), e_4\left(\omega_i\right)\ldots e_k\left(\omega_i\right)\ldots$ and then a simple function $\sum_{k=1}^{m} \alpha_k e_k\left(\omega_i\right)$ where the $\alpha_k$'s are scaling constants (converting the function into the units of interest) and *m* is the total number of characteristics the function is to consider. We then define the set function on the set *S* contained in the σ-algebra $\Sigma$. If the character of the set *S* is just the sum of the characters of the elements of *S*, then we compute

$$f\left(S\right) = \sum_{\omega_i \subset S} \sum_{k=1}^{m} \alpha_k e_k\left(\omega_i\right) \cdot$$

Note that since this is a non-parametric approach, we will not recognize a functional form for $f\left(S\right)$ (i.e., $f\left(S\right)$ is not quadratic, or trigonometric for example).

These are useful examples, and they become even more illuminating if we focus on the properties of the collections of sets $\omega_i \subset S$ of interest. The examples of toxicity, urine production and vision all have as there $\left(\Omega, \Sigma\right)$ basis the tacit assumption that there are only finitely many sets in $\Omega$ and in $\Sigma$. While this assumption is valid for these examples, (there are only finitely many nephrons in the kidney), we want to develop a theory that encompasses many more possible elements.

For example, if $\Omega$ was the sample size for the number of possible phenotypes of a hepatocyte, then how many elements would it have? Or if $\Omega$ was the set of all possible adverse events, how large would it be? In these circumstances, $\Omega$ and $\Sigma$ contain many more outcomes which are as numerous as the irrational numbers (nondenumerable). So any development must take into account that when we talk about the set of all $\omega_i \subset A \subset \Sigma$, this collection of $\omega_i$ may be infinite and nondenumerable. This consideration delivers us to the concept of limits and integration.

## Limits of simple functions

One of the disadvantages of the simple function generation process described above is that the final result does not have an easily recognizable form and therefore, the function's behavior can be difficult to describe. For example, if the function is cubic, then we know to expect a smooth trajectory that changes direction twice. Parameterization helps us to see, characterize and understand the function's behavior.

Not so with simple functions. It is quite difficult to see and appreciate their behavior in the abstract. For example, consider the retinal reception example above, where $f_n(\omega_i) = \sum_{k=1}^{n} \alpha_k 1_{\omega_i \subset A_k}$ is a measure of the sensitivity to light of the $i^{\text{th}}$ cone. What does this function look like? How do we even describe the abscissa of this function? And what about *n*? What happens as *n* gets larger and larger? Does it even make sense for us to talk about convergence of $f_n$ as $n \to \infty$? And to what function $f$ would it converge?

Stepping back to consider the physiology of vision for a moment, it clearly makes no sense to talk about the convergence of $f_n(\omega_i)$ to the same value for different values of $\omega_i$ since there is no biologically plausible argument stating that the light sensitivity function of one cone should converge to that of another cone.[*] However, if we fix *i*, then what of the long term behavior of $f_n(\omega_i)$?

Even though we do not know what the limit of this function is, we can deduce that it does converge for any given $\omega_i$. [†]As it turns out, this convergence and the fact that the constants $\alpha_k > 0$ are what we need to assert the integrability of these functions.

[*] This is another way to say that $f_n(\omega_i)$ is not uniformly convergent across all of the $\omega$ cones.

[†] Recall that a sequence $x_n$ is Cauchy if the further out in the sequence we get, the function values approach each other. Assume that we have an exceedingly fine mesh of $\{A_k\}$. Begin by recognizing that a given cone is not receptive to all frequencies of light. We can therefore reorder the light frequencies $A_k$ so that all of the frequencies of light to which the $\omega_i$ is not sensitive occur at the end of the sequence. Thus, for these frequencies $e_k(\omega_i) = \alpha_k 1_{\omega_i \subset A_k}$ are all equal to zero and the function $f_n(\omega_i)$ does not increase. Then we can find an *N* such that for all $n > N$ and $m > N$ $f_m(\omega_i) = f_n(\omega_i)$ or $|f_m(\omega_i) - f_n(\omega_i)| = 0 \le \varepsilon$ satisfying the Cauchy criteria for convergence.

But what does integrability mean? Specifically, can we integrate this simple function? What does integration even mean if we cannot write down the function's final form, but conceive of as only $\lim_{n\to\infty} f_n\left(\omega_i\right)$?

Measure theory answers these questions for us.

Advanced Probability

# Measure and its Properties

We have talked in fairly nonmathematical ways about the concept of measure, discussing it as an operation on a set that produces that set's content or value. As initially developed by Henri Lebesgue, its use was focused on the sets of numbers that constitute intervals on the real line. This tool of using the real line as the object of measure theory helps us to understand some of its implications, since we know so much about the real line (e.g., how to measure the length of an interval).

However, the intriguing nature of measure theory for us is that it applies to measuring set content in general and is not restricted to real line intervals. Therefore, measure theory can be used to measure the wealth of families, the content of microRNA in a specimen, or the quantity of radiation in a room.

But, while we have a great deal of freedom in defining our "measure", it must have some common features to be useful. Here, we introduce its important characteristics and provide their motivations.

Prerequisites
An Introduction to the Concept of Measure
Elementary Set Theory
Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Simple Functions in Public Health

## Sample space and the sigma algebra

Recall that the sample space $\Omega$ is the ultimate source of sets that concern us. The members $\omega$ of $\Omega$ are the building blocks of sets in which we hold the greatest interest. Recall that the set $\Omega$ can have a relatively small number of elements (for example the number of patients in an infectology ward on a given day), or it can have an immense number of sets (the individual cubic nanometers of atmosphere over the Pacific Ocean). The limitations of the constituents of $\Omega$ reside only within the scope of the problem and the imagination of the worker.

Once $\Omega$ is established as the foundation, the $\sigma$-algebra $\Sigma$ is generated. Remember that $\Sigma$ is nothing more than the collection of sets built from a combination of the elements in $\Omega$ using the elementary set operations of unions, intersections, and complements. While $\Sigma$ is easy and systematic to construct, the actual number of sets in $\Sigma$ can be immense.

To begin, every element $\omega_i$ that is contained in Ω is also contained in Σ. Σ also contains the null set. In addition, Σ contains every possible union of different elements in Ω, first taken two at a time $\{\omega_1 \cup \omega_2\}, \{\omega_1 \cup \omega_3\}, \{\omega_1 \cup \omega_4\}$.... , then three at a time, and so on. Next, Σ contains all of the intersections, then unions of intersections, then intersections of unions in all of their complexity. From here, the process of building Σ continues, this time including complements of sets.

Thus, even when Ω is small, Σ can be quite large[*], and when Ω is large (such as cubic mm. of soil at an industrial waste dumping ground) then the $\sigma$-algebra Σ can be quite overwhelming.

## The four core measure properties

Once Σ is identified, we are free to create set functions on it. Remember that the only property that a measurable function must have is that it must be positive, and that it must pass the inspection test, i.e., every value that it takes must map back to a set in Σ. We have tremendous freedom in defining measurable functions.

However, measure is different than a measurable function. A measurable function is a non-negative set function that passes the inspection test and assigns a value to a set. Measure does not just assign a value to a set–it assigns *content*. Thus there are properties of content assessment that go beyond simply assigning a number to the set:

### Measure property 1

*If set $A$ is a member of Σ,, then $\mu(A)$ (called "the measure of A") must be a non-negative real number.*

Just as measurable functions must be positive, so too must measure be a number greater than or equal to zero. Thus the measure of a set does not map a set to another set, or a set to a multidimensional vector. It maps the set to a real number that cannot be negative. This real number, $\mu(A)$ is the measure of, the content of, or the value of the set $A$.

How the set is converted into a number is the property of the measure. For example, if Ω is the set of all cubic centimeters of a lake, and the set $A$ is contained in its $\sigma$-algebra Σ, then one possible candidate for a measure[†] might be the oxygen content of sets $\nu(A) \quad A \subset \Sigma$; a wholly separate measure $\xi(\omega_i)$ could return the algae density of that same set $A$. In addition, the measure could simply be a 0-1 dichotomous measure, e.g., does the set $A$ contain any hydrocarbons. However, the measure or value itself must reside on the non-negative real line.

### Measure property 2

*If $\mu$ is a measure on $(\Omega, \Sigma)$ then $\mu(\varnothing) = 0$.*

This statement buttresses the notion that the measure provides value or content to sets residing in Σ by permitting no value or content to the empty set. Even though the set ∅ resides within Σ, the measure we attach to it is by definition zero. For example, while one can quite reasonably define a measure based on the number of bacteria that inhabit a cubic mm of space, it makes little sense to ask what is the measure or content of "no space". The statement $\mu(\varnothing) = 0$ is a mathematical statement of that reality, tightening the link between measure as a practical assessment of content.

[*] If for example, Ω contains three and only three elements, Σ contains over thirty elements.
[†] Assuming that the other three measure properties are upheld.

### Measure property 3

*If sets $A$ and $B$ are both elements of $\Sigma$ such that A is contained in B, then $\mu(A) \leq \mu(B)$. Another way to say this is that if B contains A, then $\mu(B) \geq \mu(A)$.*

This follows easily from consideration of elementary set theory. If $A = B$ then since $A$ contains $B$ and $B$ contains $A$, it must be so that $\mu(A) = \mu(B)$. For other circumstances, we begin with a very helpful formulation of the set $B$ as

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap \left(A \cup A^c\right) \\ &= \left\{B \cap A\right\} \cup \left\{B \cap A^c\right\}. \end{aligned}$$

Note that $B \cap A$ and $B \cap A^c$ are disjoint.

Now, if $A \subset B$, then $B \cap A = A$. The set $B \cap A^c$ is colloquially expressed as "$B$ without $A$ ", or $B - A$.[*]

Since these sets are disjoint, $\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A)$. We will return to this construction, providing a formal demonstration of this when we consider the next property of measure.

As another example, from this perspective, we can think of a collection of sets $A_n$ all in $\Sigma$, such that $A_1 \subset A_2 \subset A_3 \subset ... \subset A_n \subset ...$ . This sequence property of increasing sets, each one containing the set preceding it, is referred to as monotonicity. Then according to this measure property we must have

$\mu(A_1) \leq \mu(A_2) \leq \mu(A_3) \leq ... \leq \mu(A_n) \leq ...$

This third measure property provides for the accumulation of measure, which increasing as the set's content increases so that larger sets accumulate no less measure than smaller sets (Figure 1).

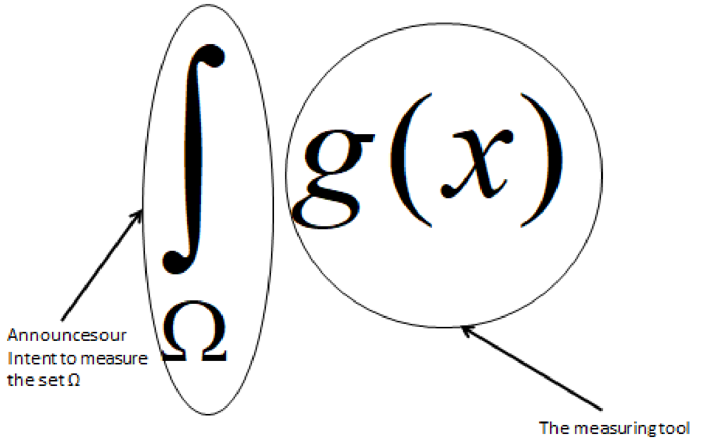


Figure 1. Use of the integral in measure theory. The integral sign announces our plan and states the set we wish to measure, while the integrand is the measuring tool.

[*] Technically, there is no set operation as $B - A$. However, since $B \cap A^c$ expresses the same concept as substraction, we will let $B - A = B \cap A^c$.

### *Measures of unions of sets*

Our work on the properties of measure along with our set theory background has put us in the position to consider the measure of the unions of sets. There are several different scenarios that we will consider.

### Equivalence

Let's begin with the concept of equivalence, that is, if $A = B,$ then $\mu(A) = \mu(B).$

If $A = B,$ then both $A \subseteq B$ and $B \subseteq A.$ From property three of measure we know that $A \subseteq B$ implies that $\mu(A) \leq \mu(B).$ However, it is also true that $B \subseteq A,$ implying that $\mu(B) \leq \mu(A).$The only way that both inequalities involving the measure are true is if $\mu(A) = \mu(B).$ Thus, equivalent sets must have equivalent measure.

We can take this further. If the sets $A$ and $B$ are the same, then we would expect that the measure of the union of the sets $\mu(A \cup B) = \mu(A) = \mu(B).$ How can we show this?

If $A = B,$ then $A \cup B = A = B$ and since the sets are equivalent, $\mu(A \cup B) = \mu(A) = \mu(B).$

### Disjoint sets

If the two sets $A$ and $B$ are disjoint, then is it true that $\mu(A \cup B) = \mu(A) + \mu(B)$?

Let's assume that it is not true. For example, that $\mu(A \cup B) \leq \mu(A) + \mu(B).$ Then there must be some set $C$ such that $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(C).$ Where must this set $C$ reside? If $C$ is disjoint from each of sets $A$ and $B$ , then $C$ is not part of the union of sets $A$ and $B$ and the proposition fail.

Alternatively, $C$ could reside within either $A$ or $B.$ Assume $C \subset A.$ Then $\mu(A) - \mu(C) = \mu(A \cap C^c),$and $A \cap C^c \cup B \neq A \cup B.$ A similar argument follows if $C \subset B.$ So $\mu(A \cup B)$ cannot be less than the sum of the measures.

Can the measure of the union of disjoint sets be greater than the sum of the measures? That would mean $\mu(A \cup B) = \mu(A) + \mu(B) + \mu(C).$ Clearly the set $C$ would reside outside each of sets $A$ and $B,$ and therefore $C \not\subset A \cup B$ , another contradiction.

Thus $\mu(A \cup B) = \mu(A) + \mu(B)$ for two disjoint sets.

### Measure of complements

The notion of disjoint sets serves well when we try assess the measure of complements. Given an $(\Omega, \Sigma)$ pair and a set $A \subset \Sigma,$ we know that $A \cap A^c = \varnothing.$ We also know that $A \cup A^c = \Omega.$ These two statements permit us to write

$$\mu(\Omega) = \mu(A \cup A^c) = \mu(A) + \mu(A^c)$$
$$\mu(A^c) = \mu(\Omega) - \mu(A),$$

### Nondisjoint Sets

There are several circumstances that must be formally considered when the sets $A$ and $B$ are not disjoint.

For example, if $\mu(A) \neq \varnothing$ and $A \subset B$ then $A \cup B = B$, and $\mu(A \cup B) = \mu(B) < \mu(A) + \mu(B)$.

However, if $A \neq B$, and $A \cap B \neq \varnothing$, some additional steps are needed to find and then bound $\mu(A \cup B)$.

We know that we can write

$$A = (A \cap B) \cup (A \cap B^c)$$
$$B = (A \cap B) \cup (B \cap A^c).$$

From this construction, we see that $A \cup B$ is the union of three disjoint sets, $A \cap B^c, B \cap A^c, A \cap B.$ Since these sets are disjoint, we can write

$$\mu(A) = \mu(A \cap B) + \mu(A \cap B^c)$$
$$\mu(B) = \mu(A \cap B) + \mu(B \cap A^c).$$

Thus $\mu(A) + \mu(B) = \mu(A \cap B^c) + 2\mu(A \cap B) + \mu(B \cap A^c).$

Now turning to the union, we note that since

$$\begin{aligned} A \cup B &= (A \cap B) \cup (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B) \\ &= (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B) \end{aligned}$$

And these are disjoint sets, then

$$\mu(A \cup B) = \mu(A \cap B^c) + \mu(A \cap B) + \mu(B \cap A^c).$$

This is less than the $\mu(A) + \mu(B)$ since the measure of the intersection is only needed once. So, in general we may write

$$\mu(A \cup B) \leq \mu(A) + \mu(B).$$

Based on this result, Property four of measure is not surprising.

### *Measure Property 4*

*If an infinite sequence of disjoint sets* $A_n$ *is contained in* $\Sigma$*, then* $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} \mu(A_n)$ *This is known as countable additivity.*

Note that the upper bound of the index is infinity. There is another concept where the upper bound is finite, termed finite additivity i.e., $\mu\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{i=1}^{k} \mu\left(A_n\right)$. This is a derivative of the countable additivity property which can be easily demonstrated.[*]

Non-null intersections of these sets requires a change based on the consideration of this complication. If the sets are not disjoint, then $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{i=1}^{\infty} \mu\left(A_n\right)$. Furthermore, using DeMorgan's law we can write this as $\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} A_n^c\right) = \sum_{n=1}^{\infty} \mu\left(A_n^c\right)$. The countable additivity property is very powerful[†]

As another example, let's assume we have two sets $A$ and $B$ each contained in the σ-algebra $\Sigma$. As is our custom, let's define the set operation of $A - B$ as $A - B = A \cap B^c$. Then the set is equal to different sets based on the relationship between the two sets (Figure 2).

<<Figure 2>>

We can easily deduce $\mu(A - B) = \mu\left(A \cap B^c\right)$ for $B \subset A$. We begin by writing $A = B \cup \left(A - B^c\right)$.[‡] Then, the sets $B$ and $A - B$ are disjoint since $B \cap (A - B) = B \cap \left(A \cap B^c\right) = (B \cap A) \cap \left(B \cap B^c\right)$ which is $B \cap \varnothing = \varnothing$. Thus $\mu(A) = \mu\left(B \cup \left(A - B^c\right)\right)$ and by finite additivity, $\mu\left(B \cup (A - B)\right) = \mu(B) + \mu(A - B)$. Therefore, $\mu(A - B) = \mu(A) - \mu(B)$ when $B \subset A$.

[*] We note that $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu\left(A_n\right)$, Now lets choose our collection of sets such that for $n$ = 1 to $k$, $A_n \neq \varnothing$. However, for all $n > k$, $A_n = \varnothing$. Then $\bigcup_{n=1}^{\infty} A_n = \left\{\bigcup_{n=1}^{k} A_n\right\} \cup \bigcup_{n=k+1}^{\infty} \varnothing = \bigcup_{n=1}^{k} A_n$. Thus, for this collection of $A_n$, the infinite union reduces to a finite union of exactly the $A_n$ sets that we want. Also, we have $\sum_{n=1}^{\infty} \mu\left(A_n\right) = \sum_{n=1}^{k} \mu\left(A_n\right) + \sum_{n=k+1}^{\infty} \mu\left(A_n\right) = \sum_{n=1}^{k} \mu\left(A_n\right) + 0$. Thereforefor any collection of finite sets contained in $\Sigma$, $\mu\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} \mu\left(A_n\right)$.

[†] For example, if we follow the proof above, this time lettting $A_n = \varnothing$ for all $n$, we find $\bigcup_{n=1}^{\infty} A_n = \varnothing$. Also $\sum_{n=1}^{\infty} \mu\left(A_n\right) = 0$. Since in this case the $A_n$ are each identifable and must be nonnegative, the only possible choice for their value is zero. Thus $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu(\varnothing) = \sum_{n=1}^{\infty} \mu\left(A_n\right) = 0$.

[‡] This follows from $B \cup (A - B) = B \cup \left(A \cap B^c\right) = (B \cup A) \cap \left(B \cup B^c\right)$. Since $B \subset A$ then $B \cup A = A$. $B \cup B^c = \Omega$, and $B \cup (A - B) = A \cap \Omega = A$.

This fourth "countable additivity" property of measure, while appearing somewhat abstract right now is actually quite important. It permits us to deconstruct the measure of a large union of sets into the measures of the individual constituents of these sets.

In addition, if the individual sets that compose the union are disjoint, we can simply sum the measures. Much of the developmental work of measure theory is based on the ability to deconstruct the union of sets into an equivalent union of different but disjoint events, then using the property of countable additivity to sum the measure of these disjoint sets. This is part of the heart of the proof of the monotone convergence theorem.

Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Monotone Convergence Theorem
Some Classic Measure Theory Results
Asymptotics
Tail Event Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables

# Working with Measure

Prerequisites
Elementary Set Theory
Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties

## Why does measure need its properties?

Most measurable set functions do not constitute a measure. In order to be so, a measurable function must have the four properties we just described: 1) map a set to a nonnegative number on the real number line, 2) map a null set to 0, 3) if the set *A* contains the set *B*, then the measure of *A* is greater than or equal to that of *B*, and 4) countable additivity.

So, why do we need these four properties, and especially, why insist on countable additivity?

The motivation for properties 1 – 3 comes from the need and desire to make measure useful. We want it to assess the content or value of an item. Properties 1-3 nicely match this intuition.

However, property four (countable additivity) has a different motivation. It is focused on aggregating measure across sets. Specifically, given that measure can be assigned to a collection of sets, it governs how measure may be assigned to other sets which can be formed from the original collection.

## Path construction

For example, suppose that we want to measure a set *A*. The procedure to obtain the measure of *A* is to "cover" *A* with sets that each have known measure. If the collection of sets whose measure we know "nicely" covers the set *A* then we can build up to the measure of *A* by 1) noting the precise union, intersections and complement operations needed to construct *A* (the path of set *A*' s construction), and then 2) build up the measure of *A* by, following the path of its construction, adding and subtracting the measure of the constituent sets that make up *A*. This is where countable additivity is important.

## Example: Measure of three sets

As a simple example. Consider the sets *A*, *B* and *C* in Figure 1.

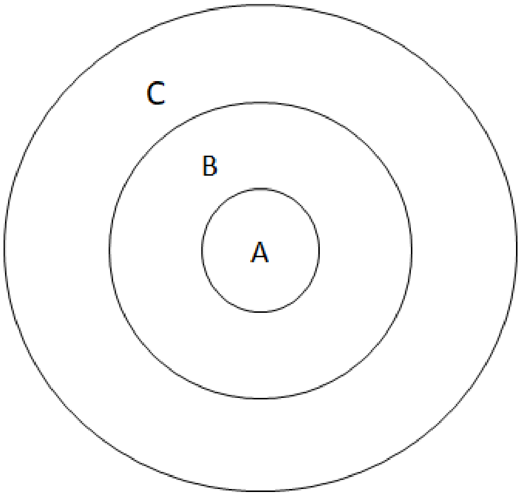


Figure 1. An example of the assertion that $A \subset B \subset C$ implies $\mu(A) \le \mu(B) \le \mu(C)$.

We are tasked with finding $\mu(A \cup B \cup C)$. It is easiest if we can disassemble the set $A \cup B \cup C$ into the union of disjoint sets which are themselves a function of the original three sets. From Figure 1, we can write

$$A \cup B \cup C = \begin{array}{l} A \cap (B \cup C)^c \ \bigcup \ B \cap (A \cup C)^c \ \bigcup \ C \cap (A \cup B)^c \ \bigcup \\ A \cap B \cap C^c \quad \bigcup \ B \cap C \cap A^c \quad \bigcup A \cap C \cap B^c \quad \bigcup \\ A \cap B \cap C \end{array}$$

We can then write the $\mu(A \cup B \cup C)$ as the sum of the measures of these seven sets. While this is not the only way to write the measure of this union, its advantage is that it breaks the union down into a collection of subsets. If the measure of these subsets can be assessed, then the measure of $\mu(A \cup B \cup C)$ is easily attained.

For example, if it takes $\bigcup_{n=1}^{k} A_n$ to nicely approximate set $A$, where all of the sets $A_n$ are mutually disjoint*, then $\mu(A) = \mu\left(\bigcup_{n=1}^{k} A_n\right)$ and since the countable additivity provides $\mu(A) = \mu\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} \mu(A_n)$ then we have the measure of the set we want.

## Outer and inner measure

The approximation of sets by collections of other sets is an important concept in set and measure theory.

Let's explore this in detail by beginning with a thought experiment. There is a set $A$ within the interval $(a,b)$, $A \subset (a,b)$. You want to know its measure, which, in this case we define as the line's length.†

You have at your disposal and infinite number of people each of whom has their own individual collections of intervals, whose lengths they know. Some have a finite collection of

* This assumption of mutual exclusivity is a big one. However, in many circumstances, as we did in the previous example, we will have to work to have the sets that cover $A$ be mutually disjoint.

† This is what is called Lebesgue measure, not to be confused with Lebesgue measure theory which is more general.

intervals, others have an infinite but countable number of intervals. But, each claims that by approximating $A$ though set operations applied to their intervals, they will be able to compute $\mu(A)$.

Since $A$ is known, you can divide the candidate collections of intervals into four different groups. The first group of candidates have families of intervals that do not intersect with $A$ at all (i.e., they can only approximate $A^c$). They cannot be used to get the length (or measure) of $A$ and are rejected out of hand.

The second group only cover part of $A$ and also cover part of $A^c$ They are also rejected.

### *Outer measure*

The third group of individuals have interval families that cover all of $A$ but also cover part of $A^c$. Lets collect these sets. It the $i^{\text{th}}$ such individual provides disjoint sets $A_{i,1}, A_{i,2}, A_{i,3}, \ldots. A_{i,n}, \ldots..$ and we know that the union of these $\bigcup_j A_{i,j}$ covers $A$, or $\bigcup_j A_{i,j} \supset A.$ Now this coverage may be precise and capture $A$ exactly, or may be coarse, and cover not just $A$ but a sizable component of $A^c$ We do this for every individual in this group, collecting their intervals $A_{i,1}, A_{i,2}, A_{i,3}, \ldots. A_{i,n}, \ldots..$ some collections doing a good job and others doing poorly.

Now, we can manage their varying ability to "tightly" cover $A$ by noting that since they all cover $A$, their intersection must cover $A$. Thus $\bigcap_i \bigcup_j A_{i,j}$ should be a tight upper bound for $A$. This is the infimum of these unions of sets. We write this as $\inf_i \left[\bigcup_j A_{i,j}\right]$. If the sets are disjoint, then the $\mu(A) < \inf_i \left[\sum_{j=1}^{\infty} \mu(A_{i,j})\right]$. We define $\inf_i \left[\sum_{j=1}^{\infty} \mu(A_{i,j})\right]$ as the outer measure of $A$.

### *Inner measure*

The fourth group of individuals each has a collection of sets, but this time they reside wholly within $A$. As before, the $k^{\text{th}}$ individual of this group produces a collection of infinite disjoint $B_k$ sets such that $\bigcup_j B_{k,j} \subset A.$ Each of these union serves as an approximation of $A$. However, these unions are each within $A$, the best coverage of $A$ is obtained by taking the union of all of these unions. Thus $\bigcup_k \bigcup_j B_{k,j} \subset A$ approximates $A$. This is the supremum or greatest lower bound, $\sup_k \left[\bigcup_j B_{k,j}\right]$ and is known as the inner measure of $A$. We write $\mu(A) \geq \sup_k \sum_{j=1}^{\infty} \mu(B_{k,j})$.

Thus, we have two ways we can use sets with known measure to cover a set of unknown measure. However, neither the $\inf_i \left[\bigcup_j A_{i,j}\right]$ nor the $\sup_k \left[\bigcup_j B_{k,j}\right]$ are guaranteed to be exactly $A$. They approximate $A$ from above and below respectively. Since we know that each of the $A_{i,j}$ and $B_{k,j}$ have measure, then $\mu\left(\sup_k \left[\bigcup_j B_{k,j}\right]\right) \leq \mu(A) \leq \mu\left(\inf_k \left[\bigcup_j A_{i,j}\right]\right)$. Using countable additivity, we have $\sup_k\left[\sum_{j=1}^{\infty}\mu\left(B_{k,j}\right)\right]\leq\mu\left(A\right)\leq\inf_i\left[\sum_{j=1}^{\infty}\mu\left(A_{i,j}\right)\right]$, but this does not tell us what the $\mu\left(A\right)$ is. We can write this as $\mu_*\left(A\right)\leq\mu\left(A\right)\leq\mu^*\left(A\right)$, where $\mu_*\left(A\right)$is the inner measure and $\mu^*\left(A\right)$ the outer measure of $A$.

Let's examine the outer measure of A. Since we know $A\subseteq\inf_i\bigcup_{j=1}^{\infty}A_j$ there is some set $E^*$ such that $E^*\subseteq\inf_i\bigcup_{j=1}^{\infty}A_j$ and $E^*\subseteq A^c$. Thus $\mu\left(A\right)+\mu\left(E^*\right)=\mu\left(\inf_i\bigcup_{j=1}^{\infty}A_j\right)$. Similarly, regarding inner measure, there is a set $E_*$ such that $\sup_k\left[\bigcup_j B_{k,j}\right]\cup E_*=A.$ then

$\sup_k\left[\sum_{j=1}^{\infty}\mu\left(B_{k,j}\right)\right]+\mu\left(E_*\right)=\mu\left(A\right)=\inf_i\left[\sum_{j=1}^{\infty}\mu\left(A_{i,j}\right)\right]+\mu\left(E^*\right)$. And if $\mu\left(E_*\right)=\mu\left(E^*\right)$, then outer measure and inner measure are equal and we say the set $A$ is measurable.

## Limits of measure

We have discussed how to compute and understand the limits of an infinite sequence of sets. A relevant question is how does measure operate term by term when we examine these sequences? For example if a sequence of sets $A_n$ converges to a set $A$[*] then under what circumstances can we say that the measure of the limit converges to the limit of the measures, i.e., $\mu\left(\lim_{n\to\infty}A_n\right)=\lim_{n\to\infty}\mu\left(A_n\right)$.

Why do we even care that this property may be true? The reasoning begins with the contribution of simple functions. As we have shown, it is possible to generate simple functions that assess the presence of characteristics of systems in public health. We will be interested in determining the value or measure of these systems. This would be akin to computing $\mu\left(\lim_{n\to\infty}A_n\right)$.

However, it is difficult to see what $\lim_{n\to\infty}A_n$ actually is, much less take its measure. However we do know what the $\mu\left(A_n\right)$ is and it is possible to compute its limit. Therefore the equality $\mu\left(\lim_{n\to\infty}A_n\right)=\lim_{n\to\infty}\mu\left(A_n\right)$ enables us to compute a desired but unrecognizable function $\mu\left(\lim_{n\to\infty}A_n\right)$ by taking the limit of a recognizable one.

However, while this equality is in general not true. it is a property of certain types of sets. One such collection of sets is those that have the monotonicity property; i.e., they are "increasing", i.e., in the sequence of sets $A_1, A_2, A_3,....,$ where $A_1\subset A_2\subset A_3\subset...\subset A_n\subset.....$ (Figure 1).

[*] Recall that, if the limit of an infinite sequence of sets $A_n$ exists, then every element of $A$ must be in all but finitely many of the $A_n$, and every element that is not in $A$ must only be in finitely many of the $A_n$, for $A$ to exist. This is clearly the case for an increasing sequence of sets for which $A$ contains all of the sets in the entire sequence..

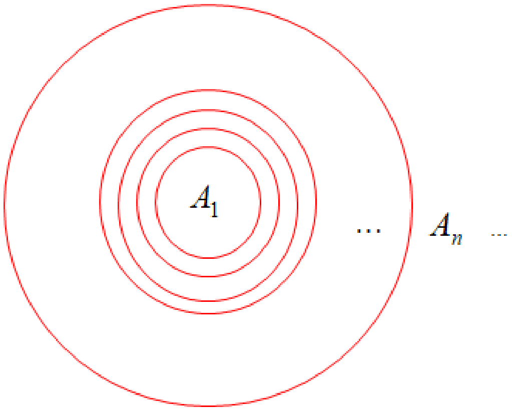


**Figure 1.** Depiction of a sequence of increasing sets

If the sequence of sets is increasing, we can demonstrate that $\mu\left(\lim_{n\to\infty}\bigcup_{i=1}^{n} A_n\right) = \lim_{n\to\infty}\mu(A_n)$.

The key is to recreate $A_n$ from a sequence of disjoint sets. Define a new but related infinite sequence of sets $\{B_n\}$ such that $B_1 = A_1$. $B_2 = A_2 - A_1$; $B_3 = A_3 - A_2$; This defines $B$ as the donut, i.e., the space between $A_n$ and $A_{n-1}$. (Figure 2)

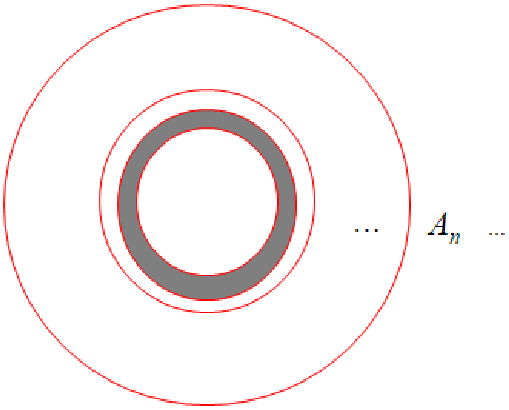


**Figure 2.** The gray “donut” is set $B_2$.

Note that the family of sets $B$ are disjoint.[*] The construction of this sequence of pairwise disjoint sets is critical because it permits us to invoke the countable additivity property of measure.

Let's begin. We note from the construction that in fact, $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$ and thus $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right)$. Now we can invoke countable additivity to write $\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n)$. This last expression is a simple infinite sum and can be written as $\lim_{n\to\infty} \sum_{i=1}^{n} \mu(B_i)$. Now, using finite additivity, $\sum_{i=1}^{n} \mu(B_i) = \mu\left(\bigcup_{i=1}^{n} B_i\right)$. However, $\bigcup_{i=1}^{n} B_i = A_n$, and thus find that $\lim_{n\to\infty} \sum_{i=1}^{n} \mu(B_i) = \lim_{n\to\infty} \mu\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \mu(A_n)$.

The complete development is as follows.

$$\mu\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mu(B_i)$$
$$= \lim_{n\to\infty} \mu\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \mu(A_n)$$

Note, that a key to this demonstration was the use of each of the concepts of countable and finite additivity. This finding will be an important ingredient in our proof of the Monotone Convergence Theorem.

This result of having the measure of the limit be the limit of the measure is quite useful. However, in the above example, it applies only to increasing sequence of sets. However, there are other constructions that are quite useful.

For example, consider the set of all analyses that are carried out in a clinical trial. Among these sets are results provided on primary endpoints, secondary endpoints, and exploratory endpoints. In addition, there are subgroup analyses, nonparametric analyses, Bayesian analyses, and analyses based on the general model.

Let $A_n$ be the knowledge gained from the $n^{\text{th}}$ analysis. Then, since the analyses are themselves not independent of one another, the information that they provide in response to a clinical question e.g., "is the therapy beneficial" overlap. If we can depict each circle as a single $A_n$, then they might appear intensely nondisjoint. (Figure 3).

[*] Consider two sets $B_i$ and $B_j$ .Then $B_i \cap B_j = (A_i - A_{i-1}) \cap (A_j - A_{j-1})$ which is $(A_i \cap A_{i-1}^c) \cap (A_j \cap A_{j-1}^c)$. While $A_i \cap A_j = A_i$, $A_i \cap A_{j-1}^c = \varnothing$, making the enire complex equal to the null set, proving the two sets $B_i$ and $B_j$ are disjoint.

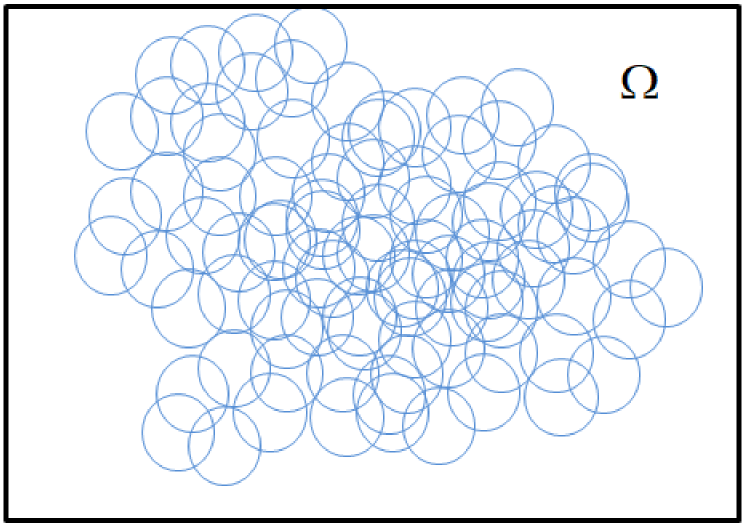


**Figure 3.** Overlap of analysis content in a clinical research effort.

We are interested in the total information in the system which we represent as $\bigcup_{n=1}^{\infty} A_n$ and we would like to compute the measure of this universe $\mu\left(\bigcup_{n=1}^{\infty} A_n\right)$. Clearly we cannot write $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ and thus $\mu\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$ because of the nondisjoint nature of the sets.

However, suppose we create the following sequence of sets.

$$
\begin{aligned}
&C_{1`} = A_1 \\
&C_{2`} = A_1 \cup A_2 \\
&C_{3`} = A_1 \cup A_2 \cup A_3 \\
&\ldots
\end{aligned}
$$

Then $C_1 \subset C_2 \subset C_3 \subset \ldots$ and $\bigcup_{n=1}^{\infty} C_n = \bigcup_{n=1}^{\infty} A_n$, Thus, $\mu\left(\bigcup_n A_n\right) = \mu\left(\bigcup_n C_n\right)$. From our previous construction, we have $\mu\left(\bigcup_n C_n\right) = \mu\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} C_i\right) = \mu\left(\lim_{n\to\infty} C_n\right)$. However, since the collection of sets $\{C_n\}$ is increasing, we can write $\mu\left(\lim_{n\to\infty} C_n\right) = \lim_{n\to\infty} \mu(C_n)$. Thus

$$\mu\left(\bigcup_n A_n\right) = \lim_{n\to\infty} \mu(C_n).$$

Exploring this at a deeper level reveals that the set of disjoint sets $\{B_i\}$ that makes this work is

$$\begin{aligned}
B_1 &= C_1 = A_1 \\
B_2 &= C_2 - C_1 = A_1 \cup A_2 - A_1 \\
B_3 &= C_3 - C_2 = A_1 \cup A_2 \cup A_3 - A_1 \cup A_2
\end{aligned}$$

Thus the disjoint sets $\{B_i\}$ are the disjoint contributions of each analysis $A_i$. We can therefore write that

$$\mu\left(\bigcup_n A_n\right) = \lim_{n\to\infty} \mu(C_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mu(B_i).$$

Thus the measure of the union we seek is the sum of the measures of the independent (i.e., disjoint) components that make up that union.

This technique of carrying out a collection of set operations to give a sequence of sets the desired monotonicity property is quite helpful.

Next sections
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Monotone Convergence Theorem
Some Classic Measure Theory Results
Asymptotics
Tail Event Measure

Probability Foundations
Basic Properties of Probability
Counting Events

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables

# Measure-Based Integration

Prerequisite
An Introduction to the Concept of Measure
Elementary Set Theory
Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

## Background

Most students of calculus have an understanding of the role of integration i.e., of "finding the area under a curve" and after a period of re-acquaintance, can integrate familiar functions easily. They can therefore be forgiven for asking "What's the big deal with integration now?" when the topic comes up in measure theory.

After all, the notion of integration as area under the curve makes very good intuitive sense and is quite practical. Smooth, continuous functions comprise a very rich field, and our ability to manipulate their integrals requires no additional consideration from measure theory. The area under the curve approach extends to many problems in probability. Use of common distributions such as the normal distribution, Laplace distribution, chi-square distribution, etc. can be managed nicely. In fact, the central limit theorem – which itself requires no deep understanding of measure theory – leads to the broad generalization of the use of normal measure for many problems where the underlying probability model is not normal. While the actual probabilities have to be found in tables or other look-up instruments, the explicit application of measure theory by the student is not necessary.

So, why do we bother with measure theory for integration in general, and for probability in particular?

In a nutshell, we need measure theory because the problems that we solve without it are only a fraction of the problems that we need to solve, and the solution to these more complicated problems requires working with functions for whom the notion of area under the curve does not help e.g., the disease burden function.

The developments in measure theory that we review in this chapter provide us the ability to not just apply integration as we usually do, but to recognize circumstances where we need to integrate but in which the classic Riemann integral fails, and we must turn from our usual standard of integration. We will see that integration is not just taking the area over the curve, but is more generally the process of accumulation, Area under the curve is just one type of accumulation among many that we will use.

## Evolution of integration

In the typical calculus course, integration is presented as a "self-contained" process, with no past and no future. It simply is, and the student memorizes its rules and tools. In order to understand the weakness of the standard "area under the curve integration," we have to understand something of its history.

Integration was first discovered not as its own process, but instead as the reverse of differentiation.

Developed independently by Isaac Newton and Gottfried Leibnitz in the late 1600's[*], integration had no underlying theory of its own; the Fundamental Theorem of Calculus formally described the integral as the reverse of the derivative. In fact, it wasn't even called the "integral" but instead called the "antiderivative" . While there was a sense that the integral could be related to the area under a curve, the theory underlying this had yet to be developed.

It was Bernhard Riemann who introduced integration as "area under the curve" using the concepts of limits. Building on the work of Cauchy (who developed the epsilon-delta approach to the limit process), Riemann defined the integral of a curvilinear function *f*(*x*) by breaking the region over which one was integrating into many vertical rectangles, and then taking the limit of the sums of their areas as the rectangles were permitted to get thinner and thinner (Figure 1).

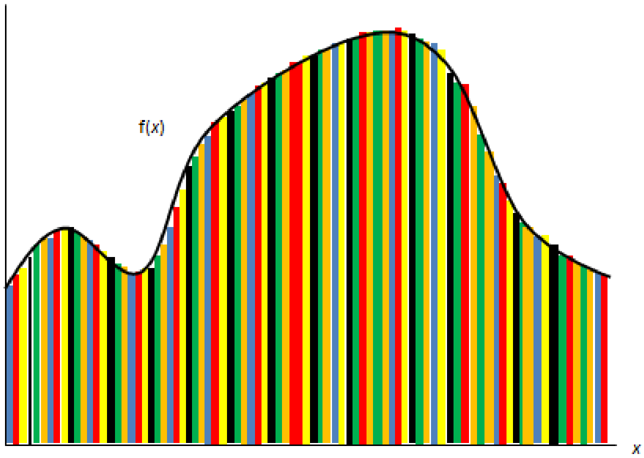


**Figure 1**. The standard area under the curve concept of integration requires a function that is smooth enough.

This underlying asymptotic approach solidified the theoretical concept of what was to be known as the Riemann integral. Essentially, one 1) partitions the domain of the function into intervals, 2) for each interval, compute the value of that function at a single point *x* within the interval, 3) compute the product of the function value and the interval width, producing an "area" then 4) sum the area of these rectangles.

This concept worked very nicely for smooth functions e.g., polynomials. In fact, in the standard calculus course, one typically works primarily if not exclusively with smooth curves, carrying out our standard integration procedures (e.g., partial fractions, integration by parts, trigonometric integration, etc.) easily, assuming that the foundation "rectangle" perspective is operating in the background.

However, they are some circumstances in which the concept of the area under the curve starts to breaks down. Consider for example, a curve that is smooth, except for a small number of discontinuities (Figure 2).

[*] Actually, the truly first theory of integration was developed by Archimedes in the 3rd century BC, using a quadrature method. However, it was tightly circumscribe by the requirement of geometric symmetry.

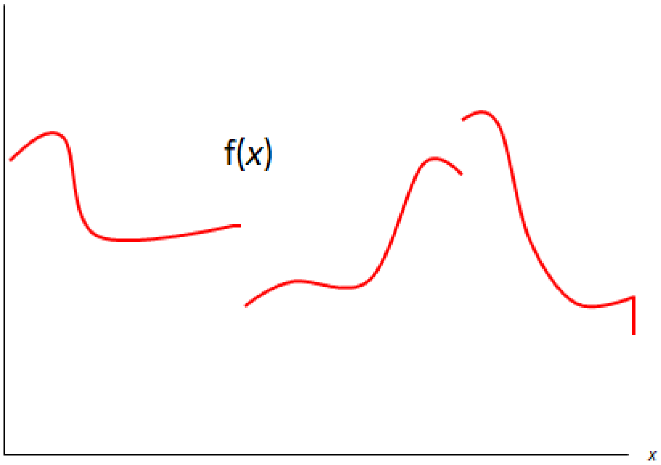


**Figure 2**. Riemann integration has area under the curve holds up with a "small" number of discontinuities.

Here, our area under the curve (Riemann) approach to integration still holds up nicely – but we have to make an adjustment. As long as we can ensure that the rectangle vertices whose areas that we need line up with the points of discontinuity, we can still create the rectangles that we need. Thus, discontinuities pose no problem for the Riemann approach − as long as there are not to many of them.

Nevertheless, this discontinuity adjustment revealed a potential weakness in the Riemann integration theory, and workers hammered away at the difficulty discontinuities could pose if there were too many of them. Ultimately they identified functions that "broke" the Riemann integration process.

The best known example is the Dirichlet function defined on the [0.1] as

$$f(x)=1 \text{ if } x \text{ is rational, and}$$
$$f(x)=0 \text{ if } x \text{ is irrational.}$$

The Dirichlet function is almost indescribably discontinuous since there are infinitely many rational numbers interspersed with uncountably many irrational numbers on the $[0,1]$ interval.

The difficulty here is not the value of the function (it is simplicity itself, taking on only the values of 0 or 1), but the domain of the function. The Dirichlet function uses the non-sparseness of the irrationals, the density of the rational numbers, and the intense inter-spersement of the rational and irrational numbers on $[0,1]$ so as to make the concept of "rectangle area as integration" break down. The problem is there is no "width" on which to base a rectangle's area.

This is an important concept and bears closer examination. The real number line is dense in the rationals, because, between any two rational numbers, however close they are, one can find another rational number. However, the rational numbers are also sparse; there is substantial space between any two rational numbers. This space is occupied by the irrational numbers. Although there are many more irrational than rational numbers the rational and irrational numbers are tightly interweaved, so much so that one cannot build a rectangle over just a set of rational numbers or over just a set of irrational numbers.[*]

However, this is just what the Riemann integral of the Dirichlet function requires. In order for it to be constructed, we must have an interval on the $x$ axis that is only rational numbers. Yet, this is impossible since every interval however tiny must contain irrational numbers. Thus, there is no way to identify a rectangle whose expanse on the $x$-axis consists of rational numbers to the exclusion of irrationals. Another way to say it is that the Dirichlet function is intensely discontinuous, having a discontinuity at every rational number in [0.1]. [†]It is this real number property of the domain of $x$ that causes the Riemann integral to break down.

The Riemann integral breaks down because of the sequence of steps required. In that process, we first focus our attention on finding contiguous values of $x$, essential to the rectangle building process. However, what if we reversed the process, instead focusing on the value of the function – not its domain. Choose a function value $f(x)$ then gather all of the values of $x$ that produce the same value of $f(x)$ and *measure the x's*.

## The Lebesgue integral

The combination of injecting measure into integration was introduced by Henri Lebesgue. Rather than build the rectangle by starting with the domain of the function, he recommended that we start with the range, i.e., the function values. He also suggested a measure that would be most useful to mathematicians.

Essentially, Lebesgue defined the measure of open sets of intervals on the real line as the length of the interval. This measure of the interval has come to be known as Lebesgue measure.[‡] This is only one of many types of measure, and we will not be restricted to it. However, it is this particular measure that is the basis for much of the application of Lebesgue measure theory to the real number line.

It is therefore important to distinguish Lebesgue measure theory (the concept of integrating by fixing the value of the function and measuring the set which produces that value) from Lebesgue measure (a particular type of measure in Lebesgue measure theory in which the measure of an interval on the real number line is the length of the interval).

How would Lebesgue measure work for the Dirichlet function? One writes the measure of this function as simply the sum of the values of this function multiplied by the measure of the set of $x$ that produce this value. In this case, it would be

$$0\text{* measure of irrationals on } \left[0,1\right] + 1\text{*measure of rationals on } \left[0,1\right].$$

The first term is zero. For the second, we note that the rational numbers, though infinite are countable. However, the Lebesgue measure of a countable set is zero. The second term is therefore 0 and the measure of the Dirichlet function is 0. Recall, that, using the Riemann concept, the integral was simply undefined.

[*] As a helpful analogy, think of a pastoral field, extending farther than the eye can see. The rational numbers are the blades of grass and the soil that surrounds and is interspersed with these blades is the set of irrational numbers.

[†] Many attempts have been made to plot the Diriclet function such as https://www.google.com/search?q=dirichlet+function+image&hl=en&tbo=u&tbm=isch&source=univ&sa=X&ei=LNHEUIuELq322QWgxYGIAw&ved=0CCsQsAQ&biw=1151&bih=849

[‡] One of the reasons that Lebesgue measure is so popular as a didactic tool is that it is the natural counterpart to Riemann integration. From the Riemann perspective, the base of the rectangle $\Delta x$ is always the same, regardless of the integrand. Defining Lebesgue measure as the length of an integral, creates a firm relationship with Riemann integration,setting a solid foundation for the demonstration that Riemann integration and Lebesgue integration are equivalent. Of course Lebesgue integration theory is much more general in that it permits many other types of measures other than interval length.

Fortunately the Lebesgue integral equals the familiar Riemann integral when the number of discontinuities is finite,[*] bringing the familiar findings of the Riemann integral as area under the Lebesgue roof. However, the Lebesgue integral and its larger measure theory construct permits the integration of many functions that had simply been intractable under Riemann.

## Purpose of the measure theory integral

If we are to approach integration from the perspective of measure theory we will need some notation. The good news is that all that is required is that we re-task the "integral" sign. We let $\int$ announce our intention to measure a set.

For us the concept of integrating a function in measure theory is easy. Measuring a set is the same as integrating an indicator function. However, in order to carry this out, we need a metric, or measure.

## Examples of measures

We have already formally developed this metric's requirements. However, there are many metrics that meet these requisites, giving us substantial freedom in formulating unique ones. In general, we will define this metric as $\mu$ and denote $\mu(A)$ as "the measure of the set *A*". If this metric conforms to the following a) it is real valued and nonnegative, b) $\mu(\varnothing)=0,$ c) if $A\subset B,$ then $\mu(A)\leq\mu(B),$ and d) $\mu\left(\bigcup_n A_n\right)\leq\sum_n\mu(A_n)$ [†] then the metric is a measure.

For example let's say that we have an $(\Omega,\Sigma)$ collection and we propose that the metric of a set $A\subset\Sigma$ is simply the number of elements of the set (so called "counting measure"). This simple counting procedure meets all four properties above, qualifying it as a measure that we may delineate as $\mu(A)$.

Another useful example of a metric that meets these criteria and therefore is a measure is to let $(\Omega,\Sigma)$ be all intervals on the real line, and define $\mu(A)=b-a$ the length of the interval *A*. We have seen that this particular measure is defined as "Lebesgue measure".

## Indicator Functions in Lebesgue Integrals

Now, with the triplet $(\Omega,\Sigma,\mu),$ where $\mu(A)$ is defined for any $A\subset\Sigma,$ we can define the integral of the indicator function $e(\omega_i)=1_{\omega_i\subset A}$ as

$$\int_\Omega 1_{\omega_i\subset A}\,d\mu=\mu(A).$$

Thus the integral of the indicator function over the entire sample space $\Omega$ is defined as the measure of the set on which the indicator function is based.

[*] One way to think of this is that since Riemann uses the rectangle approximation, we adapt the Lebesgue construct by using the Lebesgue measure of the base of the rectangle as simply the rectangle's length on the *x*-axis. A more formal proof is provided here

[†] This last property it called countable additivity.

How does this process work? We might think of the source material or the domain of the integral as all elements contained in $\Omega$. The integral then inspects these elements in order to determine which ones are contained in the set *A*. The "mass" or the "value" of the accumulated sets identified to be in the set *A* is the measure of *A*, $\mu(A)$. In some sense, the integral is a sifter of elements, tagging and weighing them, assigning nonzero value to only those sets in *A*. By this same reasoning, we can write this as $\int_{\Omega} 1_{\omega_i \subset A}\, d\mu = \int_{A} d\mu \; = \mu(A)$.[*] This is a very different perspective then the Riemann "rectangle" approach to integration.

However, we must be explicit about the measure that we are using because different measures provide different results. For example, let $(\Omega, \Sigma)$ be a sample space/ σ-algebra pair reflecting the subjects in a neurology clinic. We could assign the measure $\mu_1(A)$ as the number of males in the set *A* where $A \subset \Sigma$. This would meet our definition of measure, permitting the creation of $(\Omega, \Sigma, \mu_1)$ and $\int_{\Omega} 1_{\omega_i \subset \Omega}\, d\mu_1 = \mu_1(\Omega)$ would reflect the number of males in the neurology clinic. Alternatively, we could assign the measure $\mu_2(A)$ as the total number of stroke victims in set $A$. This also meets our criteria of a measure, and by creating $(\Omega, \Sigma, \mu_2)$ permitting the computation $\int_{\Omega} 1_{\omega_i \subset \Omega}\, d\mu_2 = \mu_2(\Omega)$ as the total number of stroke victims in the neurology clinic.

Note that the indicator function in the above examples stayed the same. All that changed was the measure against which the indicator function was assessed. The freedom that comes from selecting a measure requires us to be explicit about the measure that we have selected.

In classic integration, this notion of the variable with which the measure is taken "with respect to" is reflected in the omnipresent expression "$dx$" e.g., in the expression $\int_a^b e^{-x} dx.$ We will adapt that and use the expression $d\mu$ which we will interpret as "with respect to the measure $\mu$" with the understanding that we will be explicit about its choice.

In order to examine the role of indicator functions in the Lebesgue integral, let's define two such functions. The first is $e_M(\omega_i) = 1_{\omega_i \subset M}$, the indicator function that identifies males, where *M* is the set of all males in the clinic. Also, we define $e_S(\omega_i) = 1_{\omega_i \subset S}$ as the indicator function where *S* is the set of all stroke victims in $\Omega$. Lets also let $\mu_m$ be the measure that counts males, and the measure $\mu_s$ be the measure that counts strokes.

Let's now consider the four integrals,

[*] This follows from

$$\int_{\Omega} 1_{\omega_i \subset A}\, d\mu = \int_{A} 1_{\omega_i \subset A}\, d\mu + \int_{A^c} 1_{\omega_i \subset A}\, d\mu = \int_{A} 1_{\omega_i \subset A}\, d\mu$$
$$= \int_{A} d\mu = \mu(A).$$

$$\int_{\Omega} 1_{\omega_i \subset M}\, d\mu_m$$
$$\int_{\Omega} 1_{\omega_i \subset M}\, d\mu_s$$
$$\int_{\Omega} 1_{\omega_i \subset S}\, d\mu_m$$
$$\int_{\Omega} 1_{\omega_i \subset S}\, d\mu_s$$

The first integral, collects all of the males in $\Sigma$ and then measures that set. Since measure $\mu_m$ measures males, then we write

$$\begin{aligned}\int_{\Omega} 1_{\omega_i \subset M}\, d\mu_m &= \int_{M} 1_{\omega_i \subset M}\, d\mu_m + \int_{M^c} 1_{\omega_i \subset M}\, d\mu_m \\ &= \int_{M} 1_{\omega_i \subset M}\, d\mu_m = \mu_m(\Omega),\end{aligned}$$

the total number of males. The second integral collects all males in the clinic and then identifies the number of strokes in this set. Thus, $\int_{\Omega} 1_{\omega_i \subset M}\, d\mu_s = \mu_s(M)$ or the number of strokes given the participant is a male. Similarly, $\int_{\Omega} 1_{\omega_i \subset S}\, d\mu_m = \mu_m(S)$, the number of males given the patient has had a stroke, and $\int_{\Omega} 1_{\omega_i \subset S}\, d\mu_s = \mu_s(S) = \mu_s(\Omega) = 1$.

This exercise justifies the differentiation between the set of interest and the measure in the integral sign (Figure 3).

<<Figure 3>>

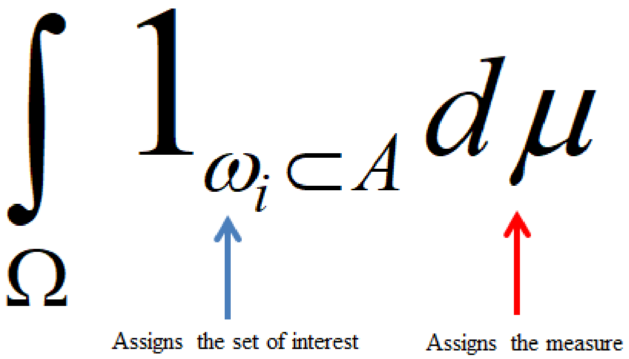

$\int_\Omega 1_{\omega_i \subset A}\, d\mu$
Assigns the set of interest
Assigns the measure

**Figure 3.** Two principal parts of the integrand of the Lebesgue integral. The indicator Assigns the set, and the measure notation tells us how to value or weigh the set.

Using this as background, what does $\int f\, d\mu$ mean? In order to compute this we will need both a function and a measure.

As an example, let $(\Omega, \Sigma)$ be the sample space and σ-algebra of the number of patients in a hospital. Define the measure $\mu$ of a set contained in $\Sigma$ as the number of patients with renal disease in that set. Then we can identify all patients in the hospital with renal disease as $\mu(\Omega) = \int_\Omega d\mu.$

Now specify the set $A \subset \Sigma.$ Then $\mu(A)$ is the number of patients in set $A$ with renal disease and $\int_\Omega 1_{\omega_i \subset A}\, d\mu = \mu(A).$ Let's also define $R$ as the set of all patients with renal disease. Then $\int_\Omega d\mu = \mu(\Omega) = \mu(R)$ since $\mu(\Omega) = \int_\Omega d\mu = \int_R d\mu + \int_{R^c} d\mu = \int_R d\mu = \mu(R).$ In addition, if we define the indicator function $e(\omega_i) = 1_{\omega_i \subset R}$ then this is a measurable function for each of the $\omega_i$ patients in the hospital and $\int_\Omega 1_{\omega_i \subset R}\, d\mu = \mu(R) = \mu(\Omega).$

Let's now develop a function defined on an element $\omega_i \subset \Sigma.$ If we assume that serum creatinine values are available we can define the measurable function $f(\omega_i)$ as the serum creatinine value for the $\omega_i^{th}$ patient. We are now poised to ask the question, on any set $A \subset \Sigma,$ what does $\int_A f(\omega_i)\, d\mu$ look like?

Let's begin by writing $\int_A f(\omega_i)\, d\mu = \int_\Omega 1_{\omega_i \subset A}\, f(\omega_i)\, d\mu.$ The indicator function in the integrand directs us to focus on the set $A$. Then, on that set, the integrand doesn't just accumulate the number of patients with renal disease, but it also accumulates their creatinine values. Thus $\int_A f(\omega_i)\, d\mu$ is the sum of the creatinine values of only those patients with renal disease in set $A$.[*]

[*] We can also write $\dfrac{\int_A f(\omega_i)\, d\mu}{\int_A d\mu}$ as the average creatine value for patients with renal disease in set $A$.

Note the contribution of $f(\omega_i)$ to the product $f(\omega_i)d\mu$. The notation $d\mu$ denotes that we assign value to a member of the set $\omega_i$ by whether or not they have renal disease. However the introduction the product $f(\omega_i)d\mu$ tells us that we must modify this process. While we must remain focused on whether the $\omega_i^{th}$ subject has renal disease, we assign the value of their creatinine value, $f(\omega_i)$. In some sense this product creates a new measure, one that assigns not just dichotomous 0 – 1 weight to set membership but instead that member's actual creatinine value (accumulating creatinine over a set satisfies the four properties of measure defined earlier). It is this modification of the measure that is credited to Johann Stieltjes, leading to the description of this process as not just Lebesgue integration, but Lebesgue-Stieltjes integration (Figure 4).

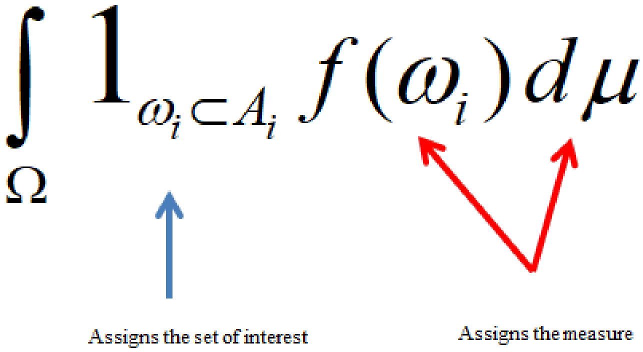

$\int_{\Omega} 1_{\omega_i \subset A_i} f(\omega_i)d\mu$
Assigns the set of interest
Assigns the measure

**Figure 4.** Lebesgue-Stieltjes integral. The function in the integrand modifies the measure.

## Example: aerosols

Let's apply these concepts to the content of aerosols. Aerosols are microscopic combinations of particles in the air. They comprise the basic element of cloud formation.

The sun, heating the ocean's surface, converts water to its gas phase where it rises into the atmosphere. There this water vapor comes in contact with and condenses on each aerosol to form a droplet of water and a dust particle. When they gather together, a cloud is formed.

Hundreds of years ago, the only aerosols in the atmosphere were salt particles from the ocean, small amounts of soot from volcanoes, or tiny bits of soil. The industrial revolution has increased the number of aerosols in the environment.

Define $\Omega$ as all of the cubic mm of space in the Earth's atmosphere, $\Sigma$ its σ–algebra and $\omega_i$ an individual element of $\Omega$. Let the measure $\mu(A)$ be the number of aerosols in a volume of space $A \subset \Sigma$. Let $f(\omega_i)$ be the pollutant content of an aerosol $\omega_i$. Then what is $\int_{\Omega} 1_{\omega_i \subset A} f(\omega_i)d\mu$?

Using the architecture described in Figure 4, we see that our concern lines with a defined sub-volume of space (which need not be contiguous) $A$. Recall that $\mu(A)$ is the aerosols contained in the cloud content of the volume of space $A$. Then $f(\omega_i)d\mu$ accumulates the pollutant content in the cloud that inhabits a cubic cc of space, and $\int\limits_{\Omega} 1_{\omega_i \subset A} f(\omega_i) d\mu$ is therefore the pollutant content or load in all clouds in the volume of space $A$.

As we have seen before, we can define a new measure which would be the pollutant content of aerosols in a volume of space $A \subset \Sigma$ as $\mu_f(A)$ and the integral becomes $\int\limits_{\Omega} 1_{\omega_i \subset A} d\mu_f = \int\limits_{A} d\mu_f$. Here the function $f(\omega_i)$ is subsumed into the measure using the Lebesgue-Stieltjes approach.

Probability Foundations

Advanced Probability

# Lebesgue Integration Theory and the Bernoulli Distribution

Having discussed the motivation and development of integration theory, we are now ready to apply it to probability. We will begin with the most elementary distribution in probability, the Bernoulli distribution.

Prerequisites
Basics of Bernoulli Trials – The Bernoulli Distribution
Measure Based Integration

## Notation and procedures

We begin by letting

$$\int_{\Omega} d\mathbf{P}$$

stand for not just an integral, but for a procedure. The term $d\mathbf{P}$ stipulates that we want to accumulate this measure with respect to a probability distribution $\mathbf{P}$. The recognition that we could let the integrand be with respect to sets other than those that comprise the real number line, permitting "$d\mathbf{P}$" rather than "$dx$" was developed by Stieltjes, earning the integral the descriptor "Lebesgue-Stieltjes"

So, one thing that we know from probability is that the total accumulation of probability over the sample space Ω should be one, or

$$\int_{\Omega} d\mathbf{P} = 1$$

In order to attempt this, we must have a measure $\mathbf{P}$. Let's first start with simplist  distribution. Define the random variable $Y$ as

$$Y = 1_{y=1}$$

Here, $Y$ takes on only one value (the value 1) with probability 1. One could be forgiven for thinking there is no probability attached to this "random" variable at all, but our goal here is to choose the simplest distribution possible to ease our first work with the use of Lebesgue integration theory. How do we demonstrate that $\int_{\Omega} d\mathbf{P} = 1$ in this example?

There are two equivalent ways that we can show this. The first is to see that there are two collections of outcomes that contribute to *Y*. The first is the value *Y* = 1 which has probability 1. All other values have probability zero. We therefore can write

$$\int_{\Omega} d\mathbf{P} = 1\mathbf{P}\left[\{all\ sets\ where\ Y = 1\}\right] + 0\mathbf{P}\left[\{all\ sets\ where\ Y = 0\}\right] = 1$$

This is the heart of the Lebesgue theory of integration. We do the integration simply by combining values of the random variable whose probability (or measure) is the same, and then just sum the probability.

An equivalent approach is to see that *Y* is a real valued random variable, so we lose nothing by writing $\int_{\Omega} d\mathbf{P} = \int_{-\infty}^{\infty} d\mathbf{P} = \int_{-\infty}^{\infty} 1_{y=1}$. The integral $\int_{-\infty}^{\infty} 1_{y=1}$ states that we will measure the $(-\infty, \infty)$ interval using the (very simple) measuring tool $1_{y=1}$. This integral we compute by beginning at -∞ and, moving in the positive direction, accumulate probability (or measure) as we go. There is no accumulation until we get to the value *y* = 1 where there is a spike in the probability. We accumulate this spike of "1" and continue to positive infinity where again we make no additional accumulations. Thus the value of $\int_{-\infty}^{\infty} 1_{y=1} = 1$. The failure of the Riemann-Stieltjes integral* here is why historically probability distributions have been divided into continuous and discrete distributions, since the area under the curve concept breaks down when probability is concentrated at one point. However, the Lebesgue approach will allow us to manage both types of distributions equally well.

## Application to bernoulli distribution

Now, let's turn to the classic Bernoulli distribution. Here, for $p$ a known constant, $0 < p < 1$, we have $\mathbf{P}[X] = p1_{x=1} + (1-p)1_{x=0}$. We now know that we can write

$$\int_{\Omega} d\mathbf{P} = \int_{\Omega} \left(p1_{x=1} + (1-p)1_{x=0}\right) = p\int_{\Omega} 1_{x=1} + (1-p)\int_{\Omega} 1_{x=0}$$
$$= p + (1-p) = 1.$$

Again, we can "measure" this set two ways. One way is to aggregate the real line into two sets. The first is all values of $x$ where the probability $p$. The second is all values of $x$ where the probability is $1-p$. The second is, to write as before $\int_{\Omega} d\mathbf{P} = \int_{-\infty}^{\infty} d\mathbf{P}$, and to accumulate measure or probability moving from $-\infty$ positively. During this process, we pick up no probability until we

[*] It is important to note that the common Riemann-Stiljes integral fails in this simple example, since it would produce $\int_{-\infty}^{\infty} 1_{y=1} = 0$ since the function has no 'area under the curve".

get to $x=0$ where we pick up probability $1-p$, then accumulate no more probability until we get to 1 where we accumulate at this one point probability or measure $p$, accumulating no more measure after that.

Note again the failure of Riemann integral. Since it requires a rectangle to compute the area, and there are no rectangles defined by this measuring tool defined by the Bernoulli distribution $\mathbf{P}[X]=p1_{x=1}+(1-p)1_{x=0}$, the Riemann integral would be zero.

Monotone Convergence Theorem
Some Classic Measure Theory Results
Asymptotics
Tail Event Measure

Probability Foundations
Basic Properties of Probability
Counting Events

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables

# Basic Properties of the Lebesgue-Stieltjes Integral

We have developed enough of an introduction to the Lebesgue-Stieltjes integral to begin to establish some of its basic principles. For each of these principles, we assume that we have a triplet $(\Omega, \Sigma, \mu)$ focusing on any set $A \subset \Sigma.$ We will begin with the assertion that we can characterize an integrable function $f(\omega_i)$ as the sum of indicator functions $\sum_{k=1}^{n} \alpha_k 1_{\omega_i \subset A_k}$ for *n* large enough where the sets $A_k$ are mutually exclusive, and that $\int_A f(\omega_i) d\mu = \int_A \sum_{k=1}^{n} \alpha_i 1_{\omega_i \subset A_k \cap A} = \sum_{k=1}^{n} \alpha_k \mu(A_k \cap A).$ We will first limit these principals to these functions, then generalize beyond them using the Monotone Convergence Theorem.

## Properties of integrals of simple functions

The following represent the major properties of the Lebesgue-Stieltjes integral. They are worth careful study with the view to understanding each of the steps in their motivation and development.

**Property 1: If** $f(\omega_i) = 0,$ **for all** $\omega_i \subset A$ **then** $\int_A f(\omega_i) d\mu = 0.$

The key to this demonstration is to see that if $f(\omega_i) = \sum_{k=1}^{n} \alpha_k 1_{\omega_i \subset A_k}$ and $f(\omega_i) = 0$ for all $\omega_i,$ then each $\alpha_k$ for which $\omega_i \subset A_k$ must be zero.* Then

* Let $(\Omega, \Sigma)$ represent community membership for each of $\omega_i$ citizens. Define $f(\omega_i) = \alpha_1 1_{\omega_i \subset M} + \alpha_2 1_{\omega_i \subset F}$ where $M \subset \Sigma$ is the set of males and $F \subset \Sigma$ is the set of females. Then since $M \cap F = \varnothing,$ $\omega_i$ can be a member of either $M$ or $F.$ For males $f(\omega_i) = \alpha_1 1_{\omega_i \subset M}$ but $f(\omega_i) = 0$ implies that $\alpha_1 = 0.$ A similar argument can be made for females, revealing that $\alpha_2 = 0.$The key part of this proof is constructing the simple function from disjoint sets.

$$\int_A f(\omega_i)d\mu = \int_A \sum_{k=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k} d\mu = \sum_{k=1}^n \alpha_k \mu(A_k) = \sum_{k=1}^n (0)\mu(A_k) = 0.$$

**Property 2 : For all $\omega_i \subset A,$ and constant $c > 0,$ then**
$\int_A cf(\omega_i)d\mu = c\int_A f(\omega_i)du.$
We write

$$\int_A cf(\omega_i)d\mu = \int_A c\sum_{k=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k \cap A} d\mu = \int_A \sum_{k=1}^n c\alpha_k \mathbf{1}_{\omega_i \subset A_k \cap A} d\mu$$

$$= \sum_{k=1}^n c\alpha_k \mu(A_k \cap A) = c\sum_{k=1}^n \alpha_k \mu(A_k \cap A)$$

$$= c\int_A \sum_{i=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k \cap A} d\mu = c\int_A f(\omega_i)d\mu$$

**Property 3: If for all $\omega_i \subset A,$ $f(\omega_i) > 0,$ then $\int_A f(\omega_i)d\mu \geq 0.$**

Begin with $f(\omega_i) = \sum_{k=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k}$. Since the family of sets $A_k$ are mutually exclusive, then the element $\omega_i \subset A_k$ falls in one and only one of the $A_k$. Call this particular $A_k$ the set $A_k^*$. Then $f(\omega_i) = \sum_{k=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k} = \alpha_k^*$. However, $f(\omega_i) = \alpha_k^* \mathbf{1}_{\omega_i \subset A_{k^*}} > 0.$ Thus $\int_A f(\omega_i)d\mu = \int_A \sum_{k=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k} d\mu = \int_A \alpha_k^* \mathbf{1}_{\omega_i \subset A_k \cap A} = \alpha_k^* \mu(A_k^* \cap A) \geq 0.$ Now, we know that $\alpha_k^* > 0,$ so $\alpha_k^* \mu(A_k^* \cap A)$ is positive if $\mu(A_k^* \cap A) > 0$ but zero if the set $A_k^* \cap A$ has zero measure.

**Property 4. If for all $\omega_i \subset A,$ $f(\omega_i) = g(\omega_i),$ then $\int_A f(\omega_i)d\mu = \int_A g(\omega_i)d\mu.$**

In order to demonstrate this, observe that we can finely partition the family of $A_k$ (which are mutually exclusive) such that the element $\omega_i \subset A_k$ falls in one and only one of the $A_k$. Call this particular $A_k$ the set $A_k^*$. Then $f(\omega_i) = \alpha_k^* = g(\omega_i)$. Thus

$$\int_A f(\omega_i)d\mu = \int_A \sum_{k=1}^n \alpha_k \mathbf{1}_{\omega_i \subset A_k} d\mu = \int_A \alpha_k^* \mathbf{1}_{\omega_i \subset A_k \cap A} = \int_A g(\omega_i)d\mu.$$

Note however, that the converse is not true. If, $\int_A f(\omega_i)d\mu = \int_A g(\omega_i)d\mu$ then it does not always follow that the functions are equal. For an example, set $f(\omega_i) = 0$ for all $\omega_i \subset A.$ Then $\int_A f(\omega_i)d\mu = \int_A 0d\mu = 0.$ Now, create sets $B_1$ and $B_2$ such that $B_1 \cap B_2 = \varnothing, B_1 \cup B_2 = A,$ and

$\mu(B_1)=\mu(B_2)$. Let $g(\omega_i)$ be positive for $\omega_i \subset B_1$ and its negative value for all $\omega_i \subset B_2$ such that $\int_{B_2} g(\omega_i)d\mu = -\int_{B_1} g(\omega_i)d\mu$. Then

$$\begin{aligned}\int_A g(\omega_i)d\mu &= \int_{B_1} g(\omega_i)d\mu + \int_{B_2} g(\omega_i)d\mu \\ &= \int_{B_1} g(\omega_i)d\mu - \int_{B_1} g(\omega_i)d\mu = 0.\end{aligned}$$

Thus $\int_A f(\omega_i)d\mu = \int_A g(\omega_i)d\mu = 0$ but clearly $0 = f(\omega_i) \neq g(\omega_i)$ for all of $\omega_i \subset A.$

**Property 5: If for all $\omega_i \subset A,\ f(\omega_i) > g(\omega_i),$ then $\int_A f(\omega_i)d\mu > \int_A g(\omega_i)d\mu.$**

This stands to reason, since, if the integrand of one is greater than the integrand for the other for all $\omega_i \subset A,$ then the integral (which is just an accumulation of the integrands) should also be greater.

Keeping in mind that both *f* and *g* are simple functions, we can write $f(\omega_i)=\sum_{k=1}^{n}\alpha_k 1_{\omega_i \subset A_k}$ and $g(\omega_i)=\sum_{k=1}^{n}\beta_k 1_{\omega_i \subset A_k}$. Then observe that we can finely partition the family of $A_k$ (which are mutually exclusive) such that the element $\omega_i \subset A_k$ falls in one and only one of the $A_k$. Call this particular $A_k$ the set $A_k^*$. Then $f(\omega_i)=\alpha_k^*$ and $g(\omega_i)=\beta_k^*$. Since $f(\omega_i)>g(\omega_i)$ we can write $\alpha_k^* > \beta_k^*$. Thus

$$\begin{aligned}&\int_A f(\omega_i)d\mu = \int_A \sum_{k=1}^{n}\alpha_k 1_{\omega_i \subset A_k \cap A} = \int_A \alpha_k^* 1_{\omega_i \subset A_k^* \cap A} = \alpha_k^*\mu\left(A_k^* \cap A\right) \\ &< \beta_k^*\mu\left(A_k^* \cap A\right) = \int_A \beta_k^* 1_{\omega_i \subset A_k^* \cap A} = \int_A \sum_{k=1}^{n}\beta_k 1_{\omega_i \subset A_k \cap A} = \int_A g(\omega_i)d\mu.\end{aligned}$$

These are examples of properties of measurable functions which can be precisely categorized as simple functions.

**Property 6: If for all $\omega_i \subset A,\ \ f(\omega_i)$ and $g(\omega_i)$ exists, then $\int_A \left(f(\omega_i)+g(\omega_i)\right)d\mu = \int_A f(\omega_i)d\mu + \int_A g(\omega_i)d\mu.$**

We define $f(\omega_i)=\sum_{k=1}^{n}\alpha_k 1_{\omega_i \subset A_k}$ and $g(\omega_i)=\sum_{k=1}^{n}\beta_k 1_{\omega_i \subset A_k}$. Then

$$\begin{aligned}&\int_A \left(f(\omega_i)+g(\omega_i)\right)d\mu = \int_A \left(\sum_{k=1}^{n}\alpha_k 1_{\omega_i \subset A_k \cap A} + \sum_{k=1}^{n}\beta_k 1_{\omega_i \subset A_k \cap A}\right)d\mu \\ &= \int_A \sum_{k=1}^{n}(\alpha_k+\beta_k)1_{\omega_i \subset A_k \cap A}\, d\mu = \sum_{k=1}^{n}(\alpha_k+\beta_k)\mu(A_k \cap A) \\ &= \sum_{k=1}^{n}\alpha_k\mu(A_k \cap A) + \sum_{k=1}^{n}\beta_k\mu(A_k \cap A) = \int_A \sum_{k=1}^{n}\alpha_k 1_{\omega_i \subset A_k} + \int_A \sum_{k=1}^{n}\beta_k 1_{\omega_i \subset A_k} \\ &= \int_A f(\omega_i)d\mu + \int_A g(\omega_i)d\mu.\end{aligned}$$

The same is true for taking the difference of two functions.

**Property 7: If for all** $\omega_i \subset A,\ \ f(\omega_i)$ exists and integrable, then $\int\limits_A \frac{1}{f(\omega_i)}d\mu$ is integrable so long as $f(\omega_i) \neq 0$ on a set of positive measure.

Since $f(\omega_i)$ is integrable, we know that $f(\omega_i)=\sum\limits_{j=1}^{n}\alpha_j 1_{\omega_i \subset A_j}$ and we can write naively that $\frac{1}{f(\omega_i)}=\frac{1}{\sum\limits_{j=1}^{n}\alpha_j 1_{\omega_i \subset A_j}}$. However, since the simple function is in cardinal form and therefore the set $\{A_i\}$ is made up of pairwise disjoint sets, then for each $\omega_i$ we can write $\frac{1}{f(\omega_i)}=\frac{1}{\alpha_j^*}1_{\omega_i \subset A_j^*}$ where the * signifies the single set $A_j^*$ in which $\omega_i$ resides. This indicator function exists as long as $\alpha_j^*$ is not equal to zero where the $\mu\left(A_j^*\right)>0$. If $\mu\left(A_j^*\right)=0$, then the value of the function on the set is immaterial since this value is not considered on a set of measure zero.

**Property 8: If for all** $\omega_i \subset A,\ \ f(\omega_i)$ **and** $g(\omega_i)$ **exists and are integrable, then** $\int\limits_A f(\omega_i)g(\omega_i)d\mu$ **is integrable.**

Since $f(\omega_i)$ is integrable, we know that $f(\omega_i)=\sum\limits_{j=1}^{n}\alpha_j 1_{\omega_i \subset A_j}$ and $\int\limits_A f(\omega_i)=\sum\limits_{j=1}^{m}\alpha_j\mu\left(A_j\right)$. Also, $g(\omega_i)=\sum\limits_{k=1}^{m}\beta_k 1_{\omega_i \subset B_k}$ and $\int\limits_A g(\omega_i)=\sum\limits_{k=1}^{m}\beta_k\mu(B_k)$. If the sets $\{A_j\}$ and $\{B_k\}$ are in cardinal form such that the $\{A_j\}$ and $\{B_k\}$ are each pairwise disjoint, then we can write

$$f(\omega_i)g(\omega_i)=\left[\sum_{j=1}^{n}\alpha_j 1_{\omega_i \subset A_j}\right]\left[\sum_{k=1}^{m}\beta_k 1_{\omega_i \subset B_k}\right]=\sum_{j=1}^{n}\sum_{k=1}^{m}\alpha_j\beta_k 1_{\omega_i \subset A_j \cap B_k}$$

This is itself a simple function in cardinal form. Since the sets $\{A_j\}$ and $\{B_k\}$ are measurable, then so are their intersections. Thus the integral exists. Furthermore, since $\{A_j\}$ and $\{B_k\}$ are in cardinal form, their intersections are also in cardinal form. Thus,

$$\begin{aligned}\int\limits_A f(\omega_i)g(\omega_i)d\mu &= \int\limits_A \sum_{j=1}^{n}\sum_{k=1}^{m}\alpha_j\beta_k 1_{\omega_i \subset A_j \cap B_k}d\mu \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m}\alpha_j\beta_k\mu\left(A_j \cap B_k\right).\end{aligned}$$

Combining properties 7 and 8, we can now deduce that if $\omega_i \subset A,\ \ f(\omega_i)$ and $g(\omega_i)$ exists and are integrable, then $\int\limits_A \frac{f(\omega_i)}{g(\omega_i)}d\mu$ is integrable.

## Integration using simple functions

From the developments of the last chapter, we can integrate set and simple functions using Lebesgue integration. This is already a solid step forward from Riemann integration since many functions provide no helpful mesh around which we can build a Riemann integral.

However, for Lebesgue integration to be of real value we must be able to integrate functions that themselves are not simple.

Ironically, the tool that we use for this extension is the simple function. Assume that we have a sample space, σ-algebra and a measure $(\Omega, \Sigma, \mu)$. Let **S** be the set of all simple functions. Then, for a function $f(\omega_i)$ and a measure $\mu$ we define the Lebesgue integral $\int_A f(\omega_i)\,d\mu$ as

$$\int_A f(\omega_i)\,d\mu = \sup_{\substack{s(\omega_i)\subset S \\ s(\omega_i)\le f(\omega_i)}} \int_A f(\omega_i)\,d\mu$$

Thus, the Lebesgue integral is the supremum of the integral over all simple functions that are $\le f(\omega_i)$. How does this take place? Here is the process:

1 - identify all simple functions that are $\le f(\omega_i)$.

2 – for each of these simple functions $s_j(\omega_i) = \sum_{j=1}^{n_j} \alpha_j 1_{\omega_i \subset A_j}$, we must compute its integral on the set $A$, $\int_A s(\omega_i)\,d\mu = \sum_{j=1}^{n_j} \alpha_j \mu(A_j \cap A)$.

3- identify the supremum of all of these integrals and that supremum is $\int_A f(\omega_i)$.

This computation has something of the feel of the "inner measure" concept that we discussed earlier. Here we first find all of the simple functions $s(\omega_i)$ where $s(\omega_i) \le f(\omega_i)$. This activity, in and of itself is a huge undertaking, since there are uncountably many such functions.

We then compute the integral of each of these simple functions, remembering that for the simple function $\sum_{k=1}^{n} \alpha_k 1_{\omega_i \subset A_k \cap A} \rightarrow \int \sum_{k=1}^{n} \alpha_k 1_{\omega_i \subset A_k \cap A}\,d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k)$.

The final task is to identify the sup of these integrals

The foundation of this argument is that a general function $f(x)$ even though itself is not a simple function, can be linked to simple functions. Also note that the role of $s(\omega_i)$ is not to approximate $f(\omega_i)$; it is not $\int_A f(\omega_i)\,d\mu = \int_A \sup_{\substack{s(\omega_i)\subset S \\ s(\omega_i)\le f(\omega_i)}} f(\omega_i)\,d\mu,$

that we seek, but $\int_A f(\omega_i)\,d\mu = \sup_{\substack{s(\omega_i)\subset S \\ s(\omega_i)\le f(\omega_i)}} \int_A f(\omega_i)\,d\mu$

We are approximating the measure of the set of $A$ directly, not indirectly by approximating $f(\omega_i)$ and then integrating this approximation.

Another way to say this is that every Lebesgue integrable function is itself the limit of a monotonically increasing sequence of simple functions. Thus, the process is to identify the simple function which most closely approximates $f(x)$ from below. Since that simple function is Lebesgue integrable, we integrate it, and have the Lebesgue integral for $f(x)$.

Lebesgue integrable functions, just like Riemann integrable functions must both be built; however, they are built from different substrates. As we have seen, Riemann integrable functions are built up from rectangles. Lebesgue integrable functions are built up from simple functions.

## Example – Protein synthesis

As an example, lets create a function based on the ribosome. The ribosome is one of the smallest organelles contained in a cell, but it plays a critical role in the synthesis of proteins. The ribosome is itself composed of ribonucleic acids and protein divided into two components.

The smaller of these two components contains the nucleic acid and is prepared to house the messenger RNA (mRNA) that comes from the cell nucleus. This mRNA strand holds the sequence of instructions to build the protein structure chain one amino acid at a time. The larger component functions more like an enzyme; it is a protein catalyst providing the energy for the lengthening of the nascent protein to take place, amino acid by amino acid.

The process for protein development begins in the cell's nucleus, where a sequence of DNA is decoded into messenger RNA (mRNA). This mRNA strand, once extruded from the nucleus into the cytoplasm, then attracts the two inactive components of the ribosome which come together around the mRNA strand so that the mRNA is aligned along their interface.

At this point transfer RNA (tRNA), a three codon sequence that is attached to a single amino acid, competes for the mRNA sites. Once the correct tRNA is chosen, the larger catalytic component generates the chemical energy necessary to put the amino acid into position in the growing protein chain. When complete, the protein is released into the cytoplasm and the two components of the ribosome fall away from each other, making themselves available for the construction of other proteins.

The human cell has on average 30 million ribosomes, many of which are embedded in the rough endoplasmic reticulum that surrounds the nucleus. Each ribosome can build a 200 amino acid sequence in a minute, an amino acid addition rate of three to four amino acids per second.

Suppose we wish to find all peptides on a sphere of radius *L* microns from the center of the ribosome. Is this a measurable function? Can we define a function that will do this for us, and if so, what does its integral mean?

First, we must ask how would we manage this parametrically.

Let's define $\omega_i$ is the *i*th peptide identified. Define $\Omega$ as the sample space containing all peptides surrounding the ribosomes, and $\Sigma$ as its σ–algebra. Here, we want to map $\omega_i{}'s$ location using as a reference point a sphere of radius *L*. Either the peptide is on the surface of this sphere or it is not. This meets the criterion of a measurable function[*]. The challenge is to find a sequence of monotonically increasing simple functions that has a limit of our desirable function $f(\omega_i)$.

There are many such sequences. Consider the elementary function $e_n(\omega_i) = 1_{\omega_i \subset \left(L-\frac{1}{n},\, L-\frac{1}{n+1}\right)}$. The argument $\omega_i \subset \left(L-\frac{1}{n}, L-\frac{1}{n+1}\right)$ just traps the peptide in a spherical annulus whose diameter is between $L-\dfrac{1}{n}$ and $L-\dfrac{1}{n+1}$ microns taking on the value 1 if the peptide location is within this region and 0 if not.

Then $f_1(\omega_i) = 1_{\omega_i \subset \left(L-1,\, L-\frac{1}{2}\right)}$, and $f_2(\omega_i) = 1_{\omega_i \subset \left(L-1,\, L-\frac{1}{2}\right)} + 1_{\omega_i \subset \left(L-\frac{1}{2},\, L-\frac{1}{3}\right)}$. This function is 1 if either $\omega_i$ is located within 1 micron and $\frac{1}{2}$ micron below the surface of the sphere or within $\frac{1}{2}$

[*] Its real valued, maps a peptide to one and only one location and its preimage is a property of $\omega_i$, namely its location.

and $\frac{1}{3}$ micron. If we let $f_n\left(\omega_i\right)=\sum_{i=1}^{n} 1_{\omega_i \subset \left(L-\frac{1}{n},L+\frac{1}{n}\right)}$, then $f_n\left(w_i\right)$ is either equal to 0 or 1, depending on how close the location of peptide $\omega_i$ is to L. Thus, $\lim_{n\to\infty} f_n\left(\omega_i\right)=1=f\left(\omega_i\right)$. Thus, for each of the billions of peptides, $f\left(\omega_i\right)$ is either 0 or 1 but whatever the limit, $\lim_{n\to\infty} f_n\left(\omega_i\right)=f\left(\omega_i\right)$.

Suppose that we now defined $e_n\left(\omega_i\right)=k\left(1-\frac{1}{n}\right)1_{\omega_i\subset\left(L-\frac{1}{n},L-\frac{1}{n+1}\right)}$. Then, there is some value $n(\omega_i)$ such that $e_n\left(\omega_i\right)=k\left(1-\frac{1}{n\left(\omega_i\right)}\right)1_{\omega_i\subset\left(L-\frac{1}{n(\omega_i)},L-\frac{1}{n(\omega_i)+1}\right)}=k\left(1-\frac{1}{n\left(\omega_i\right)}\right)$. We may write

$$f_n\left(\omega_i\right)=\sum_{j=1}^{n} e_j\left(\omega_i\right)=\sum_{j=1}^{n} k\left(1-\frac{1}{n\left(\omega_i\right)}\right)1_{\omega_i\subset\left(L-\frac{1}{n(\omega_i)},L-\frac{1}{n(\omega_i)+1}\right)}, \quad \text{and}$$

$$\lim_{n\to\infty} f_n\left(\omega_i\right)$$

$$=\lim_{n\to\infty}\sum_{j=1}^{n} k\left(1-\frac{1}{n\left(\omega_i\right)}\right)1_{\omega_i\subset\left(L-\frac{1}{n(\omega_i)},L-\frac{1}{n(\omega_i)+1}\right)}=k\left(1-\frac{1}{n\left(\omega_i\right)}\right).$$

where $k$ is the value of the function when $\omega_i$ is on the sphere with radius $L$. Note that this is an increasing function, $f_n\left(\omega_i\right)\uparrow f\left(\omega_i\right)$. Note that since we do not assess the collection of $\omega_i$ in any particular order, this function's value jumps rapidly from 0 to $k\left(1-\frac{1}{n\left(\omega_i\right)}\right)$ and back depending $\omega_i$ and its location.

Note that our function $k\left(1-\frac{1}{n\left(\omega_i\right)}\right)$ is the limit of a monotonic sequence of simple functions, which is sufficient to build its Lebesgue integral. For this function to be integrable, we must define the measure of a spherical annulus. This poses no difficulty since we can define measure as "volume measure".

However, note that building the sequence of functions which is the foundation of $f\left(\omega_i\right)$ is a different process than identifying the measure.

Finally, since the measure is a three dimensional volume, it also has a Riemann integral; the use of rectangles (or in this case three dimensional rectangles) would be the geometric basis of the Riemann integral.

In this case the Lebesgue integral and Riemann integrals each exist and they equal to each other. It is the simultaneous construction of the integral as the limit of areas of rectangles (when they exists) and the integral as a sequence of monotonically increasing simple functions that is the basis of the equivalence of the Riemann and Lebesgue integral, a finding that will be proved later.

However, suppose we define a measure on $(\Omega,\Sigma)$ called $\phi$ and define $\varphi(\omega_i)$ as the number of times that the alanine-tyrosine-lysine triplet (ATLs) occurs in the peptide. Since peptides are separate and apart from each other, then it is easy to confirm that $\varphi(\omega_i)$ is also a measure.[*] Using our previous definition of $\lim_{n\to\infty} f_n(\omega_i)=k\left(1-\frac{1}{n(\omega_i)}\right)$ as the location of the peptide on the sphere, how can we interpret $\int_\Omega f(\omega_i)d\varphi$?

Our intuition leads us to think that $\int_\Omega f(\omega_i)d\varphi$ is the accumulation of the value $k\left(1-\frac{1}{n(\omega_i)}\right)$ multiplied by the ATL's in each peptide whose location is in that annulus of the sphere . However, that intuition assumes that $\lim_{n\to\infty}\int f_n(\omega_i)d\varphi=\int\lim_{n\to\infty} f_n(\omega_i)d\varphi=\int f(\omega_i)d\varphi,$ a set of equalities that is not always the case. The role of the monotone convergence theorem is to provide an important circumstance when these equalities are correct.

We know that continuous functions can be built up from rectangles. However, how many functions can be approximated by simple functions? As it turns out, every bounded measurable function is the nonnegative limit of a sequence of monotonically increasing simple functions. This we learn from the monotone convergence theorem.

Monotone Convergence Theorem
Some Classic Measure Theory Results
Basic Properties of Probability
Counting Events

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

[*] Clearly it is positive, its measure of the null set is 0, and the measure of peptide is greater than the measure of a subpeptide that it contains. Since the peptides are separate from each other, $\varphi\left(\bigcup_{n=1}^{\infty}\omega_i\right)=\sum_{n=1}^{\infty}\varphi(\omega_i)$.

# Monotone Convergence Theorem

## What does this theorem do for us?

What we have seen is that the basic properties of the Lebesgue integral are relatively straightforward to prove as long as the function $f$ is a simple function. However, many times if not most times our function is not a simple function. The function may be unidentifiable. Under what assumptions are such functions Lebesgue integrable?

## Prerequisites

Sequences of Sets
Sequences of Functions
Set Functions in Measure Theory
Basic Properties of the Lebesgue-Stieltjes Integral

## Lebesgue integration and the MCT

The monotone convergence theorem (MCT, also known as Beppo Levi's Theorem) answers that question for us. It tells us that, although the function *f* may not be a simple function, if it is the limit of an increasing sequence of simple functions, then the function will have an integral, and the integral itself will be defined in terms of the sequence of integrals of simple functions.

The monotone convergence theorem is a powerful tool because it permits us to link simple functions (that we know to be measurable) to non-simple functions. This link allows us to essentially transfer all of the properties of integration that we have derived for simple functions onto non-simple functions that are measurable. This greatly expands both the utility of simple functions as well as the universe of measurable functions for us.

Its precise statement follows:

**Monotone Convergence Theorem**

*Assume we have a measure space* $(\Omega, \Sigma, \mu)$ *and a function* $f$. *Assume that there is monotonically increasing sequence of simple functions* $f_n(\omega_i) = \sum_{k=1}^{n} \omega_i 1_{\omega_i \subset A_k}$ *for sets* $A_k \subset \Sigma$ *such that*

$\lim\limits_{n\to\infty} f_n(\omega_i) = f(\omega_i)$. *Then this limit function,* $f(\omega_i)$, *is Lebesgue integrable and* $\int\limits_A f(\omega_i)\,d\mu = \lim\limits_{n\to\infty}\int\limits_A f_n(\omega_i)\,d\mu.$

This is an important theorem and it behooves us to understand all of its facets.

Overall, the monotone convergence theorem tells us how to determine if a function $f$ is integrable and then how to carry out its integration. In order to do this we must find the simple function of which it is a limit, integrate a member of that sequence, and then take the limit of the integral.

At first blush, this may seem somewhat awkward . For example, consider the uncomplicated function $f(x) = \sin x - \cos x$ for $0 \le x \le \frac{\pi}{2}$. If we were to apply the monotone convergence theorem to this function, we would first be required to identify a sequence of simple functions that converges to $f(x)$. Then, once this function was identified, we would find its $n^{th}$ term. We would then integrate that term $f_n(x)$ and take the limit of this integral.

While this sequence of steps would be technically accurate, the fact is that it is much simpler to just go ahead with the direct integration of $f(\omega_i) = \sin(\omega_i) - \cos(\omega_i)$.

We are able to take this more direct approach because this function is Riemann integrable.* When the function is Riemann integrable it is commonly easier to just carry out the Riemann integration than to go through the simple function construct. So from one perspective the theorem does not appear to be of practical use.

## Utility of the monotone class theorem

However, the key to the utility of the monotone convergence theorem is seeing that many problems have solutions where the function is not Riemann integrable. Commonly the only function that we are exposed to is the Diriclet function, one that admittedly does not appear commonly in public health.

However examples that we have provided in protein synthesis, toxin measurement, aerosols, etc show how useful public health functions are not Riemann integrable. If all we have is the Riemann integration, these functions could not be integrable (in fact, we would likely not even know that they exists.

Knowledge of the MCT motivates us to see beyond Riemann integrable, “parametric” functions to functions that are based on complicated simple functions that the MCT guarantees will be Lebesgue integrable, in fact, tells us how to conduct the integration.

For example, the function that we built to capture a nephron’s ability to filter waste products in the blood is nothing but a sequence of increasing simple functions. It is not possible to know the functional form of $f(\omega_i) = \lim\limits_{n\to\infty} e_n(\omega_i)$ yet our interest resides in the integral of this limit. Fortunately, we do not have to know the form of the limit $f$ in order to integrate it, and that is the key contribution of the monotone convergence theorem. We simply compute $\lim\limits_{n\to\infty}\int\limits_A e_n(\omega_i)\,d\mu$ and by this theorem know that it is $\int\limits_A f(\omega_i)\,d\mu.$

* This process could be carried out under the monotone convergence theorem. In fact it is precisely the process of following the steps outlined by the monotone class theorem that demonstrates Riemann integrals are also Lebesgue integrable and have the same value.

The second point of the monotone convergence theorem is that it provides a condition for which we can pass the limiting argument through the integral sign without disturbing the equality. Thus, $\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu = \int_A \lim_{n\to\infty} f_n(\omega_i)\,d\mu = \int_A f(\omega_i)\,d\mu.$ The ability to pass the limit through the integral sign is not a property that we can take for granted, but it does operate under MCT control. [*]

## Proving the theorem

With this as background, we are now in a position to prove the monotone convergence theorem. A key concept that we will have to deal with is the process of passing a limit through an integral sign.

In order to be in a position to proof this we will develop a key concept of measures and sets. For example if a sequence of sets $A_n$ converges to a set $A$[†] then under what circumstances can we say that the $\lim_{n\to\infty}\mu(A_n) = \mu\left(\lim_{n\to\infty} A_n\right) = \mu(A)$?

In turns out that this is not a general property of measure. However, there are some sets for which this equality is true, and one collection of such sets is sets that are "increasing", i.e., in the sequence of sets $A_1, A_2, A_3, \ldots, A_n \ldots$ the sets have the property that $A_1 \subset A_2 \subset A_3 \subset \ldots \subset A_n \subset \ldots\ldots$ ).If the sequence of sets is increasing, we have demonstrated that $\lim_{n\to\infty}\mu(A_n) = \mu\left(\lim_{n\to\infty} A_n\right) = \mu(A)$

Now having demonstrated this property of sets and measure, we can begin the formal proof of the monotone convergence theorem.

We must show that if $\{f_n(\omega_i)\}$ is an increasing sequence of simple functions, such that $\lim_{n\to\infty} f_n(\omega_i) = f(\omega_i)$, then $\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu = \int_A \lim_{n\to\infty} f_n(\omega_i)\,d\mu.$ We will accomplish this by demonstrating that both $\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu \le \int_A \lim_{n\to\infty} f_n(\omega_i)\,d\mu$ and $\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu \ge \int_A \lim_{n\to\infty} f_n(\omega_i)\,d\mu$ are both true, leaving the equality statement as the only option.

Begin with demonstrating that $\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu \le \int_A \lim_{n\to\infty} f_n(\omega_i)\,d\mu.$ This part of the proof is straightforward.

Since $\{f_n(\omega_i)\}$ is an increasing sequence of functions up to $f(\omega_i)$, then $f_n(\omega_i) \le f_{n+1}(\omega_i) \le \lim_{n\to\infty} f_n(\omega_i) = f(\omega_i)$.In addition, by our properties of integrals on set functions, then $\int_A f_n(\omega_i)\,d\mu \le \int_A f_{n+1}(\omega_i)\,d\mu \le \ldots\ldots$ Since this sequence of integrals is increasing and is bounded, its limit, $\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu$ makes sense. If this is the case, then

$$\lim_{n\to\infty}\int_A f_n(\omega_i)\,d\mu = \sup_{n\to\infty}\int_A f_n(\omega_i)\,d\mu.$$

However, since $\{f_n(\omega_i)\}$ is a sequence of simple functions each of which is less than or equal to $f(\omega_i)$, then by the definition of the Lebesgue integral it now follows that the

[*] The Lebesgue dominated integration theorm is another example.

[†] Recall that, if the limit of an infinite sequence of sets $A_n$ exists, then every element of $A$ must be in all but finitely many of the $A_n$, and every element that is not in $A$ must only be in finitely many of the $A_n$, for $A$ to exixt.

$\sup\limits_{n\to\infty}\int\limits_A f_n(\omega_i)d\mu \le \sup\limits_{s/s(\omega_i)\le f(\omega_i)}\int s(\omega_i) = \int\limits_A f(\omega_i)d\mu.$ Thus $\lim\limits_{n\to\infty}\int\limits_A f_n(\omega_i)d\mu \le \int\limits_A f(\omega_i)d\mu.$[*] This concludes the first half of the proof.

We now show $\lim\limits_{n\to\infty}\int\limits_A f_n(\omega_i)d\mu \ge \int\limits_A \lim\limits_{n\to\infty} f_n(\omega_i)d\mu = \int\limits_A f(\omega_i)d\mu.$ This may seem somewhat counterintuitive.

Let's first define a simple function $s(\omega_i)$ that is less then our function $f(\omega_i)$ for all $\omega_i$. Now lets choose an arbitrary value $\alpha$ such that $0<\alpha<1$. Then clearly $\alpha s(\omega_i)<s(\omega_i)$ for all $\omega_i$ and $\int\limits_A \alpha s(\omega_i)d\mu = \alpha\int\limits_A s(\omega_i)d\mu \le \int\limits_A f(\omega_i).$ In addition, we are reminded by definition that $\sup\limits_{s/s(\omega_i)\le f(\omega_i)}\int s(\omega_i) = \int f(\omega_i).$

Now consider the sequence of functions $\{f_n(\omega_i)\}$. As this sequence increases to $f(\omega_i)$, the elements of $\{f_n(\omega_i)\}$ must first pass through and exceed $\alpha s(\omega_i)$ and then exceed $s(\omega_i)$. Now, for a fixed value of $n$ this will not be true for all $\omega_i$ but as $n$ increases it will be true for a larger and larger set of $\omega_i$ until it is true for all $\omega_i \subset A.$ Thus there is a set of $\omega_i$ what we will denote as $E_{n,\alpha}$ for which all $\omega_i \subset E_n$ have the property that $f_n(\omega_i) > \alpha s(\omega_i)$. As $n$ increases, this set of $\omega_i^s$, $E_n$ also increases. This permits us to write $\int\limits_{E_{n,\alpha}} \alpha s(\omega_i)d\mu = \alpha\int\limits_{E_{n,\alpha}} s(\omega_i)d\mu \le \int\limits_{E_{n,\alpha}} f_n(\omega_i).$ We also know that since $E_n \subset A,$ that $\int\limits_{E_{n,\alpha}} f_n(\omega_i) \le \int\limits_A f_n(\omega_i).$ Thus

$$\int\limits_{E_{n,\alpha}} \alpha s(\omega_i)d\mu = \alpha\int\limits_{E_{n,\alpha}} s(\omega_i)d\mu \le \int\limits_{E_{n,\alpha}} f_n(\omega_i) \le \int\limits_A f_n(\omega_i).$$

Taking limits we can write

$$\begin{aligned}\lim_{n\to\infty}\int\limits_{E_{n,\alpha}} \alpha s(\omega_i)d\mu = \alpha\lim_{n\to\infty}\int\limits_{E_{n,\alpha}} s(\omega_i)d\mu \le \lim_{n\to\infty}\int\limits_{E_{n,\alpha}} f_n(\omega_i)\\ \le \lim_{n\to\infty}\int\limits_A f_n(\omega_i).\end{aligned}$$

The critical inequality here is

$$\alpha\lim_{n\to\infty}\int\limits_{E_{n,\alpha}} s(\omega_i)d\mu \le \lim_{n\to\infty}\int\limits_A f_n(\omega_i).$$

[*] The definition of the Lebesgue integral is $\int f(\omega_i) = \sup\limits_{s/s(\omega_i)\le f(\omega_i)}\int s(\omega_i).$ This supremum is over all simple functions less than $f(\omega_i)$. The reason that we can say that $\sup\limits_{n\to\infty}\int\limits_A f_n(\omega_i)d\mu \le \sup\limits_{s/s(\omega_i)\le f(\omega_i)}\int s(\omega_i)$ is because the sequence of functions $\{f_n(\omega_i)\}$ is only a fraction of all simple functions that are less than or equal to $f(\omega_i)$. Thus, its supremum must be lower than the supremum over all simple functions. The fact that $\int f(\omega_i) = \sup\limits_{s/s(\omega_i)\le f(\omega_i)}\int s(\omega_i).$ finishes this part of the proof off.

We need to convert the region of integration on the left hand side of the integral to the set $A$.

In order to do this, we can invoke an earlier result about the measure of the set of increasing function. Now call $\int_A \alpha s(\omega_i)d\mu = \mu_{\alpha s}(A)$. So we have $\int_A \alpha s(\omega_i)d\mu = \mu_{\alpha s}(A)$. Also, since $E_n$ is increasing to $A$, then $A = \bigcup_{n=1}^{\infty} E_n$. Thus $\mu_{\alpha s}(A) = \mu_{\alpha s}\left(\bigcup_{n=1}^{\infty} E_n\right)$. But from what we have demonstrated before $\mu_{\alpha s}\left(\bigcup_{n=1}^{\infty} E_{n,\alpha}\right) = \mu_{\alpha s}\left(\lim_{n\to\infty} E_{n,\alpha}\right) = \lim_{n\to\infty} \mu_{\alpha s}\left(E_{n,\alpha}\right)$.Using this result, we can write

$$\alpha \lim_{n\to\infty} \int_{E_{n,\alpha}} s(\omega_i)d\mu = \alpha \lim_{n\to\infty} \int_A s(\omega_i)d\mu \le \lim_{n\to\infty} \int_A f_n(\omega_i).$$

Since this is true for all $\alpha \subset (0,1)$, then
$\int_A s(\omega_i) \le \lim_{n\to\infty} \int_A f_n(\omega_i)$ for all simple functions $s(\omega_i) < f(\omega_i)$. Thus $\sup_{s/s(\omega_i)\le f(\omega_i)} \int s(\omega_i) \le \lim_{n\to\infty} \int_A f_n(\omega_i)$.
Now invoking $\sup_{s/s(\omega_i)\le f(\omega_i)} \int s(\omega_i) = \int f(\omega_i)$. we have $\int_A f(\omega_i) \le \lim_{n\to\infty} \int_A f_n(\omega_i)$.

Since $\int_A f(\omega_i) \ge \lim_{n\to\infty} \int_A f_n(\omega_i)$ and $\int_A f(\omega_i) \le \lim_{n\to\infty} \int_A f_n(\omega_i)$ then $\int_A f(\omega_i) = \lim_{n\to\infty} \int_A f_n(\omega_i)$ and the theorem is proved.

## How do we know that the MCT applies?

How can we demonstrate that the monotone class theorem truly applies to any measurable function?

Our approach will be to first show that we can approximate a measurable function $f$ by a simple function $e_n^*(\omega_i)$. Then we will show that the integral of this simple function is the limit of the sequence of integrals of increasing simple functions.

The first step is to demonstrate that we can create a simple function as close to $f(\omega_i)$ as we like, for any $\omega_i \subset A$, we can partition $A$ into a collection of sets $\{A_k\}$ that are pairwise disjoint such that there exist a set $A_k$ for which $\omega_i \subset A_k$ but $\omega_i$ is not in any other set in $A_k$. Then we may write $f(\omega_i) = \alpha_{k(i)} 1_{\omega_i \subset A_k}$. Then we can write this as $f(\omega_i) \sim \left(\alpha_{k(i)} - \frac{\varepsilon}{n}\right) 1_{\omega_i \subset A_k}$. As $n$ gets large and for $\varepsilon$ arbitrarily small, $\left(\alpha_{k(i)} - \frac{\varepsilon}{n}\right) 1_{\omega_i \subset A_k}$ becomes arbitrarily close to $f(\omega_i)$.

Some Classic Measure Theory Results
Basic Properties of Probability
Counting Events

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Some Classic Measure Theory Results

This chapter provides the link between the Riemann and Lebesgue integral, as well as a proof of the Lebesgue Dominated Convergence Theorem.

## Prerequisites

Sequences of Sets<br>
Set Functions in Measure Theory<br>
Simple Functions in Public Health<br>
Measure Based Integration

## Deficiencies

We have previously discussed outer and inner measure. If we define measure (better described as Lebesgue measure[*]) as the length of an interval, then it seems that we have all that we need with the notion of outer measure or $\mu^*(A)$, where $\mu^*(A) = \inf_i \left[ \sum_{j=1}^{\infty} \mu(A_{i,j}) \right]$. However as it turns out $\mu^*(A)$ does not satisfy the constraint of additivity. This seems odd, since we motivated the concept of additivity by showing how unions of sets cover our set *A*, and in fact, for many sets, outer measure does provide additivity. However, on the real number line, there are some finite disjoint sets *A* and *B*, such that

$$\mu^*(A \cup B) \neq \mu^*(A) + \mu^*(B).$$

Now, such sets are not the sets we deal with commonly on the real line. The most well-known of these sets is the Vitali set which is an uncountable set of real numbers that is nonmeasurable. If *B* is a Vitali set, then we see that $\mu^*(A \cup B) \neq \mu^*(A) + \mu^*(B)$.

This finding was quite perplexing. As it turns out, the sets that break additivity are those that have fuzzy, foggy edges with no clear dividing line between what they contain and they

[*] One of the reasons that Lebesgue measure is so popular as a didactic tool is that it is the natural counterpart to Riemann integration.

don't contain (i.e., they are not intervals). The property of measure confers more measure to these sets than they should have.

Let's explore this concept of fuzziness for a second. Consider a collection of songs. If we wanted to characterize them by length, e.g. all songs that were less than or equal to four minutes, we can come up with the precise identity of each song that meet this criteria.

Now suppose we wanted to characterize all of these songs not by length, but by ratings , e.g., all songs that received three out of five star ratings. This would also be precise. However, as we listen to the songs we are constantly reconsidering whether they are three stars, two stars, or four stars, so the ratings are constantly changing. Thus, the identities of songs in our sets are constantly changing. Even though we think that the three star boundary is precise, from the user perspective it is fuzzy.

As another example, put a piece in bread in rapid two dimensional motion and ask someone to precisely point to the bread piece's edge. Every attempt fails, because of the rapid unpredictable movement of the bread. To the human, the edge of the bread is fuzzy, i.e., it cannot be located. Thus, the concept of fuzziness is not new to us, although considering it as a property of the real number line is.

There are uncountably many Vitali sets, i.e., immeasurable sets that themselves contain uncountably many real numbers on $[0,1]$. Recognizing this, Giuseppe Vitali himself felt compelled to conclude that the problem of finding an adequate definition of measure for intervals of the real line was unsolvable.[*] The complexity of the real number line confounds every possible definition of measure.

And, there still is no resolution of this. Measure can either have additive countability and not assign measure to all sets, or it can assign measure to all sets and fail the countability proposition. When confronted with this, Lebesgue concluded that it was preferable to have the countable additivity proposition. Therefore there are some sets (in fact there are an infinite number of them) that he conceded must simply have no measure.

Thus, in the end we settle on the definition that a measurable set is a set whose inner measure equals its outer measure.

## Carathéodory extension theorem

The paradigm in which we have been working is that we develop a system of a sample space, σ-algebra, and a measure, $(\Omega, \Sigma, \mu)$ which governs the measure of sets. However, what if we do not have a σ-algebra?

In some circumstance we have collections of sets that are closed under pairwise unions and intersections, but need not be closed under countable additivity. We will call such collections of sets fields, $\Im$. These fields are subsets of σ-algebras. Without a σ-algebra, there can be no measure, so the "measure" that we want to use on the field we call a premature or pseudomeasure. The pseudomeasure that operate on the field.

What the Carathéodory extension does is, is to extend $\Im$ to the smallest σ-algebra generated by the field. Now, if $\mu_0$ is countably additive on sets that happen to be countably additive in the field $\Im$, then the measure that operates on the smallest σ-algebra generated by $\Im$, provides the same measure as the premeasure when applied to sets in $\Im$.

The statement of the theorem follows.

[*] One might say that the measure of a clinical trial effect size is "fuzzy" since it has a sampling error component that obfuscates the true population effect.

*Carathéodory Extension Theorem.*
*Let $\Omega$ be a set and $\Im$ a field on it. Let $\mu_0$ be a finite pseudo measure on $\Im$. Then if $\mu_0$ is countably additive on sets in $\Im$ that happened to be countably additive in $\Im$, then there is unique measure $\mu$ that operates on the smallest σ-algebra generated by $\Im$, and agrees with the pseudomeasure $\Im$.*

This theorem is useful for extending a measurable function that exists on a collection of sets that is not a σ-algebra to a measure on a σ-algebra.

How does this work?

Assume that we are working on $(0,1]$. We can partition this interval into a finite sequence of semiclosed intervals $(a_i, b_i]$ $i = 1,...,n$ such that each is pairwise disjoint and $\bigcup_{i=1}^{n}(a_i, b_i] = (0,1]$, Using a measurable function $\mu_0$ (commonly called a pseudo measure or premeasure), the result that we would like is $\mu_0(0,1] = \sum_{i=1}^{n}\mu_0((a_i, b_i]) = \sum_{i=1}^{n}(b_i - a_i)$.

The extension theorem allows us to extend our collection of right semiclosed intervals to the σ-algebra on $(0,1]$, and then convert $\mu_0$ to a true measure $\mu$.

## Measurable functions from simple ones

One of the important consequences of the monotone convergence theorem is that a sequence of simple functions can be identified that converges to our commonly used Riemann integrable functions.

Here, we demonstrate how to find such a sequence of simple functions on the real line. Let's first consider $f(x)$, a real valued function which is measurable on intervals of the real number line. Develop a collection of intervals $\mathbf{I}_{n,k} = \left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)$ for $k = 1, 2, 3, ...2^{2n} + 1$, and $\mathbf{I}_{n,2^{2n}+1} = \left[\ 2^{2n}, \infty\right)$.

We see that this collection of intervals is disjoint. For example, let $n = 3$. Then the collection of semiclosed intervals on $[0, 1]$ is

$$\mathbf{I}_3 = \left[0,\frac{1}{8}\right), \left[\frac{1}{8},\frac{2}{8}\right), \left[\frac{2}{8},\frac{3}{8}\right), \left[\frac{3}{8},\frac{4}{8}\right), \left[\frac{4}{8},\frac{5}{8}\right), \left[\frac{5}{8},\frac{6}{8}\right), \left[\frac{6}{8},\frac{7}{8}\right), \left[\frac{7}{8},1\right).$$

For a fixed *n*, the intervals are small in number; as *n* increases, the intervals decrease in width and increase in number. The final interval covers the rest of the positive real number line.

We next need to define some measurable sets along the real number line. Since the collection of $\mathbf{I}_{n,k}$ sets are themselves on the real number line, we need to find those sets of *x*'s that $f(x)$ maps. Define for each $\mathbf{I}_{n,k}$, the set $f^{-1}(\mathbf{I}_{n,k})$.

Now, suppose our continuous function is $f(x) = 5x$. For the first set, what was *x* such that $f(x) = 5x = \frac{1}{8}$? Clearly, any number in the interval $\left[0, \frac{1}{40}\right)$ was mapped by *f* to $\left[0, \frac{1}{8}\right)$. Thus $f^{-1}\left(\left[0, \frac{1}{8}\right)\right) = \left[0, \frac{1}{40}\right)$. Similarly, $f^{-1}\left(\left[\frac{1}{8}, \frac{2}{8}\right)\right) = \left[\frac{1}{40}, \frac{2}{40}\right)$, and so on,

Now we are ready to define the convergent sequence of functions, $f_n(x) = \sum_{k=1}^{2^{2n}+1}\frac{k-1}{2^n}1_{f^{-1}(x)\subset\mathbf{I}_{n,k}}$.

We can see how this process actually works for the well behaved function $f(x)=\sqrt{x}$ (Figure 1).

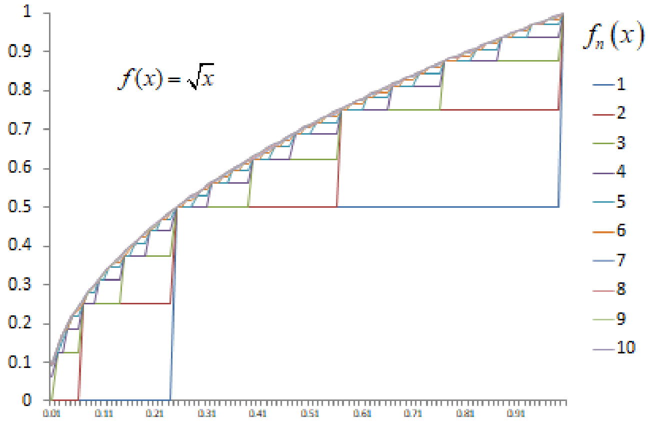


Figure 1. Building up the square root function from simple functions.

As $n$ increases the intervals $\mathbf{I}_{n,k}=\left[\frac{k-1}{2^n},\frac{k}{2^n}\right)$ become smaller and smaller, the fewer the number of values of $x$ fall in that interval, the more precise the contour of the line becomes.

One can also see how Riemann integration works with this square root function as well, easily visualizing the rectangles on which the integration is built. However, there is a fundamental difference in the two processes. Riemann builds the rectangles first, then finds a value of the function within the rectangle's horizontal vertices. The simple function approach of Lebesgue-Stieltjes evaluates the function to see if it falls within intervals. It either falls in them or it doesn't. Here, there is absolutely no area computation.

However, one can also see how tenable is the argument that for smooth functions both Riemann and Lebesgue-Stieltjes provide the same result.

### *Riemann and Lebesgue Integration Equivalence*

With the advent of Lebesgue integration, it is comforting to know that when both integrals exist, the Lebesgue-Stieltjes integral is equal to the Riemann integral. We can demonstrate this by returning to first principles for each integration definition. After these definitions, our goal will be to turn the upper and lower Riemann integrals (which are equal if the Riemann integral exists) into a sequence of step function and then demonstrate that these are Cauchy sequences and hence converge..

Let's begin with Riemann. First, define our integral function as $f(x)$. We know that by taking a sequence of meshes, we can write the upper Riemann integral for any partition P of an interval of the real line as

$$R_L(f,P)=\sum_{i=1}^{n}\inf_{x\in[x_{i-1},x_i]} f(x)(x_{i-1}-x_i)$$

$$R_u(f,P)=\sum_{i=1}^{n}\sup_{x\in[x_{i-1},x_i]} f(x)(x_{i-1}-x_i).$$

And the Riemann integral $\mathbf{R}\int f(x)=\lim_{n\to\infty}R_L\left(f,P\right)=\lim_{n\to\infty}R_U\left(f,P\right)$. This is true for any partition $P_k$ of the *x*-axis so we may write
$\mathbf{R}\int f(x)=\lim_{k\to\infty}R_L\left(f,P_k\right)=\lim_{k\to\infty}R_U\left(f,P_k\right)$.

We now have to show that this converges to the Lebesgue-Stieltjes integral. Our goal will be to define simple functions that both map to the infimum and supremum of the Riemann sums and lead to Lebesgue integrals. Let's define the two functions

$$\lambda\left(f,P,x\right)=\sum_{i=1}^{n}\inf_{x\in\left[x_{i-1},x_i\right]}f(x)\,1_{x\in\left(x_{i-1}-x_i\right)}$$

$$\omega\left(f,P,x\right)=\sum_{i=1}^{n}\sup_{x\in\left[x_{i-1},x_i\right]}f(x)\,1_{x\in\left(x_{i-1}-x_i\right)}.$$

Note that $\lambda\left(f,P,x\right)\le f(x)\le\omega\left(f,P,x\right)$ except at the points $x=x_i$ (but there are only countably many of these). Also note that these indicator functions covers the same intervals with the same mesh endpoints as the Riemann upper and lower sums. Thus the foundation of both the Riemann and Lebesgue-Stieltjes integrals is the same interval set of the *x*-axis in this construction.

Because they are defined on open intervals at the value $x=x_i$, both $\lambda\left(f,P,x\right)$ and $\omega\left(f,P,x\right)$ are each zero at those points and are indicator functions. Then, we can compute their Lebesgue integrals

$$L_L\left(f,P\right)=\int\lambda\left(f,P,x\right):\ \ L_U\left(f,P\right)=\int\omega\left(f,P,x\right).$$

Since $\lambda\left(f,P_k,x\right)$ and $\omega\left(f,P_k,x\right)$ are also functions of the partition of the *x*-axis, we may consider $P_{k+1}$ a refinement of $P_k$ and the integral based on $P_{k+1}$ is closer to $f(x)$ than that built up from $P_k$ we may write.
$\lambda\left(f,P_k,x\right)\le\lambda\left(f,P_{k+1},x\right)\ \le f(x)\le\ \omega\left(f,P_{k+1},x\right)\le\omega\left(f,P_k,x\right)$, again at the points $x=x_i$ (a countable set). Thus, we may write

$$\begin{aligned}&\int\left|\lambda\left(f,P_{k+m}\right)-\lambda\left(f,P_k\right)\right|=\int\lambda\left(f,P_{k+m}\right)-\lambda\left(f,P_k\right)\\&=\int\lambda\left(f,P_{k+m}\right)-\int\lambda\left(f,P_k\right)=L\left(f,P_{k+m}\right)-L\left(f,P_k\right)\\&=\left|L\left(f,P_{k+m}\right)-L\left(f,P_k\right)\right|\end{aligned}$$

Except for the countable set of join points that we know as Lebesgue measure zero.

As *k* goes to infinity these terms get as close together as desired, and therefore the sequence is a Cauchy sequence and converges.

Similarly, $\int\left|\omega\left(f,P_{k+m}\right)-\omega\left(f,P_k\right)\right|$ is a Cauchy sequence. Thus the functions $\lambda\left(f,P_k,x\right)$ and $\omega\left(f,P_k,x\right)$ converge almost everywhere on the interval $\left[a,b\right]$ and since $f(x)$ is trapped between them they both converge to $f(x)$. Thus the limits of these step functions converge to the Lebesgue integral of $f(x)$.

Now, we remember that the integrals of these step functions also map to the lower and upper Riemann sums (we specifically defined them that way), whose limit is the Riemann integral. Thus the step functions converge to the Riemann integral as well, and therefore the Riemann and Lebesgue integrals are equal.

From the Lebesgue-Stieltjes perspective, a function has a Riemann integral if the set on which it is discontinuous has Lebesgue measure 0. If we return to the Dirichlet function defined on 0,1, as $f(x)=1_{x=\text{rational}}$, then the set on which $f(x)$ is discontinuous has Lebesgue measure one, signifying that the Riemann integral does not exist.

## Lebesgue measure on countable sets

In fact, the only set on which the Dirichlet function takes a nonzero value is a set of zero Lebesgue measure. To show this, let there be a countable set $q_n$, $n = 1$ to $\infty$ where the sets are disjoint. We want to show that $\mu\left(\bigcup_{n=1}^{\infty} q_n\right)=0$. The key to the demonstration is to take advantage of the disjoint nature of the countable set. Around each point, build an interval from $q_n-\frac{\varepsilon}{2^n}$ to $q_n+\frac{\varepsilon}{2^n}$. Since the countable set is disjoint, these intervals are disjoint. Thus, we could write $\mu\left(\bigcup_{n=1}^{\infty} q_n\right)=\mu\left(\bigcup_{n=1}^{\infty}\left[q_n-\frac{\varepsilon}{2^n},q_n+\frac{\varepsilon}{2^n}\right]\right)$.

Recall though that the measure is the infimum of the union of these coverings. Now, applying Lebesgue measure to each of these intervals, we find that $\mu\left(\bigcup_{n=1}^{\infty}\left[q_n-\frac{\varepsilon}{2^n},q_n+\frac{\varepsilon}{2^n}\right]\right)=\sum_{n=1}^{\infty}\frac{2\varepsilon}{2^n}=\varepsilon\sum_{n=1}^{\infty}\frac{1}{2^{n-1}}=\varepsilon\sum_{n=0}^{\infty}\frac{1}{2^n}=2\varepsilon.$ Since $\varepsilon$ is arbitrarily small, we have that $\inf\left(\mu\left(\bigcup_{n=1}^{\infty} q_n\right)\right)=0.$

## Dominated convergence theorem

We have seen that one advantage of Lebesgue-Stieltjes integration is that it is a superset of functions that are Riemann integrable. When both integrals exists, they are equal. Plus there are functions for whom the Riemann integral does not exists but Lebesgue-Stieltjes does.

Another advantage of Lebesgue-Stieltjes integration is that for a sequence of convergent functions, the limit passes though the integration sign under a different set of conditions.

Now, we have seen that the same property holds for functions that are Riemann integral. The only condition is that the function must be uniformly convergent on an interval for this to take place. However, since by bounding a function on the real number line (such as the function $f(x)=x^n$ on the $[0,1)$ ) we can make many functions that are not uniformly convergent on the entire real number line uniformly convergent in the interval of interest. If this is the case, then limits pass through the Riemann integral on these intervals as well.

However, for functions that are simply too discontinuous for the Riemann integral to exist, we would still like to know what we can pass the limit through the integral sign. The statement of this feature is the dominated convergence theorem.

The dominated convergence theorem states that if we have pointwise convergence of a sequence of functions $f_n(x)$ to $f(x)$, then all that we need for $\lim_{n\to\infty}\int f_n(x)=\int f(x)$ is for every function in the sequence to be less than or equal to (i.e., "dominated") by a function $g(x)$ where the function $g$ is measurable. This is a very useful tool in integration theory.

Before we proof it though we need Fatou's lemma which makes a related statement in terms of liminfs. The proof of the dominated convergence theorem is straightforward if we can rely on Fatou's lemma, but the lemma itself is complicated so we will spend some time on it.

### *Fatou's Lemma*

For Fatou's lemma we begin with a sequence of nonnegative measurable functions $f_n(x)$ on a measure space $\Omega$. Define $f(x)$ as $\liminf\limits_{n\to\infty} f_n(x) = f(x)$.[*] Then according to Fatou's lemma.

$$\int_\Omega \liminf_{n\to\infty} f_n(x) \geq \int_\Omega f(x).$$

To prove this, begin with a set $E \subset \Omega$. Let $\psi(x)$ be a non-negative simple function such that $\psi(x) \leq f(x)$. Let's also let **M** be the maximum value of $\psi(x)$. Is it true that $\int_E \liminf\limits_{n\to\infty} f_n(x) \geq \int_E \psi(x)$? We would expect this to be true since $\psi(x) \leq f(x)$ and $\lim\limits_{n\to\infty} f_n(x) = f(x)$.

Let's define $A_n = \{x \in E \,/\, f_n(x) > (1-\varepsilon)\psi(x) \;\; \forall n \geq N\}$. Defined this way $A_n$ is an infinite number of increasing sets (since every set contains only those that are at least that far out in the sequence) whose union contains $A$. Therefore the set $A - A_n$ is a set of decreasing sets whose intersection is the empty set. Let's assume that $A$ has finite measure.[†] Then since $A = \bigcup\limits_{n=1}^{\infty} A_n$, then $\lim\limits_{n\to\infty} m(A_n) = m(A)$, which implies that for $n \geq N$ we can write $\forall_{n>N} \;\; m(A - A_n) \leq \varepsilon$. Therefore, $\int_E f_n(x) \geq \int_{A_n} f_n(x) \geq (1-\varepsilon)\int_{A_n} \psi(x)$. Also, since $\int_E \psi(x) = \int_A \psi(x)$, (because $A = \bigcup\limits_{n=1}^{\infty} A_n$,) then we know that

$\int_A \psi(x) = \int_{A_n} \psi(x) + \int_{A-A_n} \psi(x)$. So $\int_E \psi(x) = \int_{A_n} \psi(x) + \int_{A-A_n} \psi(x)$, and

$\int_{A_n} \psi(x) = \int_E \psi(x) - \int_{A-A_n} \psi(x)$. Multiplying each side by $(1-\varepsilon)$ we find

$$(1-\varepsilon)\int_{A_n} \psi(x) = (1-\varepsilon)\int_E \psi(x) - (1-\varepsilon)\int_{A-A_n} \psi(x).$$

$(1-\varepsilon)\int_{A_n} \psi(x) \geq (1-\varepsilon)\int_E \psi(x) - \int_{A-A_n} \psi(x)$. So we have two inequalities and can now write

[*] Remember for liminfs of a function we find $\liminf\limits_{k\to\infty\; m\geq k} f_m(x)$. the infimums of the tail subsequences beyond each value of the index $k$.

[†] Another proof is available for the case where $m(A) = \infty$.

$$\begin{aligned}\int_E f_n(x) \geq & \int_{A_n} f_n(x) \geq (1-\varepsilon)\int_{A_n} \psi(x) \\ & \geq (1-\varepsilon)\int_E \psi(x) - \int_{A-A_n} \psi(x) \\ & \geq \int_E \psi(x) - \varepsilon \int_E \psi(x) - \int_{A-A_n} \psi(x) \\ & \geq \int_E \psi(x) - \varepsilon \int_E \psi(x) - \mathbf{M} \\ & \geq \int_E \psi(x) - \varepsilon \left( \int_E \psi(x) + \mathbf{M} \right)\end{aligned}$$

Now we let $\varepsilon > 0$ and taking the liminf we have $\liminf_{n\to\infty} \int_E f_n(x) \geq \int_E \psi(x)$.

With Fatou's lemma in place, we can now prove the Lebesgue Dominated Convergence Theorem. As we have before, let's assume that $f_n(x)$ is a sequence of functions converging pointwise to $f(x)$, and that on the set $\Omega$, $f(x) \leq g(x)$. Then we know that

$|f_n(x) - f(x)| \leq |f(x)| + |f_n(x)| \leq 2g(x)$. We also know that $\limsup_{n\to\infty} |f_n(x) - f(x)| = 0$. Thus $\left| \int_\Omega f(x) - \int_\Omega f_n(x) \right| = \left| \int_\Omega f(x) - f_n(x) \right| \leq \int_\Omega |f(x) - f_n(x)|$.

Now, Fatou's lemma implies

$\limsup_{n\to\infty} \sup_{m>n} \left| \int_\Omega f(x) - f_n(x) \right| \leq \int_\Omega \limsup_{n\to\infty} \sup_{m>n} |f(x) - f_n(x)| = 0$. This in turn implies that the limit exists and is zero for $x \subset \Omega$. Thus $\lim_{n\to\infty} \int_\Omega |f(x) - f_n(x)| = 0$, and $\lim_{n\to\infty} \int f_n(x) = \int f(x)$.

## Return to properties of Lebesgue integrals

The Lebesgue dominated convergence theorem is a powerful tool, but let's now turn our attention to some other more ordinary but essential features of Lebesgue integrals.

First, the expression of a Lebesgue integrable function $f(x)$ as a simple function is not unique. Simple functions are so flexible that there are many ways to parameterize them. Yet, if there are multiple ways to assemble $f(x)$ from simple functions, than the value of the integral of these different simple functions should be the same.

Formally, if $f(x)$ is integrable on a closed interval (we can choose $[0,1]$ for convenience), and $f(x) = S_1 = \sum_{i=1}^{n} \alpha_i 1_{x\in A_i}$, and alternatively, $f(x) = S_2 = \sum_{j=1}^{m} \beta_i 1_{x\in B_j}$, then $\sum_{i=1}^{n} \alpha_i m(A_i)$ should be equal to $\sum_{j=1}^{m} \beta_j m(B_j)$. We can show this by partitioning the $[0,1]$ interval into a collection of subintervals such that we can we can construct the partitions of $S_1$ and $S_2$. Lets create a new simple function $S_3 = \sum_{k=1}^{L} \gamma_k \mathbf{1}_{x \in C_k}$ such that for every set $A_i$ in $S_1$, then there exists a subcollection of $\{C_k\}$ (call this $\{C_{i(k)}\}$ such that $A_i = \bigcup_{n_{i(k)}} C$. Similarly, for every set $B_j$ in $S_2$, then there exists a subcollection of $\{C_k\}$ (call this $\{C_{j(k)}\}$ such that $B_j = \bigcup_{n_{j(k)}} C$. Now set $\gamma_k = \alpha_i$ if there is an *i* such that $A_i = \bigcup_{n_{i(k)}} C$ and $\gamma_k = \beta_j$ if there is an *j* such that $B_j = \bigcup_{n_{j(k)}} C$. Then $f(x) = \sum_{k=1}^{L} \gamma_k \mathbf{1}_{x \in C_k}$. Therefore $\int f(x) = \sum_{k=1}^{L} \gamma_k m(C_k)$. But $\sum_{k=1}^{L} \gamma_k m(C_k) = \sum_{k=1}^{L} \gamma_k m(C_{i(k)}) = \sum_{k=1}^{L} \gamma_k m\left(\bigcup_{n_{i(k)}} C\right) = \sum_{i=1}^{n} \alpha_i m(A_i)$. Analogously $\sum_{k=1}^{L} \gamma_k m(C_k) = \sum_{k=1}^{L} \gamma_k m(C_{i(k)}) = \sum_{k=1}^{L} \gamma_k m\left(\bigcup_{n_{j(k)}} C\right) = \sum_{j=1}^{m} \beta_j m(B_j)$.

To show that for example for an integrable function $f(x)$, and a nonzero constant *a* $\int af(x) = a\int f(x)$, we first show that this is true for indicator functions, then simple functions, then for the limits of simple function. At that point, we invoke the Monotone Convergence Theorem to produce the result that, since integrable functions are the limits of simple functions, then any property of the limits of simple functions must hold for integrable functions.

So we begin with $f(x) = \alpha \mathbf{1}_{x=A}$. Then $\int af(x) = \int a\alpha \mathbf{1}_{x \subset A} = \int a\alpha \mathbf{1}_{x \subset A} = \int \gamma \mathbf{1}_{x \subset A}$, where $\gamma = a\alpha$. The fact that we recognize $\int \gamma \mathbf{1}_{x \subset A} = \gamma m(A)$ now permits us to finish as

$$\begin{aligned}\int af(x) &= \int a\alpha \mathbf{1}_{x \subset A} = \int a\alpha \mathbf{1}_{x \subset A} = \int \gamma \mathbf{1}_{x \subset A} \\ &= \gamma m(A) = a\alpha m(A) = a\int f(x),\end{aligned}$$

and we have the result for an indicator function.

To move to simple functions we define $f(x) = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{x \subset A_i}$. Then we proceed

$$\begin{aligned}\int af(x) &= \int a\sum_{i=1}^{n} \alpha_i \mathbf{1}_{x \subset A_i} = \int \sum_{i=1}^{n} a\alpha_i \mathbf{1}_{x \subset A_i} = \int \sum_{i=1}^{n} \gamma_i \mathbf{1}_{x \subset A_i} = \sum_{i=1}^{n} \gamma_i m(A_i) \\ &== \sum_{i=1}^{n} a\alpha_i m(A_i) = \sum_{i=1}^{n} \alpha_i m(A_i) = a\int f(x).\end{aligned}$$

We now define $f(x) = \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i \mathbf{1}_{x \subset A_i}$ followed by

$$\begin{aligned}\int af(x) &= \int a \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i \mathbf{1}_{x \subset A_i} = \int \lim_{n\to\infty} \sum_{i=1}^{n} a\alpha_i \mathbf{1}_{x \subset A_i} = \int \lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i \mathbf{1}_{x \subset A_i} \\ &= \lim_{n\to\infty} \int \sum_{i=1}^{n} \gamma_i \mathbf{1}_{x \subset A_i} = \sum_{i=1}^{n} \gamma_i m(A_i) = \lim_{n\to\infty} a\sum_{i=1}^{n} \alpha_i m(A_i) \\ &= a\lim_{n\to\infty} \int \sum_{i=1}^{n} \alpha_i \mathbf{1}_{x \subset A_i} = a\int \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i \mathbf{1}_{x \subset A_i} = a\int f(x).\end{aligned}$$

Note the use of the Lebesgue dominated convergence theorem twice here as we move the limit first outside, then inside the integral sign.

Let’s also follow the same process for demonstrating that if we have two integrable functions $f(x)$ and $g(x)$ then $\int f(x)+g(x)=\int f(x)+\int g(x).$ As before, we begin with indicator functions. Let $f(x)=\alpha\mathbf{1}_{x\subset A}$ and $g(x)=\beta\mathbf{1}_{x\subset B}.$

$$\begin{aligned}\int f(x)+g(x)=\int\alpha\mathbf{1}_{x\subset A}+\beta\mathbf{1}_{x\subset B}&=\int\sum_{i=1}^{2}\gamma_i\mathbf{1}_{x\subset C_i}=\sum_{i=1}^{2}\gamma_i m\left(C_i\right).\\ &=\gamma_1 m\left(C_1\right)+\gamma_2 m\left(C_2\right)=\alpha m\left(A\right)+\beta m\left(B\right)\\ &=\int f(x)+\int g(x).\end{aligned}$$

Here the integrand is just a simple function and since we know the measure of simple functions, we compute its measure, than rewrite in terms of the original simple functions and integrands.

We now examine this proposition for simple functions. Let $f(x)=\sum_{i=1}^{n}\alpha_i\mathbf{1}_{x\subset A_i}$ and $g(x)=\sum_{j=1}^{m}\beta_j\mathbf{1}_{x\subset B_j}$ and proceed.

$$\begin{aligned}\int f(x)+g(x)=\int\sum_{i=1}^{n}\alpha_i\mathbf{1}_{x\subset A_i}+\sum_{j=1}^{m}\beta_j\mathbf{1}_{x\subset B_j}&=\int\sum_{h=1}^{L}\gamma_L\mathbf{1}_{x\subset C_i}\\ &=\sum_{h=1}^{L}\gamma_h m\left(C_h\right)=\sum_{i=1}^{n}\alpha_i m\left(A_i\right)+\sum_{j=1}^{m}\beta_j m\left(A_j\right)\\ &=\int\sum_{i=1}^{n}\alpha_i\mathbf{1}_{x\subset A_i}+\int\sum_{j=1}^{m}\beta_j\mathbf{1}_{x\subset B_j}\\ &=\int f(x)+\int g(x).\end{aligned}$$

This works because of the flexibility simple functions afford. We can rewrite a combination of simple functions into another simple function.

Now for the limits of simple functions we write and $f(x)=\lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i\mathbf{1}_{x\subset A_i}$ and $g(x)=\lim_{m\to\infty}\sum_{j=1}^{m}\beta_j\mathbf{1}_{x\subset B_j}$ and follow with

$$\begin{aligned}\int f(x)+g(x)&=\int\left(\lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i\mathbf{1}_{x\subset A_i}+\lim_{m\to\infty}\sum_{j=1}^{m}\beta_j\mathbf{1}_{x\subset B_j}\right)=\int\lim_{L\to\infty}\sum_{h=1}^{L}\gamma_L\mathbf{1}_{x\subset C_i}\\ &=\lim_{L\to\infty}\int\sum_{h=1}^{L}\gamma_L\mathbf{1}_{x\subset C_i}=\int\lim_{L\to\infty}\sum_{h=1}^{L}\gamma_L\mathbf{1}_{x\subset C_i}=\lim_{L\to\infty}\int\sum_{h=1}^{L}\gamma_L\mathbf{1}_{x\subset C_i}\\ &=\lim_{L\to\infty}\sum_{h=1}^{L}\gamma_h m\left(C_h\right)=\lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i m\left(A_i\right)+\lim_{m\to\infty}\sum_{j=1}^{m}\beta_j m\left(A_j\right)\end{aligned}$$

Again the resilience of simple functions makes all of the difference. Continuing,

$$= \lim_{L\to\infty}\sum_{h=1}^{L}\gamma_h m\left(C_h\right) = \lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i m\left(A_i\right) + \lim_{m\to\infty}\sum_{j=1}^{m}\beta_j m\left(A_j\right)$$

$$= \lim_{n\to\infty}\int\sum_{i=1}^{n}\alpha_i 1_{x\subset A_i} + \lim_{m\to\infty}\int\sum_{j=1}^{m}\beta_j 1_{x\subset B_j}$$

$$= \int\lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i 1_{x\subset A_i} + \int\lim_{m\to\infty}\sum_{j=1}^{m}\beta_j 1_{x\subset B_j} = \int f(x) + \int g(x).$$

$$= \int f(x) + \int g(x).$$

Lebesgue dominated convergence theorem plays its usual role here.

Probability Foundations
Elementary Set Theory
Basic Properties of Probability

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Vitali Sets

When the notion of measurable sets of real numbers appeared, one of its first results was that the set of rational numbers, although themselves dense had measure zero. A natural next question to ask was whether there were any sets of dense real numbers that were not Lebesgue measurable. The Vitali set is one of the most vivid examples of a set of real numbers (not just rational numbers) that are both dense and are not Lebesgue measurable*. We will first demonstrate how the Vitali set is constructed, and then demonstrate its non-measurable property.

## Developing the Vitali set

### *Rational numbers*

We begin with the assertion that the rational numbers, although dense, are non-measurable. The density of the rational numbers simply means that every rational number is either a rationale number itself or is as close as we would like it to be to another rational number (that is, within $\varepsilon$ of other rational numbers for any $\varepsilon$ arbitrarily small). We may at first think that this means that there are "enough" rational numbers to satisfy the measurable criteria. However, because they are countable, even though there is are an infinite number of them, there is sufficient "space between" them to give the set of rational numbers on $[0,1]$ Lebesgue measure zero.

Vitali numbers build on this principle. We first develop a collection of classes of Vitali numbers on $[0,1]$ and demonstrate that they are dense. We then prove by contradiction that the union of these classes of numbers which covers $[0,1]$cannot be measured.

### *Vitali numbers*

Actually this is quite simple. For each irrational number *r* on $[0,1]$, we create a set of irrational numbers $\{b_r\}$ such that $r+b_r$ is rational.

This is quite straightforward process, since the sum of two irrational numbers can be rational†. However, since there is an infinite and countable number of rational numbers, we have

* We are not saying that these sets have Lebesgue measure zero. We are saying that they are not measurable, i.e., we do not know how to measure their content.

† $\left(1-\sqrt[3]{2}\right)+\sqrt[3]{2}=1$.

an infinite and countable collection of $b_r$'s for each irrational number *r*. This is a dense set of denumerable irrational numbers.

Now, since there are uncountably many real numbers on $[0,1]$, and every real number *r* has an infinite countable set $\{b_r\}$associated with it there are uncountably many collections of sets $\{b_r\}$. Furthermore, since every real number has a collection of sets $\{b_r\}$ the union of all these $\{b_r\}$ sets covers$[0,1]$, i.e., $\bigcup_r \{b_r\} \supset [0,1]$. The Vitali set is the set composed by selecting one element from each of the $\{b_r\}$. Thus, the Vitali set is an uncountable set of irrational numbers on $[0,1]$.

## Non-measurability of the Vitali set

We prove the non-measurability of the Vitali set indirectly. To set this up, remember that the set of rationale numbers is countable. Let's sequence these rational numbers in the interval $[-1,1]$by $q_1, q_2, q_3, ..., q_k, ...,$. Now choose a rational number $q_k$ then select a Vitali set $\{v\}$ and create a countably infinite number of sets $V_k = \{v + q_k\}$. The construction makes the $V_k$ sets pairwise disjoint. Now begin with.

$$[0,1] \subseteq \bigcup_k V_k \subseteq [-1,2].$$

The first $\subseteq$ comes from the property of Vitali sets. The second $\subseteq$ is from the simple addition of the dense set of rationals on $[-1,1]$ to the Vitali set.

If the Vitali set was measurable, then

$$\mu([0,1]) \subseteq \mu\left(\bigcup_k V_k\right) \subseteq \mu([-1,2]).$$

Since, $\mu\left(\bigcup_{k=1}^{\infty} V_k\right) = \sum_{k=1}^{\infty} \mu(V_k),$ we can write using Lebesgue measure

$$1 \le \sum_{k=1}^{\infty} \mu(V_k) \le 3$$

and since the measure is translation invariant, we have

$$1 \le \sum_{k=1}^{\infty} \mu(V) \le 3$$

However, how could this be possible? The Lebesgue measure of the Vitali set is a constant, therefore its sum is either zero or infinity. Thus, we have a contradiction and the Vitali set must be nonmeasurable.

So, why is the Vitali set, itself a set of uncountably many irrational numbers not measurable? Because the set is linked to rational numbers. Linking each irrational number *r* to a denumerable number of irrational numbers, i.e., making it countable, introduced enough "space between the irrational numbers to undermine measurability. The fact that one has uncountably many of these sets does not cure the link to the denumerable rationals.

It is easy to see what happens. Let's say that we have the irrational number $\sqrt[3]{0.2} \approx 0.585.$ Can we fill the interval $\left[\sqrt[3]{0.2}, 1\right]$ with irrational numbers. Certainly. Simply be adding a set of uncountably many irrational numbers, we can fill up the $\left[\sqrt[3]{0.2}, 1\right]$. This would be all of irrationals in the interval $\left[0,\ 1-\sqrt[3]{0.2}\right]$.However, suppose we "counted the infinite but denumerable set of rational numbers in $\left[\sqrt[3]{0.2},1\right]$ and for each irrational number in this interval, added an irrational number to $\sqrt[3]{0.2}$ to produce that rational number. This would reproduce the rationals in $\left[0,\ 1-\sqrt[3]{0.2}\right]$ and, although dense, would not permit the "space" between them to fill up. Thus the denumerable infinite set of irrational numbers has measure 0, although we never stop adding to them.

Up until this section, we knew that the set of rational numbers were both dense and had measure zero. Now we know that there are sets of irrational numbers that cover an open interval and are nonmeasurable. In fact, since we chose only one member of each of the $\{b_r\}$ sets to create our Vitali set, sand there are uncountably many Vitali sets, there are uncountably many collections of dense real numbers that are also nonmeasurable.

So, what is so special about the Vitali set? The problem is that it has no sharp edges. Sets like intervals (be they open or closed) have clear and well defined boundaries. However the complexity of the real number line generates sets of real numbers that are dense, and even nondenumerable without clear dividing lines between what is a member and a non-member. The Vitali set is one of these, and its presence confounds the assignment of measure.

# The Basics of Bernoulli Trials- Bernoulli Distribution

The Bernoulli distribution is among the simplest probability distributions yet is rich enough to provide a solid foundation for all of our work with more complicated probability laws.

We will introduce and study the Bernoulli distribution in detail, providing definitions and computations on which we will rely for all of our deeper discussions of probability. Note that a more advanced treatment of the Bernoulli distribution is also available here.

## Prerequisites for the Bernoulli Distribution

The Notion of Random Events
Sigma Notation
Factorials Permutations, and Combinations
Elementary Set Theory
Properties of Probability
Conditional Probability
Counting Events - Combinatorics

## Properties of Bernoulli trials

In the section that discussed sampling with and without replacement, we set up a series of experiments that led to the computation of probability without actually collecting any data. This focused on sampling from a particular type of experiment. That experiment had the following characteristics:

- The result of the experiment can be dichotomized. Thus only two possible outcomes can occur. They are typically characterized as either a "success" or a "failure".
- The probability of a success is a known quantity, $p$ and the same from experiment to experiment.
- The result of one experiment does not influence the likelihood of a success or failure in future experiments. Experimental results are independent of each other.

Experiments that have property 1 are known as Bernoulli trials, named for Jacob Bernoulli, from the famed Bernoulli family. If, in addition they have properties 2 and 3, they are known as independent and identically distributed (i.i.d) Bernoulli trials.

Bernoulli trials are ubiquitous in our culture. That, and the relative ease of probability computations that they induce makes this among the most popular and easily understood probability distributions.

We will introduce the concept of the Bernoulli distribution and use its simplicity to begin discussions about probability distributions in general. In addition, using Bernoulli trials will allow us to deconstruct complicated events into simple ones that are relatively easy to solve..

## Identifying the Probability Mass Function

Qualitative characterization of this collection of experiments (each with independent outcomes described as success or failure) as Bernoulli trials is a fine start. However, we will also need to be able to provide a mathematical characterization as well.

In order to completely describe the distribution of probabilities across both outcomes, we will need to characterize the outcome of the experiment. In the case of the Bernoulli distribution, this involves an experiment that produces two outcomes, either success or failure, sometimes characterized as a "1" or a "0".

For example, we might characterize the mortality finding of a surgery as follows. Let $X$ be the mortality outcome. If the patient survives the surgery, then we have a success and $X = 1$. If the patient dies, then we have a failure and $X = 0$. To characterize a distribution we simply need to provide all possible experimental outcomes and the probabilities attached to them.

## Random variables

We are used to the concept of the variable $X$ as representing an unknown quantity. In standard mathematics, this is typically an unknown, fixed quantity. However in probability we will adjust this variable to be the result of an experiment.

In the prior example we would let $X_i$ reflect the outcome of the $i^{\text{th}}$ subject undergoing surgery. Before or during the surgery, we do not know whether the patient will survive. Thus, we denote the unknown result, subject to influences both known and unknown as $X_i$, calling it a random variable.

## Mass and cumulative distributions

For a Bernoulli trial we characterize the possible outcomes as either 0, or 1. We can then write

$$\begin{aligned}&\mathbf{P}\left[X=1\right]=p\\&\mathbf{P}\left[X=0\right]=q=1-p.\end{aligned}$$

This is all we need to characterize the Bernoulli distribution. This is the probability mass function. Essentially, there are two "masses" of probability, one mass for $X=0$, the other for $X=1$. All other values of $X$ have probability zero. They are simply impossible under both the experimental model and the Bernoulli distribution.

We can also accumulate probability over different values of $X$. The most usful of these is the cumulative distribution function, $\mathbf{F}_X(x)$. It is defined as

$$\mathbf{F}_X(x)=\mathbf{P}\left[X\leq x\right].$$

In this notation, $X$ represents a Bernoulli variable. We call this a random variable because it is the outcome of an experiment. Thus we can simply write for the Bernoulli distribution.

$$\begin{aligned} \mathbf{F}_X(x) &= 0 \text{ for } x < 1 \\ &= 1 \text{ for } x \geq 1. \end{aligned}$$

## Mean and variance definition

Another quantity of interest is the expected value of *X*. This is typically known as just the "average of" *X*. Technically it is the weighted accumulation of all possible values of *X*, the weights being the probability of *X*. Described this way, the formula for the expectation of $X$, denoted as $\mathbf{E}\left[X\right]$ can be easily written as

$$\mathbf{E}\left[X\right] = \sum_{All\,x} x\mathbf{P}\left[X = x\right].$$

We can compute this easily and directly for the Bernoulli mass function.

$$\mathbf{E}\left[X\right] = 0\mathbf{P}\left[X = 0\right] + 1\mathbf{P}\left[X = 1\right] = 0q + 1p = p$$

Similarly, we could write

$$\mathbf{E}\left[X^2\right] = 0^2\mathbf{P}\left[X = 0\right] + 1^2\mathbf{P}\left[X = 1\right] = 0q + 1p = p.$$

A measure of dispersion of *X* is the variance of *X*, $\mathbf{Var}\left[X\right]$.The $\mathbf{Var}\left[X\right]$ is defined as

$$\mathbf{Var}\left[X\right] = \mathbf{E}\left[\left(X - \mathbf{E}\left[X\right]\right)^2\right]$$

The is commonly written as $\sigma^2$, and the standard deviation of *X* which is the square root of the variance, is $\sigma$. One advantage of the standard deviation is that it is in the same units as the random variable *X*.

The variance with its square term is only zero in the special circumstance when every realized value of the random variable *X* is the same, i.e., $X_i = c$ for all *i*. The general solution for i.i.d. Bernoulli trials is

$$\sigma_X^2 = \mathbf{Var}\left(X\right) = \mathbf{E}\left[X^2\right] - \mathbf{E}^2\left[X\right] = p - p^2 = p\left(1 - p\right) = pq.$$
$$\sigma_X = \mathbf{SD}(X) = \sqrt{pq}.$$

The expected value, $\mathbf{E}\left[X\right]$ is a measure of central tendency, $\mathbf{Var}\left[X\right]$ is a measure of dispersion.

## Sums of Bernoulli random variables

We can also compute expectations of sums of random variables. For example, if $X_1$ and $X_2$ be two random variables, Then we can compute the expectation of the sum of them $W = X_1 + X_2$ as $\mathbf{E}\left[W\right] = \mathbf{E}\left[X_1\right] + \mathbf{E}\left[X_2\right]$.

Similarly, if the random variables are independent, that is knowledge of one does not help us with the value of the other, then $\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2]$.

Thus, if $X_1$ follows a Bernoulli($p_1$) and $X_2$ follows a Bernoulli($p_2$), then sum of these two random variables has a mean of $p_1 + p_2$, and variance $p_1 q_1 + p_2 q_2$.

Elementary Track
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Track
Bernoulli Distribution – In Depth Discussion

# Bernoulli Measure

## Required review

The following is a more advanced treatment of the Bernoulli distribution. It will also be good to review the following sections

An Introduction to the Concept of Measure
Elementary Set Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution

In the previous section that discussed sampling with and without replacement, we set up a series of experiments that led to the computation of probability without actually collecting any data. This focused on sampling from a particular type of experiment whose characteristics were as follows:

- The result of the experiment could be dichotomized, typically characterized as either "success" or "failure or "1" or "0".
- The probability of a success is the known quantity, *p*.
- The probability *p* does not change its value from experiment to experiment.
- The result of a prior experiment does not influence the result of a subsequent experiment.

Experiments that have all four properties are known as Bernoulli trials, named for Jacob Bernoulli, from the famed Bernoulli family.

Bernoulli trials are ubiquitous in our culture. That and the ease of probability computations based on them makes this among the most popular and easily understood probability distributions. We will use the concept of the Bernoulli distribution to introduce us to the features of measures used in probability in general, and move on from there to describe some of their characteristics.

## Characterizing probability distributions.

While we can qualitatively describe distributions by the experiment, e.g., Bernoulli trials, we will need to be able to provide a mathematical characterization as well, which requires us to detail the outcome of the experiment. Remembering, that this characterization is based on a measurable space (Ω, Σ, **P**). we describe any particular outcome of the experiment as $X(\omega)$. $X(\omega)$ we can think of as both an event and a measurable function from the σ-algebra. Sometimes, all that is required is the qualitative description of the experiment, such as, "death or no death", a depiction that provides all that we need to describe the σ-algebra. To characterize a probability measure, we simply need to provide all possible experimental outcomes and the probabilities attached to them.

## Random variables

We have learned to work with variables in mathematics. Typically we think of a variable $x$ as being flexible (we can define it anyway we want, hopefully in a way that is helpful), but once defined, is fixed (e.g., let $x$ be the height of the new hospital), and of course is initially unknown (i.e., we need to find or "solve for $x$").

A random variable is the outcome of an experiment whose conduct if fixed but whose result is not (i.e., did the patient have a stroke or not). Since its value depends on the outcome of an experiment that is in the sample space Ω we call the variable not just $x$, but $X(\omega)$. Its reliance on members $\omega \subset \Omega$ make it a measurable function. For the random variable $X(\omega)$, we can choose and govern the rules by which it is assigned, but we cannot set its value − the outcome of the experiment determines that.

For example, in the Bernoulli model, we may define the experiment as the flip of a fair coin. Then the sample space $\Omega$ contains all possible outcomes of the experiment (including the null set). This is simply "a head", "a tail", or $\varnothing$.

The σ-algebra $\Sigma$, the set of all unions, complements, and intersections of the members of $\Omega$ is also quite simple.

We now define a random variable, $X(\omega)$ in a way that it is measurable with respect to $(\Omega,\Sigma)$. We choose the values $X(\omega) = 1$ if the result is a head, or $X(\omega) = 0$ if the experiment produces a tail.

Recall that the random variable $X(\omega)$ is measurable on Ω if every value that $X(\omega)$ takes can be mapped back to a member of the σ-algebra $\Sigma$. In this case $X(\omega) = 1$ maps back the result of a "head" and $X(\omega) = 0$ maps back the result of a "tail" which are each members of Σ which contain the outcomes of the Bernoulli trial. This feature is the heart of a measurable function. We say that $X(\omega)$ is a measurable function of $(\Omega,\Sigma)$.

Note, that once that assignment rule is made, the actual value of $X(\omega)$ is determined by the experiment. However, although we do not know its value in advance, we will see that we can define its average, understand the variability in its possible values, and perhaps even understand its long term behavior.

For a Bernoulli trial we characterize the possible outcomes $X(\omega)$ as either 0, or 1. We can then write

$$\begin{aligned}\mathbf{P}[X=1]&=p\\ \mathbf{P}[X=0]&=q=1-p.\end{aligned}$$

This is all we need to characterize the Bernoulli distribution. Since the argument "($\omega$)" is implied for all random variables we will typically not refer to it explicitly so as to simplify notation, but keep in mind that it is a measurable function on $(\Omega,\Sigma)$.

We can define the function that assigns measure or probability to each value of the random variable $X$ is $\mathbf{P}[X] = p\mathbf{1}_{x=1} + (1-p)\mathbf{1}_{x=0}$ . $X$ is the random variable, the outcome of the experiment in general, and and we define its probability as $\mathbf{P}[X]$ for all possible values of $X$.

Note how the indicator function works here. When $x = 1$ only the value *p* remains on the right hand side of the equation, and $\mathbf{P}[X] = p$. Similarly, when $x = 0$, the value $p$ disappears and $1 - p$ emerges.

We can use this notation to assure ourselves that probability sums to one over the entire sample space. Our intuition tells us this must be so since there are only two possible outcomes, $X(\omega) = 1$ with probability *p* and $X(\omega) = 0$ with probability 1 – *p*, and, since they are mutually exclusive $p + 1 - p = 1$. However, we can apply what we learned from measure and integration to demonstrate this formally.

## Cumulative distribution function

Once we have this function, $\mathbf{P}[X] = p\mathbf{1}_{x=1} + (1 - p)\mathbf{1}_{x=0}$ known as a probability mass function, or a measuring tool, there are other quantities that we can compute. One is the cumulative distribution function, written as

$$\mathbf{F}_X(x) = \mathbf{P}[X \le x].$$

In this notation, *X* represents the random variable, and the lower case *x* is any value that the random variable $X$ can take. For example, $\mathbf{F}_X(4) = \mathbf{P}[X \le 4]$. Once we know that *X* follows a Bernoulli distribution we can write

$$\begin{aligned}\mathbf{F}_X(x) &= 0 \text{ for } x < 1 \\ &= 1 \text{ for } x \ge 1.\end{aligned}$$

and depict this distribution (Figure 1). Note the discontinuity at $x = 1$.

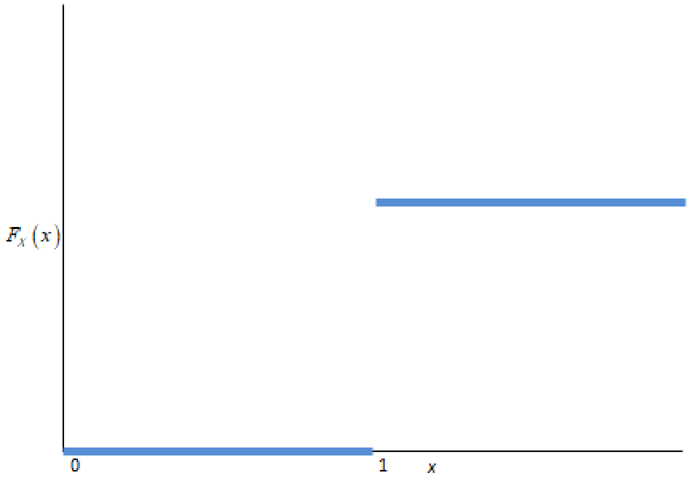


**Figure 1**. The cumulative distribution function for the Bernoulli distribution

## Expectation

Another quantity of interest is the expected value of *X,* which we note as $\mathbf{E}[X]$. It is a weighed sum of all of the possible values the random variable can take, where the weights are the probabilities of these values. We write this as

$$\mathbf{E}\left[X\right]=\int_{\Omega} x d\mathbf{P}.$$

From our work with the Lebesgue integral, we know how to accumulate Bernoulli measure over this integral. For each value of $X = x$, we multiply by the measure of that value and accumulate, then move on to the next value of $X$.

For the Bernoulli distribution, this becomes

$$\mathbf{E}\left[X\right]=\int_{\Omega_x} x\,\mathrm{d}\mathbf{P}=0\mathbf{P}\left[X=0\right]+1\mathbf{P}\left[X=1\right]=0q+1p=p$$

The $\mathrm{d}\mathbf{P}$ component requires us to ignore all values of the random variable $X$ that assigned probability 0. The expected value is a measure of central tendency of the probability distribution **P.** The expectation itself is commonly referred to by $\mu$. Thus, we see that the expectation, $\mu$ is a function of the parameter $p$.

We can also find the expected value of functions of the random variable. Let $c$ be a known constant, and compute $W = cX$ as a new random variable[*], and we should be able to find $\mathbf{E}\left[W\right]$. We write

$$\mathbf{E}\left[W\right]=\mathbf{E}\left[cX\right]=\int_{\Omega} cX\,d\mathbf{P}=c\int_{\Omega} X\,d\mathbf{P}=c\mathbf{E}\left[X\right]$$

We know from our work on properties of the Lebesgue integral that $\int_{\Omega} cX\,d\mathbf{P}=c\int_{\Omega} X\,d\mathbf{P}$, the equations proceeds smoothly.

A simple equality that will give us some practice in manipulating both summations and the integral stems from an examination of the quantity $X-\mathbf{E}\left[X\right]$. Keep in mind that $X$ is random, taking on values with probability governed by **P.** On the other hand, $\mathbf{E}\left[X\right]$ is not random, but fixed, and is a function of the parameters of the probability distribution **P,** which in the case of the Bernoulli distribution, is $p$.

So, the quantity $X-\mathbf{E}\left[X\right]$ is also random. Let's try to find $\mathbf{E}\left[X-\mathbf{E}\left[X\right]\right]$. If we write expectation using summation, we would write

$$\mathbf{E}\left[X-\mathbf{E}\left[X\right]\right]=\sum_{all\,x}\left[x-\sum_{all\,x}x\mathbf{P}\left[X=x\right]\right]\mathbf{P}\left[X=x\right]$$

The key to the simplification here is to see that the expression $\sum_{all\,x}x\mathbf{P}\left[X=x\right]$ while a sum over $x$, once the sum is taken, is no longer a function of $x$. In fact, we can just write this as the constant $\mu$, simplifying to

[*] Clearly $W$ is measurable, since for every value of $W = w$ we can find an $\frac{w}{c}$ which has a member in the σ-algebra $\Sigma$.

$$\mathbf{E}\left[X-\mathbf{E}[X]\right]=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]$$

Now we just distribute the summation sign to write

$$\begin{aligned}\mathbf{E}\left[X-\mathbf{E}[X]\right]&=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]\\&=\sum_{all\,x}x\mathbf{P}[X=x]-\sum_{all\,x}\mu\mathbf{P}[X=x].\end{aligned}$$

The first term on the second line of the above equation is $\mu$. The second term involves $\mu$, but since $\mu$ is a constant with respect to the sum over all possible values of $x$, it can be moved to outside the summation sign. Thus we have

$$\begin{aligned}\mathbf{E}\left[X-\mathbf{E}[X]\right]&=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]\\&=\sum_{all\,x}x\mathbf{P}[X=x]-\sum_{all\,x}\mu\mathbf{P}[X=x]\\&=\mu-\mu\sum_{all\,x}\mathbf{P}[X=x]\end{aligned}$$

Observing that $\sum_{all\,x}\mathbf{P}[X=x]=1$ permits us to finish the derivation.

$$\begin{aligned}\mathbf{E}\left[X-\mathbf{E}[X]\right]&=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]=\sum_{all\,x}[x-\mu]\mathbf{P}[X=x]\\&=\sum_{all\,x}x\mathbf{P}[X=x]-\sum_{all\,x}\mu\mathbf{P}[X=x]\\&=\mu-\mu\sum_{all\,x}\mathbf{P}[X=x]\\&=\mu-\mu\\&=0.\end{aligned}$$

We can also approach $\mathbf{E}\left[X-\mathbf{E}[X]\right]$ using our notion of the integral from measure theory. Remembering that the mean is a constant with respect to $d\mathbf{P}$**,** we can say that

$$\mathbf{E}\left[X-\mathbf{E}[X]\right]=\int_{\Omega_x}\left(x-\mathbf{E}[X]\right)\mathrm{d}\mathbf{P}=\int_{\Omega_x}\left(x-\int_{\Omega_x}x\mathrm{d}\mathbf{P}\right)\mathrm{d}\mathbf{P}$$

$$=\int_{\Omega_x}(x-\mu)\mathrm{d}\mathbf{P}=\int_{\Omega_x}x\,\mathrm{d}\mathbf{P}-\int_{\Omega_x}\mu\,\mathrm{d}\mathbf{P}=\mu-\mu\int_{\Omega_x}\mathrm{d}\mathbf{P}=\mu-\mu=0.$$

Use of the probability measure d**P** ensures that we only consider values of $X$ with non-zero probability.*

One way to think about the zero value for this expression $\mathbf{E}\left[X-\mathbf{E}[X]\right]=\mathbf{E}[X-\mu]$ is that, since $X$ is random, values of $X$ can be larger than $\mu$ while other values of $X$ can be smaller. In the long term, that is for large sample sizes, we might expect these large values of $X$

* This computation assumes that $\mathbf{E}[X]$ exixts. We will see later that some probability distributions are so disbursed that they do not even have this expectation.

to be balanced by small values of $X$. Thus, their long run average will be $\mathbf{E}[X]$, and $\mathbf{E}\left[X-\mathbf{E}[X]\right]$ will be zero.[*]

Following in the same vein, we can write

$$\mathbf{E}\left[X^2\right]=\int_{\Omega_x} x\,\mathrm{d}\mathbf{P}=0^2\,\mathbf{P}[X=0]+1^2\,\mathbf{P}[X=1]=0q+1p=p$$

We call $\mathbf{E}[X]$ and $\mathbf{E}\left[X^2\right]$ as the first and second moments of the random variable $X$. For the Bernoulli distribution, they happen to be equal.

## Variance

$\mathbf{E}[X]$ is an important measure of a central tendency. A useful measure of dispersion of $X$ is called the variance of $X$, $\mathbf{Var}[X]$, The $\mathbf{Var}[X]$ is defined as

$$\mathbf{Var}[X]=\mathbf{E}\left[(X-\mu)^2\right]$$

$\mathbf{Var}[X]$ commonly written as $\sigma^2$, and the standard deviation of $X$ is the square root of the variance, is σ. One advantage of the standard deviation is that its units are the same as the units of the original random variable.

We already know that $\mathbf{E}[X-\mu]=0.$ The variance with its square term is only zero in the special circumstance when every realized value of the random variable $X$ is the same, i.e., $x_i = c$ for all $i$. In general, we can write

$$\mathbf{Var}(X)=\mathbf{E}\left[(X-\mu)^2\right]=\mathbf{E}\left[X^2-2X\mu+\mu^2\right].$$

Passing the expectation through the terms on the right hand side we find

$$\begin{aligned}\mathbf{Var}(X)&=\mathbf{E}\left[(X-\mu)^2\right]=\mathbf{E}\left[X^2-2X\mu+\mu^2\right]\\&=\mathbf{E}\left[X^2\right]-2\mu^2+\mu^2\\&=\mathbf{E}\left[X^2\right]-\mu^2.\end{aligned}$$

This is one of the most useful formulas for $\mathbf{Var}[X]$. To find the variance for the Bernoulli distribution we first need to find the $\mathbf{E}\left[X^2\right]$. We saw earlier that

$$\mathbf{E}[X]=\int_{\Omega_x} x\,\mathrm{d}\mathbf{P}=0\mathbf{P}[X=0]+1\mathbf{P}[X=1]=0q+1p=p$$

$$\mathbf{E}\left[X^2\right]=\int_{\Omega_x} x\,\mathrm{d}\mathbf{P}=0^2\,\mathbf{P}[X=0]+1^2\,\mathbf{P}[X=1]=0q+1p=p$$

so we may write

[*] We will have a lot more to say about long run averages when we discuss asymptotic theory.

$$\sigma_X^2 = \mathbf{Var}(X) = \mathbf{E}\left[X^2\right] - \mu^2 = p - p^2 = p(1-p) = pq.$$
$$\sigma_X = \mathbf{SD}(X) = \sqrt{pq}.$$

A the expected value, $\mathbf{E}[X]$ is a measure of central tendency, $\mathbf{Var}(X)$ is a measure of dispersion.

Assume for example, that we have a Bernoulli distribution with $p = 0.01$. This is distribution where the 0's are generating in overwhelming numbers. We might therefore anticipate that the expected value will be small, but also the variance is small as well. We can compute $\mathbf{E}\left[X\right] = p = 0.01$, and $\sigma_X = \sqrt{pq} = \sqrt{(0.01)(0.99)} = \sqrt{0.0099} = 0.099$.

Now suppose *Y* is another random variable following the Bernoulli distribution, but this time, the parameter *p* = 0.50. In this case we would expect just as many zeros as ones generated with no preponderance of one value over the other. The "spread" of the values is greater than for the random variable *X*. We compute $\mathbf{E}\left[Y\right] = p = 0.50$, and its standard deviation $\sigma_X = \sqrt{pq} = \sqrt{(0.50)(0.50)} = 0.50.$ This is almost fifty times the standard deviation of the Bernoulli (*p*=0.01) random variable.

We can also other quantities that are of interest based on the different expectations or moments of the Bernoulli random variable *X*. We can write skewness as

$$S(X) = \frac{\mathbf{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

Skewness may be thought of as lack of symmetry of the distribution.

Kurtosis expresses how heavy the tails of the distribution are. A platykurtic distribution has thin tails (i.e., tails that have a relatively small degree of measure), while a leptokurtic distribution has tails that are heavier. The formula for kurtosis is*

$$\mathbf{K}_a(X) = \mathbf{E}\left[(X-\mu)^4\right] - 3$$

Some algebra reveals that the Bernoulli distribution,

$$\mathbf{S}(X) = \frac{(q-p)}{\sqrt{pq}}, \text{ and } \mathbf{K}_a(X) = \frac{(1-6pq)}{pq}.$$

## Generating functions

While generating functions are covered in detail, some brief introductory comments will be made here.

Two other concepts useful in characterizing generating functions are the moment generating function $\mathbf{M}_X(t)$, and the probability generating function, $\mathbf{G}_X(s)$. Generating functions provide an independent track of doing what otherwise would be complicated calculations. We should consider these as tools as in our probability tool kit – sometimes they are indispensible, while other times we can do the job without them.

* The subtraction of 3 is a correction for the kurtosis of the normal distribution.

$\mathbf{M}_X(t)$ is an expectation of a function of the random variable *X*, the function being $e^{tX}$. If *X* follows a Bernoulli(*p*) distribution, then we simply compute

$$\mathbf{M}_X(t) = \mathbf{E}\left[e^{tx}\right] = e^{t(0)}\left(1-p\right) + e^{t(1)}p = \left(1-p\right) + pe^t.$$

It's hallmark is that the derivative of the moment generating function, for *t*= 0 provides the mean of the distribution. In this simple case,

$$\frac{d\mathbf{M}_X(t)}{dt} = \frac{d\left(\left(1-p\right)+pe^t\right)}{dt} = pe^t$$

which is simply *p* for *t* = 0. This may seem to be "the long way around" to find the mean for the Bernoulli distribution, but we will see that for complicated distributions, this is a useful way to identify its moments.

Another use of the moment generating function is that it uniquely identities the distribution. Once we have the moment generating function, we then have the distribution itself. Finally, when we study asymptotic properties of distributions, identifying the large sample behavior and ultimately appearance of $\mathbf{M}_X(t)$ will allow us to identify the distribution itself. This is the principle behind the proof of one of the most useful theorems in probability, the Central Limit Theorem.

## Bernoulli probability generating function

Another related quantity of interest is the probability generating function. Just as the moment generating function generates moments, the probability generating function generates probabilities. It is defined as

$$\mathbf{G}_X(s) = \mathbf{E}\left[s^X\right].$$

For the Bernoulli distribution, we find

$$\mathbf{G}_X(s) = \mathbf{E}\left[s^X\right] = \left(1-p\right)s^0 + ps^1 = 1 - p + ps = 1 - p\left(1-s\right).$$

We will rely on this finding when we get to the probability generating function for the binomial distribution.

## Addition of Bernoulli random variables

Finding the probability distribution of the sum of random variables may be complicated. However, if all we need are the moments of the sum of two random variables, the work is substantially easier. For example, Let $X_1$ and $X_2$ be two random variables. Then we can compute the expectation of the sum of them $W = X_1 + X_2$

$$\mathbf{E}\left[W\right] = \int_\Omega w\,d\mathbf{P} = \int_\Omega \left(x_1 + x_2\right)d\mathbf{P} = \int_\Omega x_1 d\mathbf{P} + \int_\Omega x_2\,d\mathbf{P}$$
$$= \mathbf{E}\left[X_1\right] + \mathbf{E}\left[X_2\right].$$

What makes this computation work is the reversibility of the integral and the sum in the term $x_1 + x_2$. This is a basic property of the Lebesgue integral. This computation works finite and infinite* sums of random variables.

If the random variables are independent, we can also sum their variances. Thus

$$\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2].$$

To see this, let's write $W = X_1 + X_2$. Then we know $\mathbf{Var}[W] = \mathbf{E}\left[W^2\right] - \mathbf{E}^2[W]$. Therefore

$$\mathbf{Var}[X_1 + X_2] = \mathbf{E}\left[(X_1 + X_2)^2\right] - \left[\mathbf{E}[X_1] + \mathbf{E}[X_2]\right]^2$$
$$= \mathbf{E}\left[X_1^2 + X_2^2 + 2X_1X_2\right] - \mathbf{E}^2[X_1] - \mathbf{E}^2[X_2] - 2\mathbf{E}[X_1]\mathbf{E}[X_2]$$

Since we know expectation passes through sums, we can write

$$\begin{aligned}
&\mathbf{E}\left[X_1^2 + X_2^2 + 2X_1X_2\right] - \mathbf{E}^2[X_1] - \mathbf{E}^2[X_2] - 2\mathbf{E}[X_1]\mathbf{E}[X_2] \\
&= \mathbf{E}\left[X_1^2\right] + \mathbf{E}\left[X_2^2\right] + 2\mathbf{E}[X_1X_2] - \mathbf{E}^2[X_1] - \mathbf{E}^2[X_2] - 2\mathbf{E}[X_1]\mathbf{E}[X_2] \\
&= \mathbf{E}\left[X_1^2\right] - \mathbf{E}^2[X_1] + \mathbf{E}\left[X_2^2\right] - \mathbf{E}^2[X_2] + 2\mathbf{E}[X_1X_2] - 2\mathbf{E}[X_1]\mathbf{E}[X_2] \\
&= \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\mathbf{E}[X_1X_2] - 2\mathbf{E}[X_1]\mathbf{E}[X_2].
\end{aligned}$$

Since the random variables $X_1$ and $X_2$ are independent, $\mathbf{E}[X_1X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$. Thus the variance of a sum of independent random variables is the sum of the variances.

Let's look at the last of this development involving $\mathbf{E}[X_1X_2]$.

As we have seen, our solution for $\mathbf{Var}[X_1 + X_2]$ becomes easier if $\mathbf{E}[X_1X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$. However, this assertion is not always the case.

Yet, it is a property of independence. Just as we saw that if $\mathbf{P}[X_1 \cap X_2] = \mathbf{P}[X_1]\mathbf{P}[X_2]$, then similarly $\mathbf{E}[X_1X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$, This is because the joint measure of $X_1$ and $X_2$, $\mathbf{P}$, factors into a measure for $X_1$, $\mathbf{P}_1$ and a measure for $X_2$, $\mathbf{P}_2$. Thus

$$\begin{aligned}
\mathbf{E}[X_1X_2] &= \int_{\Omega} X_1X_2 d\mathbf{P} = \int_{\Omega} X_1X_2 d\mathbf{P}_1 d\mathbf{P}_2 \\
&= \int_{\Omega_1} X_1 d\mathbf{P}_1 \int_{\Omega_2} X_2 d\mathbf{P}_2 = \mathbf{E}[X_1]\mathbf{E}[X_2].
\end{aligned}$$

With this final step, we see that the variance of a sum of independent random variables is the sum of the variances.

Thus, if $X_1$ follows a Bernoulli($p_1$) and $X_2$ follows a Bernoulli($p_2$), then sum of these two random variables has a mean of $p_1 + p_2$, and variance $p_1q_1 + p_2q_2$.

Suppose we were dealing with not the sum but instead $X_1 - X_2$? Well, the expectations subtract as we might expect. However, the finding for the variance is surprising.

* By infinite here, we mean $\sum_{k}^{\infty} x_k$ where the summands are indexed by the integers. Another way to say this is that the summands are countable.

$$\mathbf{Var}\left[X_1 - X_2\right] = \mathbf{E}\left[\left(X_1 - X_2\right)^2\right] - \left[\mathbf{E}\left[X_1\right] \text{-} \mathbf{E}\left[X_2\right]\right]^2$$
$$= \mathbf{E}\left[X_1^2 - X_2^2 - 2X_1X_2\right] - \mathbf{E}^2\left[X_1\right] - \mathbf{E}^2\left[X_2\right] + 2\mathbf{E}\left[X_1\right]\mathbf{E}\left[X_2\right]$$
$$= \mathbf{E}\left[X_1^2\right] + \mathbf{E}\left[X_2^2\right] - 2\mathbf{E}\left[X_1X_2\right] - \mathbf{E}^2\left[X_1\right] - \mathbf{E}^2\left[X_2\right] + 2\mathbf{E}\left[X_1\right]\mathbf{E}\left[X_2\right]$$
$$= \left(\mathbf{E}\left[X_1^2\right] - \mathbf{E}^2\left[X_1\right]\right) + \left(\mathbf{E}\left[X_2^2\right] - \mathbf{E}^2\left[X_2\right]\right) - 2\mathbf{E}\left[X_1X_2\right] + 2\mathbf{E}\left[X_1\right]\mathbf{E}\left[X_2\right]$$
$$= \mathbf{Var}\left[X_1\right] + \mathbf{Var}\left[X_2\right] - 2\left[\mathbf{E}\left[X_1X_2\right] - \mathbf{E}\left[X_1\right]\mathbf{E}\left[X_2\right]\right].$$

If independent, then the variances of the random variables whose difference we wish to take add – they do not subtract. However, when there is dependence, substantial reduction in the variance can be achieved.

## Practical use of Lebesgue integration

Let's assume that a viral epidemic is spreading through a community. This infection can kill, but the probability that a patient survives depends on the age of the patient. If the patient is young, the probability the patient survives is $p_1$. If they are young or middle aged, this probability changes to $p_2$. Older patients have probability $p_3$ of survival

What is the probability, $\mathbf{P}_S$ of survival for a patient?

Using a Bernoulli model, the easiest answer to this question is that survival depends on the age epoch of the subject, However, in order to compute the probability of death without knowledge of age, we need to know the measures of age, i.e., how likely an individual will be in each of these three epochs. Let's call these probabilities, $p_c$, $p_m$, and $p_o$ respectively.

Then if we define our random variable as Bernoulli, we can write our solution as $\mathbf{P}_S = \mathbf{P}\left[X = 0\right]$.

The patient can be a child, and the child survives, or the patient can be young/middle, and the patient dies, or the patient can be older, and the child dies. This is a problem in conditional probability, and recalling the Law of Total Probability, we can write $\mathbf{P}_S = p_c p_1 + p_m p_2 + p_O p_3$.

However, we can also examine this as a problem in Bernoulli trials. Recall that Lebesgue integration requires that we first aggregate a common event and measure it. The common event is survival, be the patient a child, young/middle, or older. These three events are a survival and we aggregate them. There probability is $\mathbf{P}_S = \int X\left(\omega\right) d\mathbf{P}$. So what are the ways that $X = 1$?

$$\mathbf{P}_S = \int X\left(\omega\right) d\mathbf{P} = (1)\left(p_c p_1 + p_m p_2 + p_O p_3\right).$$

The Bernoulli Brothers
Skewness and Kurtosis for the Bernoulli Distribution.
Moment Generating and Probability Generating Functions

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Moment and Probability Generating Functions

Moment and probability generating functions provide a unique perspective on the use of probability distributions. They are quite powerful techniques; however, although the mathematics are quite precise, manipulating generating functions can sometimes seem far afield and often contribute little insight into understanding the solution for the underlying probability question at hand. However, like stepping into the fourth dimension to solve a three dimensional problem, these functions bring a new perspective to probability problems.

Prerequisites
Factorials Permutations, and Combinations
Binomial Theorem
The Concept of the Limit
Convergent Series
Exponential Functions

## Foundation of MGF's

The basis of the moment generating function (MGF) is the infinite series expansion of the function $e^{tx}$. We write this as

$$e^{tx} = 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \ldots +$$

We write $\mathbf{M}_X(t)$ as

$$\mathbf{M}_X(t) = \mathbf{E}\left[e^{tx}\right] = \int_{\Omega_x} \left(1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \ldots +\right) d\mathbf{P}$$

$$= \sum_{k=0}^{\infty} \int_{\Omega_x} \frac{(tx)^k}{k!} d\mathbf{P} = \sum_{k=0}^{\infty} \int_{\Omega_x} \frac{t^k}{k!} x^k d\mathbf{P} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{\Omega_x} x^k d\mathbf{P}$$

$$= 1 + t\mathbf{E}[X] + \frac{t^2}{2!}\mathbf{E}\left[X^2\right] + \frac{t^3}{3!}\mathbf{E}\left[X^3\right] + \ldots +$$

This is the heart of the moment generating function. Each term in the final expression involves a moment of the distribution of the random variable *X*. If we take one derivative of $\mathbf{M}_X(t)$ with respect to *t*, we find

$$\mathbf{M}_X(t) = 1 + t\mathbf{E}[X] + \frac{t^2}{2!}\mathbf{E}\left[X^2\right] + \frac{t^3}{3!}\mathbf{E}\left[X^3\right] + \ldots +$$

$$\frac{d\mathbf{M}_X(t)}{dt} = 0 + \mathbf{E}[X] + \frac{2t^1}{2!}\mathbf{E}\left[X^2\right] + \frac{3t^2}{3!}\mathbf{E}\left[X^3\right] + \ldots +$$

and, evaluating this derivative where *t* = 0 we find $\left.\frac{d\mathbf{M}_X(t)}{dt}\right|_{t=0} = \mathbf{E}[X]$. A second derivative produces

$$\frac{d^2\mathbf{M}_X(t)}{dt^2} = \frac{2}{2!}\mathbf{E}\left[X^2\right] + \frac{(3)2t}{3!}\mathbf{E}\left[X^3\right] + \ldots +$$

and $\left.\frac{d^2\mathbf{M}_X(t)}{dt^2}\right|_{t=0} = \mathbf{E}\left[X^2\right]$.

Continuing in this fashion, we can find all of the nonnegative integer valued moments of the random variable *X*. This infinite collection of moments is enough to completely specify the probability distribution of *X*, i.e., no two distinct collection of probability distributions can have all of their moments the same. Thus, knowing $\mathbf{M}_X(t)$ is mathematically equivalent to knowing the probability distribution of $X$. Also, in some circumstances, taking the derivative of the $\mathbf{M}_X(t)$ is easier than integrating the probability measure $\mathbf{P}$, and provides an alternative procedure for identifying the measure's moments.

## Probability generating functions (PGFs)

The probability generating function $\mathbf{G}_X(s)$ is developed similarly.

$$\mathbf{G}_X(s) = \mathbf{E}\left[s^X\right] = \int_{\Omega_x} s^X d\mathbf{P} = \sum_{k=0}^{\infty} s^k \mathbf{P}_k(t)$$

The derivative with respect to *s* is

$$\frac{d\,\mathbf{G}_X(s)}{ds} = \frac{d\,\mathbf{E}\left[s^X\right]}{ds} = \frac{d\sum_{k=0}^{\infty} s^k \mathbf{P}_k(t)}{ds} = \sum_{k=0}^{\infty} \frac{ds^k}{ds}\mathbf{P}_k(t)$$

Note the reversal of the derivative and summation sign in $\frac{d\sum_{k=0}^{\infty} s^k \mathbf{P}_k(t)}{ds} = \sum_{k=0}^{\infty} \frac{ds^k}{ds} \mathbf{P}_k(t)$, a procedure that is only permissible if the sum $\sum_{k=0}^{\infty} s^k \mathbf{P}_k(t)$ is uniformly convergence. However, we can see that the uniform continuity of generating functions is demonstrated by

$\sum_{k=0}^{\infty} s^k \mathbf{P}_k(t) \leq \sum_{k=0}^{\infty} \mathbf{P}_k(t) = 1$ for $0 \leq s < 1$.[*]

We evaluate $\frac{ds^k}{ds} = ks^{k-1}$ which, when evaluated at $s$ = 1 is $k$. Thus

$$\left. \frac{d\,\mathbf{G}_X(s)}{ds} \right|_{s=0} = \int_{\Omega_x} x\,d\mathbf{P} = \mathbf{E}[X]$$

The second moment provides an unusual moment

$$\begin{aligned}
\frac{d^2\mathbf{G}_X(s)}{ds^2} &= \sum_{k=0}^{\infty} \frac{d\left(ks^{k-1}\right)}{ds} \mathbf{P}_k(t) = \sum_{k=0}^{\infty} k(k-1)s^{k-2}\mathbf{P}_k(t) \\
\left. \frac{d^2\mathbf{G}_X(s)}{ds^2} \right|_{s=1} &= \left. \sum_{k=0}^{\infty} k(k-1)s^{k-2}\mathbf{P}_k(t) \right|_{s=1} \\
&= \sum_{k=0}^{\infty} k(k-1)\mathbf{P}_k(t) = \mathbf{E}[x(x-1)].
\end{aligned}$$

So, while successive derivatives of $\mathbf{M}_X(t)$ provide the moments of the random variable $X$, successive derivatives of $\mathbf{G}_X(s)$ provide the factorial moments. We will find these factorial moments quite useful for the binomial, Poisson, and negative binomial distributions. In fact, probability generating functions are particularly useful for discrete distributions.

## Continuity theorem

The probability generating function is the basis on which the utility of generating functions rests. A theorem known as the Continuity Theorem allows us to link generating functions and probability distributions. Let's begin this examination by noting another property of $\mathbf{G}_X(s)$.

$$\begin{aligned}
\mathbf{G}_X(s) &= \mathbf{E}\left[s^X\right] = \sum_{k=0}^{\infty} s^k \mathbf{P}[X = k] = \sum_{k=0}^{\infty} s^k \mathbf{P}_k \\
\frac{d\mathbf{G}_X(s)}{ds} &= \sum_{k=0}^{\infty} ks^{k-1}\mathbf{P}_k = 0s^{-1}\mathbf{P}_0 + 1s^0\mathbf{P}_1 + 2s^1\mathbf{P}_2 + 3s^2\mathbf{P}_3 + \dots \\
\left. \frac{d\mathbf{G}_X(s)}{ds} \right|_{s=0} &= \mathbf{P}_1.
\end{aligned}$$

[*] This assessment is formally known as the Wierstrauss $M$ test.

We can take a second derivative to observe

$$\frac{d^2\mathbf{G}_X(s)}{ds^2} = \sum_{k=0}^{\infty} k\left(k-1\right)s^{k-2}\mathbf{P}_k = 0(-1)s^{-2}\mathbf{P}_0 + (1)(0)s^{-1}\mathbf{P}_1 + (2)(1)s^0\mathbf{P}_2$$
$$+ (3)(2)s^1\mathbf{P}_3 + \ldots$$

$$\left.\frac{d^2\mathbf{G}_X(s)}{ds^2}\right|_{s=0} = 2!\mathbf{P}_2.$$

In general $\left.\frac{d^m\mathbf{G}_X(s)}{ds^m m!}\right|_{s=0} = \mathbf{P}\left[X = m\right]$. The entire sequence of probabilities $\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \ldots$ is generated from $\mathbf{G}_X(s)$. Thus, while successive derivatives identify moments for the moment generating function $\mathbf{M}_X(t)$, they identify probabilities for the probability generating function $\mathbf{G}_X(s)$.

Now, a result from analysis asserts that if there are two infinite series, say $A(s) = \sum_{k=0}^{\infty} s^k a_k$, and $B(s) = \sum_{k=0}^{\infty} s^k b_k$ and there is an open interval of set of values of $s$, such that on that interval $A(s) = B(s)$, then the $a_k$'s and $b_k$'s must also match, i.e., $a_k = b_k$ for all $k \geq 0$.

This finding from analysis can have wide ranging implications for the practice of probability. Now suppose we have two probability generating functions $\mathbf{G}_X(s) = \sum_{k=0}^{\infty} s^k\mathbf{P}_k$, and $\mathbf{G}_Y(s) = \sum_{k=0}^{\infty} s^k\mathbf{R}_k$. Then, if there exists a region of $s$ (e.g., $-1 < s < 1$) in which $\mathbf{G}_X(s) = \mathbf{G}_Y(s)$, then $\mathbf{P}_k = \mathbf{R}_k$ for all $k \geq 0$. This is the basis of the continuity theorem.

## Generating Function Inversion

We understand that $\mathbf{G}_X(s) = \sum_{k=0}^{\infty} s^k\mathbf{P}_k$. So given $\mathbf{G}_X(s)$, we can simply takes its successive derivatives for $s = 0$ in order to identify successive probabilities. However, in many cases, we can actually inspect the functional form of $\mathbf{G}_X(s)$ to identify the set of probabilities $\left\{\mathbf{P}_k\right\}$ directly. This process of moving directly from $\mathbf{G}_X(s)$ to the set of probabilities is termed inversion.

Inverting generating functions is typically based on observation and simple deduction. Practice facilitates their use.

## A first generating function

We begin with a very simple infinite series, and perhaps the simplest of all generating functions. Let $S = \sum_{k=0}^{\infty} s^k$. The coefficient of $s^k$ is simply 1 here, i.e. $S = \sum_{k=0}^{\infty} 1s^k$. What is the generating function?*

We can simply write (as appears in the discussion of the

$$S_k = 1 + s + s^2 + s^3 + s^4 + \ldots + s^k.$$
$$sS_k = s + s^2 + s^3 + s^4 + s^5 + \ldots + s^{k+1}$$

* This follows the development of Moyé and Kapadia, Difference Equations with Public Health Applications, Marcel-Dekker. 2000.

Subtract the second partial sum from the first to find $\left(1-s\right)S_k = 1-s^{k+1}$, or $S_k = \frac{1-s^{k+1}}{\left(1-s\right)}$. Now we take a limit to find $\lim\limits_{k\to\infty} S_k = \lim\limits_{k\to\infty}\left(\frac{1-s^{k+1}}{\left(1-s\right)}\right) = \frac{1}{1-s}$.[*]

Thus, we write $\mathbf{G}(s) = \frac{1}{1-s} = \sum\limits_{k=0}^{\infty} s^k$. The coefficient of $s^k$ for $\mathbf{G}(s)$ is 1[†]. We introduce the notation $\triangleright$ as meaning "generates the family coefficients of $s^k$ as equal to", and write $\mathbf{G}(s) \triangleright \left\{1\right\}$.

However, suppose our sequence was not $S = \sum\limits_{k=0}^{\infty} s^k$, but $S = \sum\limits_{k=0}^{\infty} a^k s^k$. Then the approach above reveals

$$\begin{aligned} S_k &= 1 + as + \left(as\right)^2 + \left(as\right)^3 + \left(as\right)^4 + \ldots + \left(as\right)^k . \\ asS_k &= as + \left(as\right)^2 + \left(as\right)^3 + \left(as\right)^4 + \left(as\right)^5 + \ldots + \left(as\right)^{k+1} \end{aligned}$$

Subtracting the second partial sum from the first, we have $\left(1-as\right)S_k = 1-\left(as\right)^{k+1}$ or $S_k = \frac{1-\left(as\right)^{k+1}}{\left(1-as\right)}$. Taking limits as before we find $\lim\limits_{k\to\infty} S_k = \frac{1}{1-as}$.

So, if we are given $\mathbf{G}(s) = \frac{1}{1-as}$, then we can see that $\mathbf{G}(s) \triangleright \{a^k\}$, that is, the coefficient of $s^k$ is $a^k$.

For another simple example, let $\mathbf{G}(s) = \sum\limits_{k=0}^{\infty} d_k s^k$. Then $c\mathbf{G}(s) = c\sum\limits_{k=0}^{\infty} d_k s^k = \sum\limits_{k=0}^{\infty} cd_k s^k$, and $\mathbf{H}(s) = c\mathbf{G}(s) \triangleright \left\{cd_k\right\}$.

Typically, we will be given a generating function $\mathbf{G}(s)$ and have to identify the series it generates. For example, suppose we are given $\mathbf{G}(s) = \frac{1}{a+bs}$. What is the coefficient of $s^k$ (denoted $a^k$) for which $\mathbf{G}(s) = \sum\limits_{k=0}^{\infty} a_k s^k$? We write

$$\mathbf{G}(s) = \frac{1}{a+bs} = \frac{1/a}{1+\frac{b}{a}s} = \frac{1/a}{1-\left(\frac{-b}{a}\right)s} \triangleright \left\{\left(\frac{1}{a}\right)\left(\frac{-b}{a}\right)^k\right\}.$$

[*] Note that the continuity of $\frac{1-s^{k+1}}{\left(1-s\right)}$ permits us to pass the limit through the function to write

$$\lim_{k\to\infty}\left(\frac{1-s^{k+1}}{\left(1-s\right)}\right) = \frac{1-\lim\limits_{k\to\infty} s^{k+1}}{\left(1-s\right)} = \frac{1}{1-s}.$$

[†] This entire development is for $0 < s < 1$ to assure convergence of the series. Also note that this is not a probability generating function, just a generating function.

If $\mathbf{G}_1(s) \triangleright \{a_k\}$, and $\mathbf{G}_2(s) \triangleright \{b_k\}$, then the addition of coefficients of the same powers of $s^k$ gets us $\mathbf{G}(s) = \mathbf{G}_1(s) + \mathbf{G}_2(s) \triangleright \{a_k + b_k\}$.

## Multiplying generating functions

We can easily invert the product of two generating functions by collecting coefficients. Let $\mathbf{G}_1(s) \triangleright \{a_k\}$ and $\mathbf{G}_2(s) \triangleright \{b_k\}$. Then

$$\mathbf{G}(s) = \mathbf{G}_1(s)\ \mathbf{G}_2(s) = \left(\sum_{k=1}^{\infty} a_k s^k\right)\left(\sum_{k=1}^{\infty} b_k s^k\right)$$
$$= \left(a_0 + a_1 s + a_2 s^2 + a_3 s^3 + \ldots\right)\left(b_0 + b_1 s + b_2 s^2 + b_3 s^3 + \ldots\right).$$

and we simply need to collect coefficients. for example, the coefficient of $s^0$ is $a_0 b_0$. The coefficient of $s$ is $a_0 b_1 + a_1 b_0$. Continuing in this fashion, we find that

$$\mathbf{G}(s) \triangleright \left\{\sum_{j=0}^{k} a_j b_{k-j}\right\}.$$

## Difference equations

Difference equations are equations that describe the underlying structure or relationship between sequence elements; solving this family of equations means using the information about the sequence element interrelationship that is contained in the family to reveal the identity of members of the sequence. Generating functions are essential for their solution.

In its most general form a difference equation can be written as

$$a_0(k)y_{k+n} + a_1(k)y_{k+n-1} + a_2(k)y_{k+n-2} + \cdots + a_n(k)y_k = R(k)$$

It consists of terms involving members of the $\{y_k\}$ sequence, and, in addition, coefficients $a_j(k)$, which are the coefficients of the $\{y_k\}$ sequence in the equation. These coefficients may or may not be functions of *k*.

## Solutions using generating functions

Generating functions can be quite helpful in solving difference equations. Consider the very simple example,

$$y_{k+1} = a y_k - b$$

for $k = 0$ to $\infty$. While this family of equations can be solved easily by recursion, or induction, we will introduce the concept of the generating function approach.

Let's first choose a value of $s$ such that $0 < s < 1$, necessary for us to later interchange the summation and derivative signs (uniform convergence is required for this interchange), and write.

$$s^k y_{k+1} = a s^k y_k - b s^k$$

for $k = 0$ to $\infty$. We are interested in identifying the coefficient of $s^k$ in this equation for all $k$, so it is reasonable to consider solving for the generating function $\mathbf{G}(s) = \sum_{k=0}^{\infty} s^k y_k$ and invert it.

We continue by summing both the left and right hand sides of this equation to write.

$$\sum_{k=0}^{\infty} s^k y_{k+1} = a\sum_{k=0}^{\infty} s^k y_k - b\sum_{k=0}^{\infty} s^k.$$

The second summation we identify as $a\mathbf{G}(s)$. The first term we can write as

$$\begin{aligned}\sum_{k=0}^{\infty} s^k y_{k+1} &= s^{-1}\sum_{k=0}^{\infty} s^{k+1} y_{k+1} = s^{-1}\sum_{k=1}^{\infty} s^k y_k \\ &= s^{-1}\left[\sum_{k=0}^{\infty} s^k y_k - s^0 y_0\right] = s^{-1}\left[\mathbf{G}(s) - y_0\right]\end{aligned}$$

and the final term we see as $b\sum_{k=0}^{\infty} s^k = \frac{b}{1-s}$. We can now write the infinite collection of equations involving $s$ as a single equation involving $\mathbf{G}(s)$.

$$s^{-1}\left[\mathbf{G}(s) - y_0\right] = a\mathbf{G}(s) - \frac{b}{1-s}.$$

Solving for $\mathbf{G}(s)$ reveals

$$\begin{aligned} s^{-1}\left[\mathbf{G}(s) - y_0\right] &= a\mathbf{G}(s) - \frac{b}{1-s} \\ \left[\mathbf{G}(s) - y_0\right] &= as\mathbf{G}(s) - \frac{bs}{1-s} \\ \mathbf{G}(s)\left(1-as\right) &= y_0 - \frac{bs}{1-s} \\ \mathbf{G}(s) &= \frac{y_0}{1-as} - \frac{bs}{\left(1-s\right)\left(1-as\right)} \end{aligned}$$

And, using our tools of inversion, we can write at once

$$\begin{aligned} \mathbf{G}(s) &= \frac{y_0}{1-as} - \frac{bs}{\left(1-s\right)\left(1-as\right)} \\ \mathbf{G}(s) &\rhd \left\{ y_0\, a^k - b\sum_{j=0}^{k-1} a^j \right\} \\ \mathbf{G}(s) &\rhd \left\{ y_0\, a^k - \frac{b}{1-a}\left(1-a^k\right) \right\} \end{aligned}$$

and the solution is $y_k = y_0 a^k - \frac{b}{1-a}\left(1-a^k\right).$

### *Derivatives of generating functions*

One of the most useful features is the ability to take derivatives (with respect to *s*) term by term of a generating function, and have it be the equivalent to taking the derivative of the generating function, i.e.,

$$\sum_{k=0}^{\infty}\left[a_k \frac{ds^k}{ds}\right]=\frac{d\,\mathbf{G}(s)}{ds}.$$

This is not an ability that we can take for granted. It makes no sense to attempt this for divergent series, and in fact, in general is not true for pointwise convergent series. However, we can carry this crucial operation out for series that are uniformly convergent.

Since *s* is a variable that we can choose to be as small as we want, and can bound it from above. By doing this, we know that $\sum_{k=0}^{\infty} a_k s^k$ will be uniformly convergent, permitting us to take its derivative term by term and have the resulting expression be $\frac{d\,\mathbf{G}(s)}{ds}$.

We can begin with the simple geometric series to show the value of taking derivatives. This we have the geometric series, we know that $\mathbf{G}(s)=\frac{1}{1-s}$. The $k^{\text{th}}$ term of the series is simply $s^k$, and its derivative $\frac{ds^k}{ds}=ks^{k-1}$. if we call $\mathbf{G}_1(s)=\frac{d\,\mathbf{G}(s)}{ds}=\frac{1}{(1-s)^2}$, then $\mathbf{G}_1(s)\triangleright\{k+1\}$.[*]

We can go one step further, recognizing that

$$\frac{1}{(1-s)^2}=1+2s+3s^2+4s^3+\ldots+(k+1)s^k+\ldots$$

Multiplying each side by *s* we get

$$\frac{s}{(1-s)^2}=1+2s^2+3s^3+4s^4+\ldots+ks^k+\ldots$$

and we see that $\frac{s}{(1-s)^2}\triangleright\{k\}$.

Yet one more derivative generates

$$\frac{2s}{(1-s)^3}+\frac{1}{(1-s)^2}=\frac{s+1}{(1-s)^3}=1+\ldots+k^2s^{k-1}+(k+1)^2s^k+\ldots$$
$$\frac{s(s+1)}{(1-s)^3}=1+\ldots+k^2s^k+\ldots$$

[*] You can multiply the two identical generating functions $\left(\frac{1}{1-s}\right)\left(\frac{1}{1-s}\right)$ to obtain the same result.

and $\mathbf{G}(s) = \dfrac{s(s+1)}{(1-s)^3} \rhd \{k^2 s^k\}$.

We can continue this development, differentiating each side of the preceding equation to find

$$\frac{2}{(1-s)^3} = 1 + 2s + (3)(2)s^1 + (4)(3)s^2 + \ldots$$
$$+ (k+1)ks^{k-1} + (k+2)(k+1)s^k + \ldots$$

and once more to see

$$\frac{(3)(2)}{(1-s)^4} = + \ldots + (k+3)(k+2)(k+1)s^k + \ldots$$

Recognizing that $(k+3)(k+2)(k+1) = \dfrac{(k+3)!}{k!}$, we can write

$$\frac{(3)(2)}{(1-s)^4} \rhd \frac{(k+3)!}{k!} \text{ or } \frac{1}{(1-s)^4} \rhd \left\{\frac{(k+3)!}{3!k!}\right\} = \binom{k+3}{3}.$$

A simple induction argument produces

$$\frac{1}{(1-s)^r} \rhd \left\{\binom{k+r-1}{r-1}\right\}.$$

This will be central to our development of the negative binomial measure.

## Collecting coefficients

Another useful tool for generating function is the process of collecting coefficients. For example, let's say we have the following generating function.

$$\mathbf{G}(s) = \frac{1}{1-s-s^2}.$$

We could rewrite this as

$$\mathbf{G}(s) = \frac{1}{1-s(1+s)} = \frac{1}{1-sa_s}$$

Permitting us to write the sequence as
$1 + a_s s + a_s^2 s^2 + a_s^3 s^3 + \ldots + a_s^k s^k$.

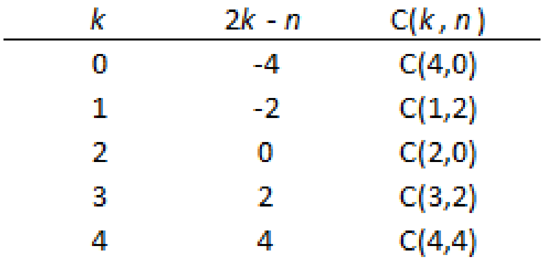
However, the problem with just writing $\mathbf{G}(s) \rhd \{a_k\}$ is that $a_k$ is a function of $s$. For example $a_s^3 = (1+s)^3 = 1 + 3s + 3s^2 + s^3$, so there are additional coefficients of $s^k$ to collect.

For this sequence we choose to write $\mathbf{G}(s) \rhd_s \left\{(1+s)^k\right\}$, the subscript on the symbol $\rhd_s$ signifying that the inversion is incomplete, the coefficient of $s^k$ at this point remains a function of $s$.

We now write the $k$+1st term of the series as $(1+s)^k s^k$, expanding this to $\sum_{j=0}^{k}\binom{k}{j}s^{k+j}$ through the use of the binomial theorem.

How does this function accumulate coefficients? For $k$ = 0, $j$ = 0 and the coefficient of $s^0$ is simply $\binom{0}{0}=1$. For $k$ = 1, and $j$ = 0 we have the coefficient of $s^{k+j}=s^1$ as $\binom{1}{0}$. $k$ = 1 and $j$ = 1 generates a coefficient of $s^{1+1}=s^2$ as $\binom{1}{1}$. We continue in this fashion, first increasing $k$, then allowing $j$ to move from 0 to $k$, accumulating all of the coefficients of $s^{k+j}$ (Table 1).

Table 1. Example of Collecting Coeficients for $n$ =4



| $k$ | $2k$ - $n$ | C($k$, $n$) |
|---|---|---|
| 0 | -4 | C(4,0) |
| 1 | -2 | C(1,2) |
| 2 | 0 | C(2,0) |
| 3 | 2 | C(3,2) |
| 4 | 4 | C(4,4) |

From Table 1, the third column, we see that there is only one way to generate $s^0$ and $s^1$, and we already have those coefficients. For the term $s^2$, we observe that there are two coefficients, and we sum them, $\binom{1}{1}+\binom{2}{0}$.

For $s^3$, we create $\binom{2}{1}+\binom{3}{0}$. For $s^4$, we compute $\binom{2}{2}+\binom{3}{1}+\binom{4}{0}$. The process is wholly mechanical, and we can summarize it by creating a new variable $m$ and writing the coefficient of $s^k$ as $a_k$ where

$$a_k = \sum_{m=0}^{k}\sum_{j=0}^{m}\binom{m}{j}1_{m+j=k}.$$

and we have reached our goal of having the coefficient be free of any terms involving $s$.

For example, if we have the more general generating function

$$\mathbf{G}(s) = \frac{1}{c-bs-as^2},$$

We proceed as follows.

$$\mathbf{G}(s)=\frac{1}{c-bs-as^2}=\frac{\frac{1}{c}}{1-\frac{a}{c}s\left(s+\frac{b}{a}\right)}\rhd_s\left\{\frac{1}{c}\left(\frac{a}{c}\right)^k\right\}\left(s+\frac{b}{a}\right)^k s^k$$

To collect coefficients from $\left(s+\frac{b}{a}\right)^k s^k=\sum_{j=0}^{k}\binom{k}{j}s^{k+j}\left(\frac{b}{a}\right)^{k-j}$. Following the previous example, the collection process yields $\sum_{m=0}^{k}\sum_{j}^{m}\binom{m}{j}\left(\frac{b}{a}\right)^{m-j}1_{m+j=k}$ and

$$\mathbf{G}(s)=\frac{1}{c-bs-as^2}=\rhd\left\{\frac{1}{c}\left(\frac{a}{c}\right)^k\sum_{m=0}^{k}\sum_{j}^{m}\binom{m}{j}\left(\frac{b}{a}\right)^{m-j}1_{m+j=k}\right\}$$

Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Skewness and Kurtosis for Bernoulli Measure

Prerequisites
Properties of Probability
Bernoulli Distribution – In Depth Discussion

We can also compute other quantities that are of interest based on the different expectations or moments of the Bernoulli random variable *X*. We can write skewness as

$$S(X)=\frac{\mathbf{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

and kurtosis as

$$K(X)=\frac{\mathbf{E}\left[(X-\mu)^4\right]}{\sigma^4}.$$

## Skewness for Bernoulli distribution

Some algebra reveals

$$\begin{aligned}\sigma^{-3}S(X)&=\mathbf{E}\left[(X-\mu)^3\right]=\mathbf{E}\left[X^3-3X^2\mu+3X\mu^2-\mu^3\right]\\&=\mathbf{E}\left[X^3\right]-3\mu\mathbf{E}\left[X^2\right]+3\mu^2\mathbf{E}\left[X\right]-\mu^3.\end{aligned}$$

Our work is eased by noting that for any positive integer k.

$$\mathbf{E}\left[X^k\right]=\int_{\Omega_x} x^k\,\mathrm{d}\mathbf{P}=0^k\,\mathbf{P}\left[X=0\right]+1^k\,\mathbf{P}\left[X=1\right]=0q+1p=p$$

Thus

$$
\begin{aligned}
\sigma^{-3}\mathbf{S}(X) &= \mathbf{E}\left[\left(X-\mu\right)^3\right]=\mathbf{E}\left[X^3-3X^2\mu+3X\mu^2-\mu^3\right]\\
&= \mathbf{E}\left[X^3\right]-3\mu\mathbf{E}\left[X^2\right]+3\mu^2\mathbf{E}\left[X\right]-\mu^3\\
&= p-3p^2+3p^3-p^3=p-3p^2+2p^3\\
&= p-p^2-2p^2+2p^3=p(1-p)-2p^2(1-p) \qquad\qquad \text{Thus}\\
&= p\left(1-p\right)-2p^2\left(1-p\right)=\left(1-p\right)\left(p-2p^2\right)=p\left(1-p\right)\left(1-2p\right)\\
&= p\left(1-p\right)-2p^2\left(1-p\right)=\left(1-p\right)\left(p-2p^2\right)=p\left(1-p\right)\left(1-2p\right)\\
&= p\left(1-p\right)\left(1-2p\right)=p\left(1-p\right)\left(p+q-2p\right)=pq\left(q-p\right).
\end{aligned}
$$

$$
\begin{aligned}
\sigma^{-3}S(X) &= pq\left(q-p\right)\\
S(X) &= \frac{pq\left(q-p\right)}{\sigma^3}=\frac{pq\left(q-p\right)}{pq\sqrt{pq}}=\frac{\left(q-p\right)}{\sqrt{pq}}.
\end{aligned}
$$

## Kurtosis for Bernoulli distribution

Following the development for skewness, we write

$$
\begin{aligned}
\sigma^{-4}\mathbf{K}(X) &= \mathbf{E}\left[X-\mu\right]^4=\mathbf{E}\left[\left(X-\mu\right)^2\left(X-\mu\right)^2\right]\\
&= \mathbf{E}\left[\left(X^2-2X\mu+\mu^2\right)\left(X^2-2X\mu+\mu^2\right)\right]\\
&= \mathbf{E}\left[X^4-2X^3\mu+X^2\mu^2-2X^3\mu+4X^2\mu^2\right]\\
&+\mathbf{E}\left[-2X\mu^3+X^2\mu^2-2X\mu^3+\mu^4\right]\\
&= \mathbf{E}\left[X^4-4X^3\mu+6X^2\mu^2-4X\mu^3+\mu^4\right]
\end{aligned}
$$

Recalling that $\mathbf{E}\left[X^k\right]=p,$ we write

$$
\begin{aligned}
\sigma^{-4}\mathbf{K}(X) &= \mathbf{E}\left[X^4-4X^3\mu+6X^2\mu^2-4X\mu^3+\mu^4\right]\\
&= p-4p^2+6p^3-4p^4+p^4
\end{aligned}
$$

Which simplifies to

$$
\begin{aligned}
\sigma^{-4}\mathbf{K}\left(X\right) &= p-4p^2+6p^3-4p^4+p^4\\
&= p-4p^2+6p^3-3p^4\\
&= p-p^2-3p^2+3p^3-3p^3-3p^4\\
&= p\left(1-p\right)\left(1-3p+3p^2\right)\\
&= p\left(1-p\right)\left(1-3p\left(1-p\right)\right)\\
&= pq\left(1-3pq\right).
\end{aligned}
$$

Since $\sigma^2=pq$ for the Bernoulli distribution, we write $\mathbf{K}\left(X\right)=\dfrac{pq(1-3pq)}{\left(pq\right)^2}=\dfrac{(1-3pq)}{pq}.$

Finally, workers typically subtract 3 from **K**(*X*) since this is the average kurtosis, $\mathbf{K}_a(X)$. We finish by writing

**Kurtosis for Bernoulli distribution-**

$$\mathbf{K}_a\left(X\right)=\frac{(1-3pq)}{pq}-3=\frac{(1-3pq)}{pq}-\frac{3pq}{pq}=\frac{1-6pq}{pq}$$

Which is how the kurtosis for the Bernoulli distribution is commonly reported.

Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Basics of the Binomial Distribution

The binomial distribution is a simple complication of the Bernoulli distribution. Binomial events are merely the sum of Bernoulli events.

## Prerequisite

Properties of Probability
Counting Events
Basics of Bernoulli Trials.

## Building the Binomial from the Bernoulli

Recall that we ended the section on the Bernoulli distribution with the expectation and variance of the sum of two independent Bernoulli variables. Let's now assume that two random variables $X_1$ and $X_2$ are independent and each come from the same Bernoulli($p$) distribution. We call such variables i.i.d. for *independent and identically distributed*.

In general, if we have $n$ i.i.d. Bernoulli trials, and $W$ is their sum (often times expressed as the number of successes in $n$ i.i.d Bernoulli trials, then

$$\mathbf{P}\left[W_n = k\right] = \binom{n}{k} p^k q^{n-k} \mathbf{1}_{k=\mathrm{I}_{0,n}}.$$

This is the binomial distribution probability function. Unlike Bernoulli random variables which take on only the values 0 and 1, random variables that follow the binomial distribution can take on any integer value between and including 0 and $n$.

The formula itself is made up of two components. The portion $p^k q^{n-k}$ is simply the probability of a sequence of $n$ Bernoulli trials in which any $k$ of them are successes. However, there are multiple ways to obtain $k$ successes in $n$ Bernoulli trials, each of them independent of the other. The exact number of ways to do this is $\binom{n}{k}$ which comes from our section on permutations and combinations.

## Expectation and variance

One way to easily find the mean and variance of the binomial distribution is from our previous discussion of the Bernoulli distribution. Recall that a binomial ($n$, $p$) random variable $W$ is the sum of $n$ i.i.d. Bernoulli($p$) random variables, $X_1$, $X_2$, $X_3$, ..., $X_n$. Then, if $W = \sum_{i=1}^{n} X_i$, where $\mathbf{E}\left[X_i\right] = p$ and $\mathbf{Var}\left[X_i\right] = pq$, then

$$\mathbf{E}\left[W_i\right] = \sum_{i=1}^{n} \mathbf{E}\left[X_i\right] = np$$

$$\mathbf{Var}\left[W_i\right] = \sum_{i=1}^{n} \mathbf{Var}\left[X_i\right] = npq$$

Similarly, we can see that for $W_1 + W_2$ where $W_1$ binomial ($n_1, p$ ) is independent of $W_2$ binomial ($n_2, p$). we know that $W_1 + W_2$ is the sum of $n_1 + n_2$ Bernoulli($p$) random variables which of course is binomial($n_1 + n_2$, $p$).

### *Example: Hurricanes*

Assume that in any given year, the probability that a tropical storm will become a hurricane is 0.28. If there are fifteen tropical storms in a given year, what can we say about the occurrence of hurricanes that year?

If we assume that tropical storms occur independently of one another, we can treat the generation of a hurricane as a Bernoulli trial. We can then use the binomial distribution to compute the probability of any particular number of hurricanes. If we let $k$ be the number of hurricanes, then we know that $0 \le k \le 15$, and

$$\mathbf{P}\left[X = k\right] = \binom{15}{k}(0.28)^k (0.72)^{n-k}.$$

Thus, the probability of 3 hurricanes is $\binom{15}{3}(0.28)^3 (0.72)^{12} = 0.194.$ If we wanted to compute the probability of at least three hurricanes, we could compute $\sum_{k=3}^{15}\binom{15}{k}(0.28)^k (0.72)^{n-k} = 0.835,$ although it can be easier to compute

$$\begin{aligned}\sum_{k=3}^{15}\binom{15}{k}(0.28)^k (0.72)^{n-k} &= 1 - \sum_{k=0}^{2}\binom{15}{k}(0.28)^k (0.72)^{n-k} \\ &= 1 - 0.165 = 0.835.\end{aligned}$$

The mean number of hurricanes is simply $(15)(0.28) = 4.2.$ The standard deviation is $\sqrt{(15)(0.28)(0.72)} = 1.74.$

Now, suppose the occurrence of hurricanes in one year is independent of hurricane occurrences in any other year. In a ten year span, what is the probability that in at least 5 years, there were be at least three hurricanes given that there are fifteen storms per year.

We know the probability of at least three hurricanes in a given year is 0.835. This is now the probability of a success for a new Bernoulli trial, i.e., the occurrence of at least three hurricanes in a given year. We now compute $\sum_{k=5}^{10}\binom{10}{k}(0.835)^k (0.165)^{n-k} = 0.998.$ It is a virtual certainty that there will be at least five years (not necessarily consecutive) each of which will have at least 3 hurricanes.

Elementary Track
Basics of the Poisson Distribution
Basics of Normal Measure

# Binomial Measure

The binomial distribution is one of the most ubiquitous distributions in applied probability. Based on our experience with the simple Bernoulli model, we will build up the simple 0-1 random variable to a more intricate reflection of more complex combinations of events and measure them.

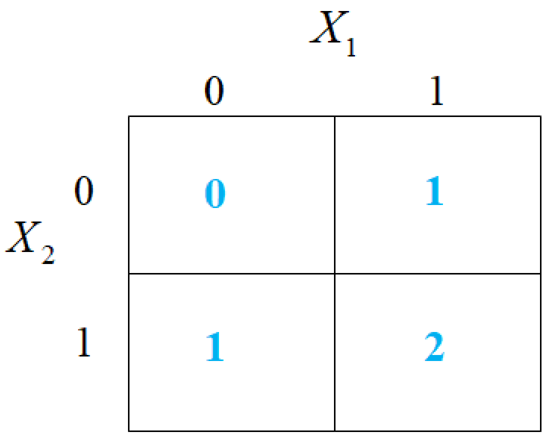
## Prerequisites

Why Probability
Factorials Permutations, and Combinations
Binomial Theorem
An Introduction to the Concept of Measure
Elementary Set Theory
Sequences of Sets
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Properties of Probability
Conditional Probability
Counting Events
Bernoulli Measure

## Building the binomial from the Bernoulli

It is commonly easier to understand the underlying event to which a measure is applied than to simply memorize a formula. Remember that we ended the discussion of the Bernoulli distribution with the sum of two Bernoulli random variables to identify the mean and variance of this sum. Let's now assume that two random variables $X_1$ and $X_2$ are independent and each come from the same Bernoulli($p$) measure. We call such variables i.i.d. for *independent and identically distributed*. Define $W_2 = X_1 + X_2$.

What experiment does $W$ represent? Its possible values are 0, 1, or 2, since both $X_1$ and $X_2$ can only take on values 0 or 1. (Figure 1).



| $X_2$ \ $X_1$ | 0 | 1 |
|---|---|---|
| 0 | 0 | 1 |
| 1 | 1 | 2 |

Figure 1. The sum of two Bernoulli random variables.

We see that $\mathbf{P}\left[W_2=2\right]=p^2$ since the only way for the value two to occur is when each of $X_1$ or $X_2$ is equal to one. Similarly, $\mathbf{P}\left[W_2=0\right]=q^2$.

However, $\mathbf{P}\left[W_2=1\right]$ is a little more complicated. $W$ takes on the value 1 when either $X_1$ or $X_2$ is equal to one, but not both. The measure of the event that exactly one of them being one is simply $pq$. However there is more than one way to have a value of one since "1" could appear first or appear second in the $X_1X_2$ sequence. There are two possible sequences and each sequence occurs with measure $pq$. Since these sequences are mutually exclusive, we write

$$\mathbf{P}\left[W_2=1\right]=2pq.$$

This is essentially a two-step process. The first step requires us to compute the measure of the sequence of zeros and ones (or successes and failures). The second step is to count the number of ways this sequence occurs, multiplying by that final number.

Assume now that we have five i.i.d. Bernoulli ($p$) random variables. We want to compute the probability that the sum of them is equal to three, i.e., $\mathbf{P}\left[W_5=3\right]$.

The measure of the event that there are three ones (or three successes) in a sequence of trials is simply $p^3q^2$. Now, how many ways are there to produce these sequences? We could simply count

SSSFF, SSFFS, SSFSF, SFFSS, SFSSF,<br>
SFSFS, FFSSS, FSFSS, FSSFS, FSSSF

A simple way to count these is to think of the process first as permuting the three successes through the five trials. However, since we are not ordering successes (i.e., $S_1S_2S_3FF$ is the same to us as $S_2S_3S_1FF$) we must move to a combination to remove the duplicates. Thus the number of ways to count these sequences is to ensure that the order of the successes and failures does not matter, as long as we have three successes and two failure. This is $\binom{5}{3}=10$. Thus, our solution is

$$\mathbf{P}[W_5 = 3] = \binom{5}{3} p^3 q^2.$$

Note, one of the assumptions that allowed this computation is that the Bernoulli random variables have the same parameter *p*, i.e., wherever the success occurs, the probability of that success is always *p* regardless of the trial number.

In general, if we have *n* i.i.d. Bernoulli trials, and $W_n$ is their sum (often times expressed as the number of successes in *n* i.i.d. Bernoulli trials, then the measure of the event $W_n = k$

$$\mathbf{P}[W_n = k] = \binom{n}{k} p^k q^{n-k} \mathbf{1}_{[k=\mathrm{I}_{0,n}]}.$$

This is binomial measure. It is our measuring tool to compute the probability of binomial random variables. The indicator $\mathbf{1}_{[k=\mathrm{I}_{0,n}]}$ indicates the function if zero if *k* is anything but an integer in [0, *n*]. This mechanism permits us to use the measure theory integral for all of our upcoming work, following this development for the Bernoulli distribution.

In order to show this is truly a probability measure, we must demonstrate that the probability or measure over the entire space $\int_\Omega d\mathbf{P} = 1$. Begin by recognizing that $\int_\Omega d\mathbf{P} = \int_\Omega \binom{n}{k} p^k q^{n-k} \mathbf{1}_{[k=\mathrm{I}_{0,n}]}$. Here $\int_\Omega \binom{n}{k} p^k q^{n-k} \mathbf{1}_{k=\mathrm{I}_{0,n}} = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$.

The symbology $\int_\Omega \binom{n}{k} p^k q^{n-k} \mathbf{1}_{k=\mathrm{I}_{0,n}}$ may look strange at first. However, when you recall that the symbol $\int_\Omega$ is simply our announcement that we will be taking the measure of the set $\Omega$ where $\binom{n}{k} p^k q^{n-k} \mathbf{1}_{k=\mathrm{I}_{0,n}}$ is the measuring tool, the formulation $\int_\Omega \binom{n}{k} p^k q^{n-k} \mathbf{1}_{k=\mathrm{I}_{0,n}}$ takes on a rather obvious meaning and we can proceed with the computation.

One of the easiest proofs that this sum is one requires the invocation of the binomial theorem, from Blaise Pascal. The binomial theorem states that for any constants *a* and *b*, and for nonnegative integer *n*, then

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

An equality easily proved using induction. We simply need to set *a* = *p* and *b* = *q*. Applying this function to the binomial measure reveals

$$(p+q)^n = 1^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

## Computing using binomial measure

Unlike Bernoulli random variables which take on only the values 0 and 1, random variables that follow the binomial distribution can take on any integer value between 0 and $n$, allowing for a very rich σ-algebra.

### *Example: satellite clinics*

As an example, assume a clinical trial has a center with twelve clinical satellites recruiting for it. The probability that a clinic reaches its recruitment goal of patients is $p = 0.75$. Let $W$ be the number of clinics that reaches their quota.

Then assuming the clinical centers recruit independently of each other, it is reasonable to operate under the assumption that $W$ follows binomial measure with parameters $n = 12$ and $p = 0.75$. Let's call this random variable $W_{12}(0.75)$. We can compute the probability that exactly nine of the satellites successfully reach their respective goals as

$$\mathbf{P}\left[W_{12}(0.75)=9\right]=\binom{12}{9}(0.75)^9(0.25)^3=0.258.$$

This may seem surprisingly low since $p= 0.75$. What is happening here?

Examining the measuring tool for W provides some illumination (Figure 2)

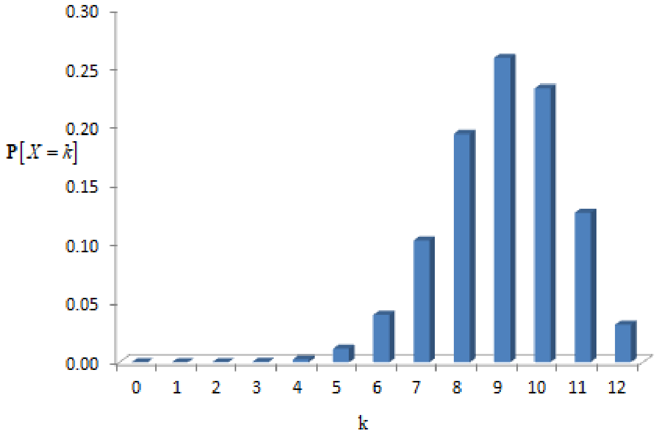


**Figure 2.** Probability function of the Binomial (12, 0.75) distribution.

We see that most of the measure is concentrated for values of $W \geq 6$. We can demonstrate this by computing

$$\mathbf{P}\left[W_{0.75}\geq 6\right]=\int_{W\geq 6} d\mathbf{P}=\sum_{k=6}^{12}\binom{12}{k}(0.75)^k(0.25)^{12-k}=0.986.$$

If the goals were set more ambitiously such that $p = 0.25$, then $W_{12}(0.25)$ would follow a binomial (12, 0.25) with a very different appearance (Figure 3).

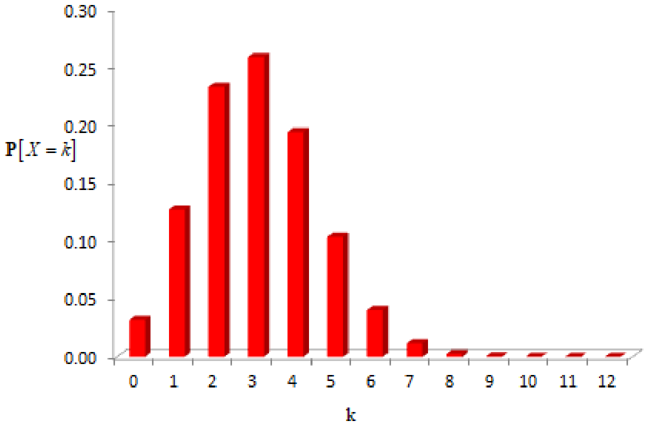


**Figure 3.** Probability function of the Binomial (12, 0.25) distribution.

We can also compute the cumulative probability distribution function for each of these measures, anticipating that the probabilities will be higher for smaller values of the random variable when taken from the binomial measuring rule with the smallest probability $p$ (Figure 4).

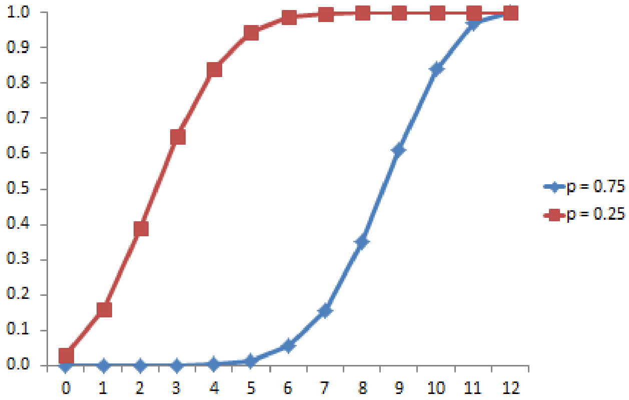


Figure 4. Comparison of the cumulative measure of two binomial random variables.

Other probabilities based on the binomial distribution are easily available. For example, we can compute the probability that between seven and ten clinics meet their goals. This is

$$\mathbf{P}\left[7 \le W_{12}(0.75) \le 10\right] = \int_{7 \le W \le 10} d\mathbf{P} = \sum_{k=7}^{10} \binom{12}{k} (0.75)^k (0.25)^{12-k}$$

We can relieve some of the computational burden by seeing that

$$\mathbf{P}\left[7 \le W_{12}(0.75) \le 10\right]$$
$$= \mathbf{P}\left[W_{12}(0.75) \le 10\right] - \mathbf{P}\left[W_{12}(0.75) < 7\right] = 0.842 - 0.054 = 0.788.$$

Note the strict inequality in $\mathbf{P}\left[W_{12}(0.75)<7\right]$ to ensure that the interval contains $\mathbf{P}\left[W_{12}(0.75)=7\right]$. The measure of the interval is equal to the difference of two probabilities taken from the CDF, a useful tool since the CDF is easy to compute*.

Similarly, we can compute the measure that no more than nine satellite clinics reach their goal as simply

$$\mathbf{P}\left[W_{12}(0.75)\leq 9\right]=\int_{W\leq 9} d\mathbf{P} =\sum_{k=0}^{9}\binom{12}{k}(0.75)^{k}(0.25)^{12-k}=0.609.$$

These computations demonstrate the ease of calculation with the binomial distribution.

## Expectation and variance

While it may be somewhat difficult to deduce the relationship between the variance of $W_n(0.25)$ and $W_n(0.75)$ at this point, we might expect that the $\mathbf{E}\left[W_n(0.25)\right]<\mathbf{E}\left[W_n(0.75)\right]$. One way to easily find the mean and variance of the binomial distribution is from our discussion of mean and variance of the sum of independent random variables. Recall that a binomial $(n, p)$ random variable $W_n(p)$ is the sum of $n$ i.i.d. Bernoulli($p$) random variables, $X_1, X_2, X_3, \ldots, X_n$. Then if

$W_n(p)=\sum_{i=1}^{n}X_i$, where $\mathbf{E}\left[X_i\right]=p$ and $\mathbf{Var}\left[X_i\right]=pq$, then

$$\mathbf{E}\left[W_n(p)\right]=\sum_{i=1}^{n}\mathbf{E}\left[X_i\right]=np$$
$$\mathbf{Var}\left[W_n(p)\right]=\sum_{i=1}^{n}\mathbf{Var}\left[X_i\right]=npq$$

and we can see that $\mathbf{E}\left[W_n(0.25)\right]=(12)(0.25)=3 < \mathbf{E}\left[W_n(0.75)\right]=(12)(0.75)=9$. As it turns out,

$$\mathbf{Var}\left[W_n(0.25)\right]=(12)(0.25)(0.75)=2.25$$
$$\mathbf{Var}\left[W_n(0.75)\right]=(12)(0.75)(0.25)=2.25.$$

Another way to compute the mean and variance of $W_n(p)$ is to carry out a direct calculation with no reliance on Bernoulli trials. We start with

$$\mathbf{E}\left[W\right]=\int_{\Omega_W} wd\mathbf{P}=\sum_{k=0}^{n}k\mathbf{P}\left[W=k\right]=\sum_{k=0}^{n}k\binom{n}{k}p^{k}q^{n-k}=\sum_{k=1}^{n}k\binom{n}{k}p^{k}q^{n-k}$$

We write $k\binom{n}{k}=\frac{kn!}{k!(n-k)!}=\frac{n!}{(k-1)!(n-k)!}=n\frac{(n-1)!}{(k-1)!(n-k)!}=n\binom{n-1}{k-1}$.

Thus

[*] For example, Excel and apps for portable devices provides this.

$$\mathbf{E}\left[W_n(p)\right]=\sum_{k=1}^{n}k\binom{n}{k}p^kq^{n-k}=\sum_{k=1}^{n}n\binom{n-1}{k-1}p^kq^{n-k}$$
$$=np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{n-k}$$

Now just let $j=k-1, m=n-1,$ and recognize that the sum on the right hand side of the previous expression is the sum of binomial ($n-1, p$) random variable from 0 to $n-1$. Thus

$$\mathbf{E}\left[W_n(p)\right]=np\sum_{j=0}^{m}\binom{m}{j}p^jq^{m-j}=np.$$

In order to find the $\mathbf{Var}\left[W_n(p)\right]=\mathbf{E}\left[W_n^2(p)\right]-\mathbf{E}^2\left[W_n(p)\right]$ directly, we will proceed analogously. Let's simplify notation and simply let $W=W_n(p)$. We will also focus on the factorial moment, $\mathbf{E}\left[W(W-1)\right]$. Write

$\mathbf{E}\left[W(W-1)\right]=\mathbf{E}\left[W^2\right]-\mathbf{E}[W]$, or reconfigure as $\mathbf{E}\left[W^2\right]=\mathbf{E}\left[W(W-1)\right]+\mathbf{E}[W]$.

We can write now write

$$\mathbf{Var}[W]=\mathbf{E}\left[W^2\right]-\mathbf{E}^2[W]$$
$$=\mathbf{E}\left[W(W-1)\right]+\mathbf{E}[W]-\mathbf{E}^2[W]$$

Since we know the $\mathbf{E}[W]=np,$ we can write

$$\mathbf{Var}[W]=\mathbf{E}\left[W(W-1)\right]+np-n^2p^2$$

We proceed as we did for the direct calculation of the expectation.

$$\mathbf{E}\left[W(W-1)\right]=\int_{\Omega_W}w(w-1)d\mathbf{P}=\sum_{k=0}^{n}k(k-1)\mathbf{P}[W=k]$$
$$=\sum_{k=0}^{n}k(k-1)\binom{n}{k}p^kq^{n-k}\ .$$

Continuing

$$\sum_{k=0}^{n}k(k-1)\binom{n}{k}p^kq^{n-k}$$
$$=\sum_{k=2}^{n}k(k-1)\binom{n}{k}p^kq^{n-k}=\sum_{k=2}^{n}n(n-1)\binom{n-2}{k-2}p^kq^{n-k}$$
$$=n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}q^{n-k}=n(n-1)p^2\sum_{j=0}^{m}\binom{m}{j}p^jq^{m-j}$$
$$=n(n-1)p^2.$$

Thus

$$\begin{aligned}\mathbf{Var}[W] &= \mathbf{E}\left[W(W-1)\right] + np - n^2p^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np - np^2 \\ &= np(1-p).\end{aligned}$$

The inclusion of the factorial in the binomial probability function induced us to first find a factorial moment.

## Skewness

We can also compute the skewness and kurtosis for binomial random variables. Let $W$ follow a binomial($n$, $p$) distribution. Then $\mathbf{S}(X) = \frac{\mathbf{E}\left[(W-\mu)^3\right]}{\sigma^3} = \frac{(1-2p)}{\sqrt{npq}}$ Note that when $p = 0.50$, the skewness is zero, indicating a symmetric distribution. (Figure 5)

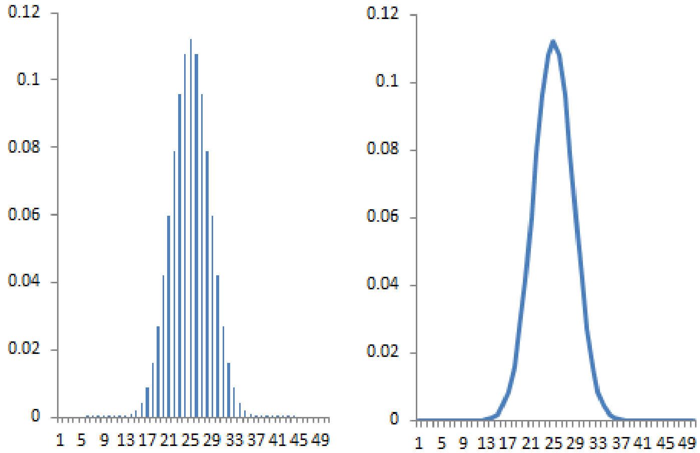


**Figure 5**. Example of a symmetric binomial distribution ($p = 0.50$). In this case $n = 50$. Note how when we have enough bars, their individual heights can be approximated by a smooth line.

We can show that kurtosis for a binomial ($n$, $p$) random variable is $\mathbf{K}(x) = \frac{1-6p(1-p)}{np(1-p)}$.

Also note in Figure 5 how when we have enough bars, we can begin to approximate their individual heights by a smooth line. Of course, this is an inexact process, since each bar has a probability associated with it, while a point on a curve that is not an integer has probability according to binomial measure.

However, we can imagine that if we go far enough out in our mind's eye, $n$ in the hundreds, then the thousands, then the tens of thousands, the bar widths get smaller and smaller.

There comes a point where the concept of probability as the height of a bar breaks down and the concept of probability as the area under the curve becomes attractive. In this case the classic Riemann integral can substitute for binomial measure to determine probability.

Traditionally, this represents a major transition. However, when we use the Lebesgue perspective, we see that all we have done is change the tool that we used to accumulate probability, approximating binomial measure with a Riemann integrable function.

## Binomial generating functions

$\mathbf{M}_W(t)$ and $\mathbf{G}_W(s)$ are easily computed for the binomial distribution. For the moment generating function we find that,

$$\mathbf{M}_W(t) = \mathbf{E}\left[e^{tw}\right] = \int_{\Omega_W} e^{tw} d\mathbf{P} = \sum_{k=0}^{n} e^{tk}\binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n}\binom{n}{k}\left(e^t p\right)^k q^{n-k}$$
$$= \left(q + pe^t\right)^n.$$

The last equality represents a use of the binomial theorem.
The probability generating function calculation proceeds analogously.

$$\mathbf{G}_W(s) = \mathbf{E}\left[s^W\right] = \int_{\Omega_W} s^w d\mathbf{P} = \sum_{k=0}^{n} s^k\binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n}\binom{n}{k}(ps)^k q^{n-k}$$
$$= (q + ps)^n.$$

Note, however, that these tools permit another generation of the binomial distribution from the Bernoulli.

Remember that we first generate the binomial random variable as the sum of i.i.d. Bernoulli trials, i.e., $W = \sum_{i=1}^{n} X_i$. We can now note that

$$\mathbf{G}_W(s) = \mathbf{E}\left[s^W\right] = \mathbf{E}\left[s^{X_1+X_2+X_3+\ldots+X_n}\right] = \prod_{i=1}^{n}\mathbf{E}\left[s^{X_i}\right]$$

However, since the $X_1$, $X_2$, $X_3$, …, $X_n$ are i.i.d., $\mathbf{E}\left[s^{X_i}\right] = q + ps$ for $i = 1, 2, 3, \ldots, n$. Therefore

$$\mathbf{G}_W(s) = \prod_{i=1}^{n}\mathbf{E}\left[s^{X_i}\right] = (q + ps)^n.$$

This is a valuable way to identify what otherwise can be a complicated probability distribution analysis of the sums of random variables. If one random variable's generating function is the product of the generating functions of several random variables, the first random variable is the sum of the random variables whose generating functions comprise the product.

### *Pulling the binomial distribution "out of a hat"*

As an aside, we can commonly identify an inverted generation function as related to the negative binomial distribution.

Lets start with a generating function, $\mathbf{G}_t(s) = (as + b)^n$. Then rewrite as

$\mathbf{G}_t(s) = (as + b)^n = (a + b)^n\left(\frac{a}{a+b}s + \frac{b}{a+b}\right)^n$, which is

$\mathbf{G}_t(s) = (a+b)^n (ps+q)^n$ . Inverting this provides the coefficient of X to the K as being a binomial probability multiplied by the constant $(a+b)^n$ .

## Random walk

Random walks are among one of the simplest of stochastic processes, which themselves are changes in random variables over time. Encryption keys, signals through circuit board, radar detection, and advanced stealth technology, are each examples of the application of complicated stochastic processes. However, the simplest process is that of the random walk.

### *Simple random walk*

We are used to binomial random variables as being the sum of Bernoulli random variables whose hallmark is that they take on the value of either one or zero with fixed probabilities. Consider a random variable *X* that takes on the value of 1 with probability *p* and -1 with probability *q* such that $p+q=1$.[*] Then what does $W_n = \sum_{i=1}^{n} X_i$ look like, assuming that the $X_i$'s are each i.i.d?

Before we try to find the exact measuring tool of $W_n$, we can think of its general properties. For each *i*, $W_n$ either increases or decreases one unit (Figure 6.)

While it is possible that it could always increase, or always decrease it is likely to meander, increasing for a time then decreasing for a time. The likelihood that it increases depends on the value of *p.*

We will use the probability generating function to discover the actual probability distribution for $W_n$. But first, let's see if we can identify some of its characteristics. We begin by noting that $\mathbf{P}[X_i = 1] = p$ and

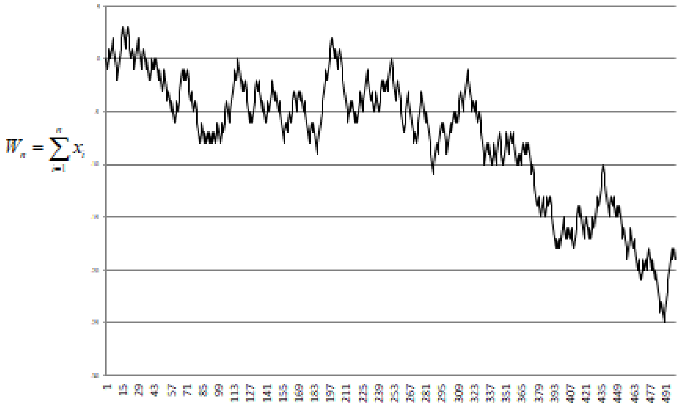


Figure 6. An example of a random walk, $p = 0.50$

[*] Note that this is not a Bernoulli trial, however it is closely related to one.

$\mathbf{P}\left[X_i = -1\right] = 1 - p = q.$ We can then take the next step to compute $\mathbf{P}\left[X = k\right] = p\mathbf{1}_{k=1} + q\mathbf{1}_{k=-1}.$ We can easily satisfy ourselves that $\int\limits_{\Omega} d\mathbf{P} = \int\limits_{-\infty}^{\infty} p\mathbf{1}_{k=1} + q\mathbf{1}_{k=-1} = p + q = 1.$ We can also compute

$$\mathbf{E}\left[X\right] = \int\limits_{\Omega} x d\mathbf{P} = \int\limits_{-\infty}^{\infty} k\left(p\mathbf{1}_{k=1} + q\mathbf{1}_{k=-1}\right) = p\left(1\right) + q\left(-1\right) = p - q.$$

This value is positive, zero, or negative depending on the relative values of $p$ and $q$. We can proceed to compute $\mathbf{Var}\left[X\right]$ by first computing

$$\mathbf{E}\left[X^2\right] = \int\limits_{\Omega} x^2 d\mathbf{P} = \int\limits_{-\infty}^{\infty} k^2\left(p\mathbf{1}_{k=1} + q\mathbf{1}_{k=-1}\right) = p\left(1\right) + q\left(1\right) = p + q.$$

Thus

$$\begin{aligned}\mathbf{Var}\left[X\right] &= \mathbf{E}\left[X^2\right] - \mathbf{E}^2\left[X\right] = 1 - \left(p - q\right)^2 \\ &= \left(p + q\right)^2 - \left(p - q\right)^2 = 4pq.\end{aligned}$$

Then if $W_n = \sum\limits_{i=1}^{n} X_i$ then

$$\mathbf{E}\left[W_n\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i-1}^{n} \mathbf{E}\left[X_i\right] = n\left(p - q\right), \text{ and } \mathbf{Var}\left[W_n\right] = \mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i-1}^{n} \mathbf{Var}\left[X_i\right] = \mathbf{4}npq$$

(independence of the $X_i$'s from each other permits interchanging the variance and summation in this computation).

We should note that while there is little surprise about the expectation of $W_n$, the fact that its variance increases over time regardless of the mean is worthy of comment. This is not a process that hovers close to its mean as *n* increases. In fact the longer the random walk is permitted to run, the more extreme its excursions, as the variance is unbounded.

We can now find the probability function for $W_n = \sum\limits_{i=1}^{n} X_i$. Recall that we had a similar problem for computing the distribution for the binomial distribution. There, we found the probability generating function for $X_i$, $\mathbf{G}_X\left(s\right)$, and computed $\mathbf{G}_W\left(s\right) = \left[\mathbf{G}_X\left(s\right)\right]^n$. We proceed analogously here, computing $\mathbf{G}_X\left(s\right) = \int\limits_{\Omega} s^x dP = \int\limits_{-\infty}^{\infty} s^k\left[p\mathbf{1}_{k=1} + q\mathbf{1}_{k=-1}\right] = ps + qs^{-1}.$ Then $\mathbf{G}_W\left(s\right) = \left(ps + qs^{-1}\right)^n$. Using the binomial theorem, we can write this as

$$\begin{aligned}\mathbf{G}_W\left(s\right) = \left(ps + qs^{-1}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \left(ps\right)^k \left(qs^{-1}\right)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} s^{2k-n}.\end{aligned}$$

We now have to simply collect the coefficients of $s$. Let's now denote $\binom{n}{k}p^k q^{n-k}$ as $C(n,k)$. We can write $\mathbf{G}_W(s)=\sum_{k=0}^{n} C(k,n)\,s^{2k-n}$, and can collect the coefficients that are identified with each exponent of $s$. Let's begin with $n = 4$. Then we can quickly see how the coefficients of $s$ match with powers of $s$ (Table 1).

<Table 1>>

And we see that negative exponents of $s$ are entirely legitimate. This is because negative values of the random walk are quite possible (indeed, probable for small values of $p$). However, only even values of its exponent (positive or negative) are permitted. Since $G_W(s)=\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}s^{2k-n}$ we could write $\mathbf{P}\left[W_4=2k-4\right]=\binom{4}{k}p^k q^{4-k}$, but this is somewhat awkward. Instead we can write $j=2k-4$, or $k=\frac{j+4}{2}$ and $4-k=\frac{4-j}{2}$. Then we can write

$$\mathbf{P}\left[W_4=j\right]=\binom{4}{\frac{j+4}{2}}p^{\frac{j+4}{2}}q^{\frac{4-j}{2}}1_{j=-4,-2,-0,2,4},$$ or of course

$$\mathbf{P}\left[W_4=k\right]=\binom{4}{\frac{k+4}{2}}p^{\frac{k+4}{2}}q^{\frac{4-k}{2}}1_{k=-4,-2,-0,2,4}.$$ We can generalize this for any positive even value of $n$ as

$$\mathbf{P}\left[W_n=k\right]=\binom{n}{\frac{n+k}{2}}p^{\frac{n+k}{2}}q^{\frac{n-k}{2}}1_{-n\le k\le n,\frac{n+k}{2}\bmod 2=0}.$$

This solution also works for odd values of $n$.

### *Simple random walk with rest*

A solution is also available for a random walk when there is a resting state, i.e., $\mathbf{P}\left[X_i=0\right]=r$. Following our development of random walk, we can easily write

$$\begin{aligned}\mathbf{E}\left[X\right]&=\int_{\Omega} x\,dP=\int_{-\infty}^{\infty} k\left[p1_{k=1}+r1_{k=0}+q1_{k=-1}\right]\\&=p(1)+r(0)+q(-1)=p-q.\end{aligned}$$

Similarly, $\mathbf{Var}\left[X\right]=4pq$, and as before, if $W_n=\sum_{k=1}^{n}X_i$, then $\mathbf{E}\left[W_n\right]=n\left(p-q\right)$, and $\mathbf{Var}\left[W_n\right]=4npq$. Thus, the moments of the original random walk and random walk at rest are equivalent.

We can now find the probability function for $W_n=\sum_{i=1}^{n}X_i$. As before, we write

$\mathbf{G}_W\left(s\right)=\left[\mathbf{G}_X\left(s\right)\right]^n$. We proceed analogously here, computing

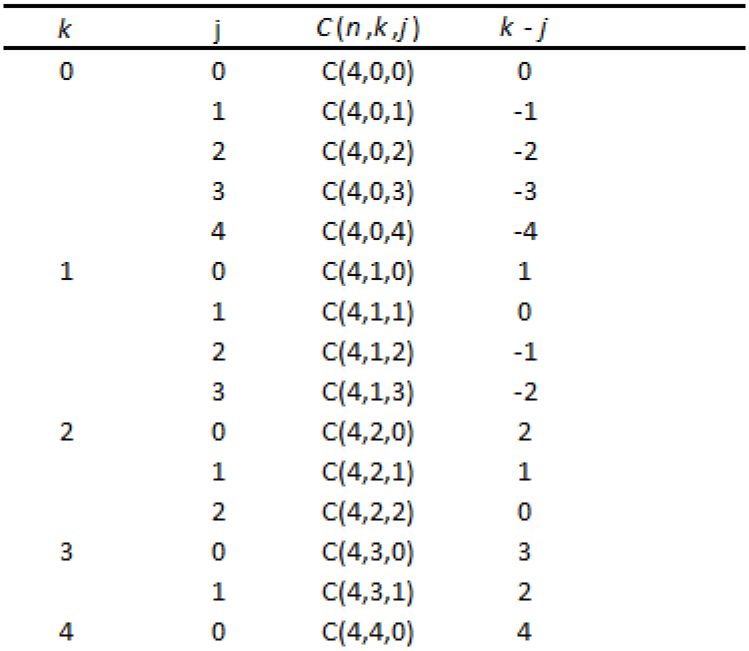
$$\begin{aligned}\mathbf{G}_W\left(s\right)&=\int_{\Omega}s^x dP=\int_{-\infty}^{\infty}s^k\left[p\mathbf{1}_{k=1}+r\mathbf{1}_{k=0}+q\mathbf{1}_{k=-1}\right]\\&=ps+qs^{-1}+r.\end{aligned}$$

Then $G_W(s)=\left(ps+qs^{-1}+r\right)^n$. Using the multinomial theorem,[*] we write

$$\begin{aligned}\mathbf{G}_W\left(s\right)&=\left(ps+qs^{-1}+r\right)^n\\&=\sum_{k=0}^{n}\sum_{j=0}^{n}\binom{n}{j\quad k}\left(ps\right)^k\left(qs^{-1}\right)^j r^{n-k-j}\mathbf{1}_{0\le k\le n,0\le j\le n}\\&=\sum_{k=0}^{n}\sum_{j=0}^{n}\binom{n}{j\quad k}p^k q^j r^{n-k-j}s^{k-j}\mathbf{1}_{0\le k\le n,0\le j\le n}.\end{aligned}$$

We now do a simple tabulation to see how powers of *s* are generated (Table 2).

Table 2. Example of Collecting Coeficients for *n* =4
Random walk with rest



| *k* | j | *C*(*n*,*k*,*j*) | *k* - *j* |
|---|---|---|---|
| 0 | 0 | C(4,0,0) | 0 |
| | 1 | C(4,0,1) | -1 |
| | 2 | C(4,0,2) | -2 |
| | 3 | C(4,0,3) | -3 |
| | 4 | C(4,0,4) | -4 |
| 1 | 0 | C(4,1,0) | 1 |
| | 1 | C(4,1,1) | 0 |
| | 2 | C(4,1,2) | -1 |
| | 3 | C(4,1,3) | -2 |
| 2 | 0 | C(4,2,0) | 2 |
| | 1 | C(4,2,1) | 1 |
| | 2 | C(4,2,2) | 0 |
| 3 | 0 | C(4,3,0) | 3 |
| | 1 | C(4,3,1) | 2 |
| 4 | 0 | C(4,4,0) | 4 |

[*] This is a generalization of the binomial theorem.

In this table $k-j$ reflects the powers of $s$. For example, the coefficient of $s^0$ is $C(4,0,0)+C(4,1,1)$. For $s^1$, we have $C(4,1,0)+C(4,2,1)$.For $s^3$ there is only $C(4,3,0)$, and the pattern becomes clear. The coefficients of $s^k$ are the sum of all $C(n,m,j)$ for which $m+j\le n$ and $k=m-j$. We may write this as

$$\mathbf{P}\left[W_n=k\right]=\sum_{m=0}^{n}\sum_{j=0}^{n}\begin{pmatrix} & n & \\ m & & k\end{pmatrix}p^m q^j r^{n-m-j}1_{m+j\le n, m-j=k}.$$

In the case of the random walk, the use of generating function permitted us to see how exponents of $s$ were assembled. Then, we simply observed the pattern and though the use of the indicator function, summarized how to carry out the collection.

## Sum of binomial random variables

The discussion from the previous section tells us how to find the probability distribution of $W_1+W_2$ where $W_1$ binomial ($n_1, p$ ) is independent of $W_2$ binomial ($n_2, p$). Since each is itself the sum of Bernoulli ($p$), then we know that $W_1+W_2$ is the sum of $n_1 + n_2$ Bernoulli($p$) random variables which of course is binomial($n_1 + n_2$, $p$).

### *Example: Satellite clinics (continued)*

From the previous example of satellites clinics, assume that clinic 1 has 12 satellites each of which achieves its goal with probability $p = 0.75$. Clinic 2 has 8 satellites, and Clinic 3 has 7 satellites that achieve their goal with the same probability. What is the probability that no more than 18 satellites reach their goal?

Each of the satellites represent an i.i.d. Bernoulli trial, and there are 12 + 8 + 7 = 27 satellites. The probability that 18 reach their goal follows a binomial distribution (27, 0.75). We therefore compute

$$\int_0^{18} d\,\mathbf{P}=\sum_{k=0}^{18}\binom{27}{k}(0.75)^k(0.25)^{27-k}=0.214.$$

■

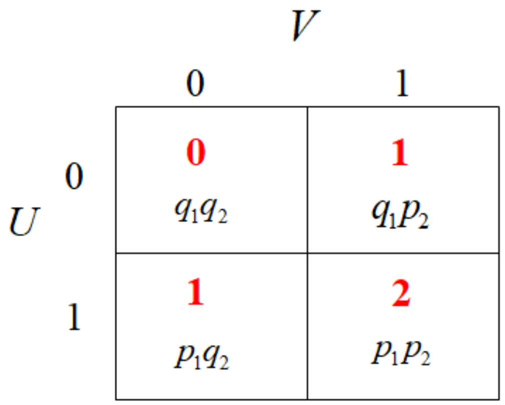
The situation is more complicated when we consider the sum of independent binomial random variables for whom the probability $p$ is different. To begin to see how to address this, let's begin with two Bernoulli random variables $U$ following Bernoulli $p_1$ and $V$ following Bernoulli $p_2$. Then what is the distribution of $Y=U+V$?

A simple table shows the possible values for the random variable $Y$ (Figure 7)

Figure 7 provides the complete elaboration of the possible values of $Y$ and the probability of those values. For example,

$$\mathbf{P}\left[Y=0\right]=\mathbf{P}\left[U=0\cap V=0\right]=\mathbf{P}\left[U=0\right]\mathbf{P}\left[V=0\right]=q_1q_2$$

The independence of $U$ and $V$ permit us to multiply the probabilities, however, the different value of $p$ mean we do not accumulate powers of $p$, but instead products of the different parameters. We will not be able to get one overarching formula, but will have to have formulas for the different possible values of $Y$. We can write



| | | $V$ | |
|---|---|---|---|
| | | 0 | 1 |
| $U$ | 0 | **0**<br>$q_1q_2$ | **1**<br>$q_1p_2$ |
| | 1 | **1**<br>$p_1q_2$ | **2**<br>$p_1p_2$ |

**Figure 7**. Distribution of the sum of two Bernoulli random variables.

$$\mathbf{P}[Y=k]=q_1q_2\mathbf{1}_{k=0}+(p_1q_2+q_1p_2)\mathbf{1}_{k=1}+p_1p_2\mathbf{1}_{k=2}.$$

Note the use of the indicator function for each of the possible values of $k$ allows us to write this complicated measure function as one function. It is easy to see that the sum of these probabilities is one. So calculating the sum of independent Bernoulli random variables with different probability parameters required enumeration, multiplying probabilities, and the use of an indicator function. We will use this experience to guide our computation of the sum of binomial random variables.

Let $W_1$ follow a binomial $(n, p_1)$, and $W_2$ follow a binomial $(n, p_2)$ . Then what does $W=W_1+W_2$ look like? We start by noting that the range of values is $0\le k\le 2n$. It pays to think of how values of $k\le n$ are produced versus values of $k>n$.

There are several ways that we can generate values of $W_1$ and $W_2$ such that $W_1+W_2=k\le n$. One is for $W_1$ =0 and $W_2=k$. We can easily calculate this probability as

$$\mathbf{P}[W_1=0\cap W_2=k]$$
$$=\mathbf{P}[W_1=0\,]\mathbf{P}[W_2=k]=\binom{n}{0}p_1^0\,q_1^n\binom{n}{k}\,p_2^k\,q_2^{n-k}.$$

However, there are more circumstances where $W_1+W_2=k$. In fact values of $(W_1,\,W_2)$ that meets this requirement are (0, $k$), (1, $k$ −1), (2, $k$–2), (3, $k$–3), …($k$, $n-k$). The probability of these possibilities is

$$\sum_{j=0}^{k}\binom{n}{j}p_1^j q_1^{n-j}\binom{n}{k-j}p_2^{k-j}q_2^{n-k+j}.$$

However, another set of possibilities are ($k$, 0), ($k$ −1, 1), ($k$ −2, 2), ($k$ −3, 3), …($n-$ $k$, $k$), producing the following probabilities

$$\sum_{j=0}^{k}\binom{n}{j}p_2^{k-j}q_2^{n-k+j}\binom{n}{k-j}p_1^{j}q_1^{n-k+j}$$

Thus we can write

$$\mathbf{P}\left[W=k\right]=$$
$$\left[\begin{array}{l}\displaystyle\sum_{j=0}^{k}\binom{n}{j}p_1^{j}q_1^{n-j}\binom{n}{k-j}p_2^{k-j}q_2^{n-k+j}\\ \displaystyle+\sum_{j=0}^{k}\binom{n}{j}p_2^{j}q_2^{n-j}\binom{n}{k-j}p_1^{k-j}q_1^{n-k+j}\end{array}\right]1_{0\le k\le n}.$$

For $n < k \le 2n$, we proceed as we did before, identifying pairs of $\left(W_1,W_2\right)$ such that $W_1+W_2=k$ for $n < k \le 2n$. For example, if $n$= 8 and $k$ = 13, then one set of collection of different values of $\left(W_1,W_2\right)$ are (5, 8), (6, 7), (7, 6), and (8,5). In general $W_1$ will go from $k-n$ to $n$, as $W_2$ moves down from $n$ to $k-n$. This is equivalent to saying that $W_2$ values are governed by $n+(k-n)-W_1=k-W_1$ However, since $p_1 \ne p_2$ we have to reverse the $\left(W_1,W_2\right)$pairs.

Thus we can write this probability as

$$\mathbf{P}\left[W=k\right]$$
$$=\left[\begin{array}{l}\displaystyle\sum_{j=k-n}^{n}\binom{n}{j}p_1^{j}q_1^{n-j}\binom{n}{k-j}p_2^{k-j}q_2^{n-k+j}\\ \displaystyle+\sum_{j=k-n}^{n}\binom{n}{j}p_2^{j}q_2^{n-j}\binom{n}{k-j}p_1^{k-j}q_1^{n-k+j}\end{array}\right]1_{n<k\le 2n}.$$

And our solution is

$$\mathbf{P}\left[W=k\right]$$
$$=\left[\sum_{j=0}^{k}\binom{n}{j}p_1^{j}q_1^{n-j}\binom{n}{k-j}p_2^{k-j}q_2^{n-k+j}+\sum_{j=0}^{k}\binom{n}{j}p_2^{j}q_2^{n-j}\binom{n}{k-j}p_1^{k-j}q_1^{n-k+j}\right]1_{0\le k\le n}.$$
$$+\left[\sum_{j=k-n}^{n}\binom{n}{j}p_1^{j}q_1^{n-j}\binom{n}{k-j}p_2^{k-j}q_2^{n-k+j}+\sum_{j=k-n}^{n}\binom{n}{j}p_2^{j}q_2^{n-j}\binom{n}{k-j}p_1^{k-j}q_1^{n-k+j}\right]1_{n<k\le 2n}$$

## Difference of binomial random variables.

Here we will compute the difference of two independent binomial random variables each with a different probability of success $p$, We will develop this probability as we did for the sum of binomial random variables by exploring the findings for the Bernoulli random variable.

To begin to see how to address this, let's begin with two Bernoulli random variables $U$ following Bernoulli $p_1$ and $V$ following Bernoulli $p_2$. Then what is the distribution of $Y=U-V$?

A simple table shows the possible values for the random variable $Y$ (Figure 8).

Following the work of the previous section, we can write the probability distribution of $Y=U-V$ as

$$\mathbf{P}[Y=k]=q_1p_2\mathbf{1}_{k=-1}+(q_1q_2+p_1p_2)\mathbf{1}_{k=0}+p_1q_2\mathbf{1}_{k=1}.$$

We will proceed in this manner for the binomial distribution. Assume $W_1$ follows a binomial distribution $(n, p_1)$ and $W_2$ follows a binomial distribution $(n, p_2)$, and $W=W_1-W_2$. We must consider two cases, for $W<0$ and $W\geq 0$. For $W=k<0$, $W_2$ is greater than $W_1$ and we must consider the following $(W_1, W_2)$ pairs $(0, -k)$, $(1, -k+1)$, $(2, -k+2)$.... $(n+k, n)$. The probability of this collection of $(W_1, W_2)$ pairs is $\sum_{j=0}^{n+k}\binom{n}{j}p_1^jq_1^{n-j}\binom{n}{j-k}p_2^{j-k}q_2^{n-j+k}$. For $k\geq 0$ we must consider the following $(W_1, W_2)$ pairs: $(k, 0)$, $(k+1, 1)$, $(k+2, 2)$, $(k+3, 3)$, …$(n, n-k)$. Their probability is $\sum_{j=k}^{n}\binom{n}{j}p_1^jq_1^{n-j}\binom{n}{j-k}p_2^{j-k}q_2^{n-j+k}$. Thus we can write

|  |  | $V$ |  |
|---|---|---|---|
|  |  | 0 | 1 |
| $U$ | 0 | **0**<br>$q_1q_2$ | **-1**<br>$q_1p_2$ |
|  | 1 | **1**<br>$p_1q_2$ | **0**<br>$p_1p_2$ |

**Figure 8.** Distribution of the difference of two Bernoulli random variables.

$$\mathbf{P}[W_1-W_2=k]=\sum_{j=0}^{n+k}\binom{n}{j}p_1^jq_1^{n-j}\binom{n}{j-k}p_2^{j-k}q_2^{n-j+k}\mathbf{1}_{k<0}$$
$$+\sum_{j=k}^{n}\binom{n}{j}p_1^jq_1^{n-j}\binom{n}{j-k}p_2^{j-k}q_2^{n-j+k}\mathbf{1}_{k\geq 0}$$

Using these probabilities, we can compute the more general probability of the event $W_1\geq W_2$. Assuming that $n$ is the same for both group, we might expect that if $p_1>p_2$, then $\mathbf{P}[W_1\geq W_2]$ will tend to be larger as well. We can compute $\mathbf{P}[W_1\geq W_2]$ by recognizing $\mathbf{P}[W_1\geq W_2]=\mathbf{P}[W_1-W_2=k]$ for all $k>0$. Since the events are disjoint for each $k$, then the desired probability $\mathbf{P}[W_1\geq W_2]$ is simply the sum of $\mathbf{P}[W_1-W_2=k]$ for each $k>0$. Therefore

$$\mathbf{P}[W_1\geq W_2]=\sum_{k=0}^{n}\sum_{j=k}^{n}\binom{n}{j}p_1^jq_1^{n-j}\binom{n}{j-k}p_2^{j-k}q_2^{n-j+k}$$

## Conditional binomial measure

We can now apply our earlier discussion of conditional probability to our work with the Bernoulli and binomial distribution. Assume we have two independent Bernoulli random variables, $X_1$ distributed Bernoulli($p$), and $X_2$ distributed Bernoulli($p$). Assume that $X = X_1 + X_2 = 1.$ What is the probability that $X_1 = 1$?

Before we carry out a formal computation, we can try to deduce the solution. There are two possibilities for $X = 1$, either $X_1 = 1$ or $X_2 = 1$. Since $X_1$ and $X_2$ are i.i.d., then the probability that either is one is the same, so we might expect $\mathbf{P}\left[X_1 = 1 \,|\, X = 1\right]$ should be $\frac{1}{2}.$

The formal computation proceeds as follows.

$$\begin{aligned}\mathbf{P}\left[X_1 = 1 \,|\, X = 1\right] = \mathbf{P}\left[X_1 = 1 \,|\, X_1 + X_2 = 1\right] &= \frac{\mathbf{P}\left[X_1 = 1 \cap X_1 + X_2 = 1\right]}{\mathbf{P}\left[X_1 + X_2 = 1\right]} \\ &= \frac{\mathbf{P}\left[X_1 = 1 \cap X_2 = 0\right]}{\mathbf{P}\left[X_1 + X_2 = 1\right]} = \frac{\mathbf{P}\left[X_1 = 1\right]\mathbf{P}\left[X_2 = 0\right]}{\mathbf{P}\left[X_1 + X_2 = 1\right]} \\ &= \frac{pq}{\binom{2}{1}pq} = \frac{1}{2}.\end{aligned}$$

Note the numerator simplified to the product of probabilities since the value of $X_1$ fixes the value of $X_2$, and $X_1$ and $X_2$ are independent.

We can generalize this problem to find

$$\begin{aligned}&\mathbf{P}\left[X_1 = 1 \,|\, X_1 + X_2 + X_3 + \ldots + X_n = 1\right] \\ &= \frac{\mathbf{P}\left[X_1 = 1 \cap X_1 + X_2 + X_3 + \ldots + X_n = 1\right]}{\mathbf{P}\left[X_1 + X_2 + X_3 + \ldots + X_n = 1\right]} \\ &= \frac{\mathbf{P}\left[X_1 = 1 \cap X_2 + X_3 + \ldots + X_n = 0\right]}{\mathbf{P}\left[X_1 + X_2 + X_3 + \ldots + X_n = 1\right]} = \frac{p\binom{n-1}{0}q^{n-1}}{\binom{n}{1}pq^{n-1}} = \frac{1}{n}.\end{aligned}$$

This is, as before, the solution we might have intuited.

We can now proceed with the solution to the following problem. Let $W_1$ be a random variable that follow a binomial$(n_1, p)$, and $W_2$ be a random variable that follow a binomial $(n_2, p)$.Given than $W_1 + W_2 = m,$ what is the probability that $W_1 = k$? Clearly the probability is zero for $k < 0$ or $k > n$. For $0 \le k \le n,$ we compute,

$$\mathbf{P}\left[W_1 = k \mid W_1 + W_2 = m\right] = \frac{\mathbf{P}\left[W_1 = k \cap W_1 + W_2 = m\right]}{\mathbf{P}\left[W_1 + W_2 = m\right]}$$

$$= \frac{\mathbf{P}\left[W_1 = k \cap W_2 = m - k\right]}{\mathbf{P}\left[W_1 + W_2 = m\right]} = \frac{\binom{n_1}{k} p^k q^{n-k} \binom{n_2}{m-k} p^{m-k} q^{n-k-(m-k)}}{\binom{n_1+n_2}{m} p^m q^{n-m}}$$

$$= \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n_1+n_2}{m}}$$

This solution we have seen before as the result of sampling without replacement and we will see later is the hypergeometric distribution.

### *Example: Clinical trial recruitment*

Two separate consortiums of clinics recruit subjects to participate in a clinical trial. Consortium 1 has sixty patients from which it could recruit, and Consortium 2 has seventy patients from which it could recruit patients for the study. Each has the same probability of recruiting a subject to the study. The total number of subjects recruited by the consortiums combined is 87. What is the probability that forty patients came from consortium 1?

We can write

$$\frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n_1+n_2}{m}} = \frac{\binom{60}{40}\binom{70}{47}}{\binom{130}{87}} = \frac{\left(4.19\text{x}10^{15}\right)\left(1.79\text{x}10^{18}\right)}{\left(5.08\ \text{x}10^{34}\right)} = 0.148.$$

## Introduction to random variables functions

These computations that we have just completed for sums and differences of Bernoulli and binomial random variables are the introduction to managing not just random variables, but functions of random variables.

Our experience reminds us that we must ensure that the function is measurable, which requires us to focus on the σ-algebra.

Consider the following elementary example. Let $W$ follow a binomial ($n$, $p$). Let $Y = -$W. What is the distribution of the new random variable $Y$?

Before we apply mathematics to this problem, we should think about its construction. We know $Y$ cannot follow the binomial distribution, since it is negatively valued. However, probabilities for its values are related to the binomial distribution. For example $\mathbf{P}\left[Y = -3\right] = \mathbf{P}\left[W = 3\right] = \binom{n}{3} p^3 \left(1-p\right)^{n-3}$. We can use this relationship to find the probability that $Y = k$, for $-n \le k \le 0$. However the $Y$ σ-algebra (which allows no positive values) is separate and distinct from the $W$ based σ-algebra. Thus, in order to compute the probabilities for $Y$ we have to first map the $Y$ σ-algebra to $W$ σ-algebra, and use this map to find the probabilities. We begin with

$$\mathbf{P}\left[Y = y\right] = \mathbf{P}\left[-W = y\right] = \mathbf{P}\left[W = -y\right].$$

We can then finish the computation in general.

$$\mathbf{P}\left[Y=k\right]=\mathbf{P}\left[W=-k\right]=\binom{n}{-k}p^{-k}\left(1-p\right)^{n+k}.$$

In this case mapping negative integers to positive integers is straightforward, and we actually give little thought to mapping the probabilities.

Consider though the following mapping. Let *W* follow a binomial (2, *p*). If $Y=W^2$, what is the probability distribution of *Y*?

Following the previous example, we can write

$$\mathbf{P}\left[Y=y\right]=\mathbf{P}\left[W^2=y\right]=\mathbf{P}\left[W=\sqrt{y}\ \cap\ W=-\sqrt{y}\right]$$
$$=\mathbf{P}\left[W=\sqrt{y}\ \right]$$

The evaluation of *Y*, due to the nature of *W*, produces two possible values of *Y,* but only one has nonzero probability.

$$\mathbf{P}\left[Y=k\right]=\mathbf{P}\left[W^2=k\right]=\mathbf{P}\left[W=\sqrt{k}\ \right]=\binom{n}{\sqrt{k}}p^{\sqrt{k}}q^{n-\sqrt{k}}.$$

For $k$ = 0, 1, 4. We will build on this concept of functions of random variables in later sections.

## Mixtures of binomial random variables

We have described several combinations of binomial random variables. However, one additional combination is to simply combine the distributions of independent binomial random variables.

For example, we understand the characteristics of $W_1$ that follows a binomial (20, 0.35) distribution, as we do the features of $W_2$ that follows a binomial (20, 0.80) distribution. We can compute their moments, and easily provide a graph of both (Figure 9)

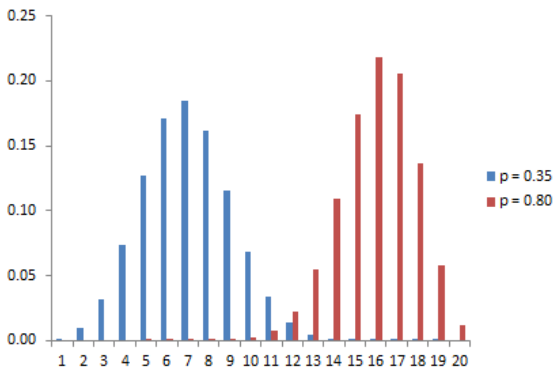


**Figure 9**. Distribution of the two independent binomial distributions as a prelude to mixing.

Previous sections have shown how to create new random variables from each of $W_1$ and $W_2$. However, suppose we wished to actually combine the two distributions. Essentially, we would modify the experiment that generates our outcome. Initially, for example, we select an observation from a binomial (20, 0.35).

However, now, we conduct two experiment. First, select an outcome from a Bernoulli experiment. If the outcome is 1, we select a realization of $W_1$. It the outcome is 0, we select the random variable $W_2$. This two stage selection process has important implications for the distribution of the result, producing  a rich combination of probability distributions (Figure 10).

How can we even write this mathematically? We could try to write the probability function simply as

$$\left[\binom{20}{k}(0.35)^k(0.65)^{n-k}+\binom{20}{k}(0.80)^k(0.20)^{n-k}\right]1_{0\le k\le 20}$$

However, this we know does not integrate one.

$$\begin{aligned}\int_\Omega d\mathbf{P} &= \int_\Omega\left[\binom{20}{k}(0.35)^k(0.65)^{n-k}+\binom{20}{k}(0.80)^k(0.20)^{n-k}\right]1_{0\le k\le 20}\\ &= \int_\Omega\binom{20}{k}(0.35)^k(0.65)^{n-k}1_{0\le k\le 20}+\int_\Omega\binom{20}{k}(0.80)^k(0.20)^{n-k}1_{0\le k\le 20}\\ &=1+1=2\end{aligned}$$

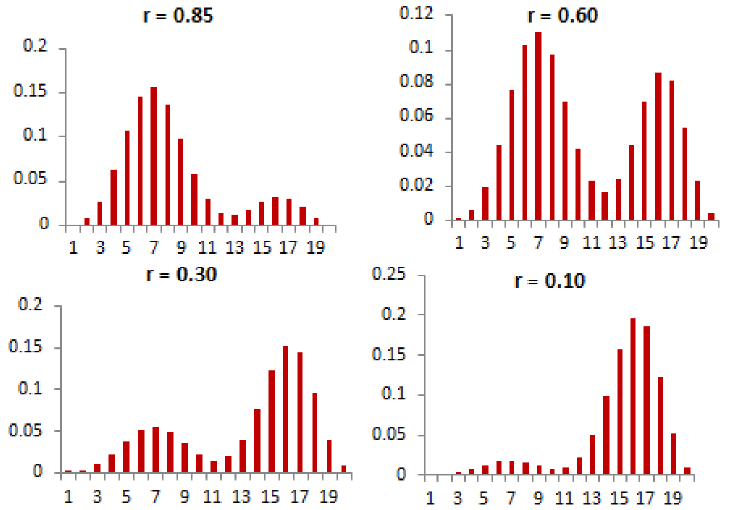


**Figure 10**. Different mixtures of the two binomial (20, 0.35) and binomial (20, 0.80). The parameter $r$ is the probability of the binomial (20, 0.35).

We therefore have to include a parameter that ensures the probability function integrates to one. We introduce a parameter $r$, $0 \le r \le 1$, and write

$$\begin{aligned}\mathbf{P}[W=k] &= r\binom{20}{k}(0.35)^k(0.65)^{n-k}1_{0\le k\le 20}\\ &\quad+(1-r)\binom{20}{k}(0.80)^k(0.20)^{n-k}1_{0\le k\le 20}\end{aligned}$$

and we now have a probability function that integrates to one, using the concept of the Lebesgue integral

$$\begin{aligned}\int_\Omega d\mathbf{P} &= \int_\Omega r\binom{20}{k}(0.35)^k(0.65)^{n-k}\mathbf{1}_{0\le k\le 20} + (1-r)\binom{20}{k}(0.80)^k(0.20)^{n-k}\mathbf{1}_{0\le k\le 20}\\ &= \int_\Omega r\binom{20}{k}(0.35)^k(0.65)^{n-k}\mathbf{1}_{0\le k\le 20} + \int_\Omega (1-r)\binom{20}{k}(0.80)^k(0.20)^{n-k}\mathbf{1}_{0\le k\le 20}\\ &= r+(1-r)=1.\end{aligned}$$

### *Example: Contemporaneous cell therapy*

Cardiac cell therapy studied the direct injection of mesenchymal cells into myocardial tissue. Consider the following: A subject receives a complete set of twenty injections that can be placed either in the scar itself (scar-based) or at the periphery of the scar (border-based).

The probability that the border based injections generate new cardiac cells is 0.65. However, the probability a patient receives border-based injections is 0.35. The likelihood of a patient getting scar-based injections is 0.65, and the probability that these injections of produce new cells at the injection site is 0.25. What is the probability that there were be at least ten injections that produce new cell growth?

We compute

$$\begin{aligned}\int_{k\ge 10} d\mathbf{P} &= \int_{k\ge 10} r\binom{20}{k}p_1^k(1-p_1)^{n-k} + \int_{k\ge 10}(1-r)\binom{20}{k}p_2^k(1-p_2)^{n-k}\\ &= \int_{k\ge 10}(0.35)\binom{20}{k}(0.65)^k(0.35)^{n-k} + \int_{k\ge 10}(0.65)\binom{20}{k}(0.25)^k(0.75)^{n-k}\\ &= \int_{k\ge 10}(0.35)\binom{20}{k}(0.65)^k(0.35)^{n-k} + \int_{k\ge 10}(0.65)\binom{20}{k}(0.25)^k(0.75)^{n-k}\\ &= (0.35)(0.946)+(0.65)(0.014)\\ &= 0.34.\end{aligned}$$

■

Yet another example of the binomial distribution is as the probability distribution of survivors from what is classically known as the death process.

These are just some examples from the rich bounty of the binomial distribution. The concepts developed here will be key in our future discussions.

Next Sections

Multinomial Distribution

Hypergeometric Measure

Geometric and Negative binomial measures

General Poisson Process

Survival Measure: Exponential, Gamma, and Related

Cauchy, Laplace, and Double Exponential

Continuous Probability Measure

Moment and Probability Generating Functions

Variable Transformations

Uniform and Beta Measure

Normal Measure

Compounding

F and T Measure

Ordering Random Variables
Asymptotics
Tail Event Measure

# Skewness and Kurtosis for the Binomial Distribution

We can also compute the skewness and kurtosis for binomial random variables. Let $W$ follow a binomial($n$, $p$) distribution. Then we recall that

$$\mathbf{S}(X) = \frac{\mathbf{E}\left[(W-\mu)^3\right]}{\sigma^3},$$

and

$$\begin{aligned}
\mathbf{E}\left[(W-\mu)^3\right] &= \mathbf{E}\left[W^3\right] - 3\mathbf{E}\left[W^2\right]\mu + 3\mathbf{E}[W]\mu^2 - \mu^3 \\
&= \mathbf{E}\left[W^3\right] - 3\mathbf{E}\left[W^2\right]\mu + 3\mu^3 - \mu^3 \\
&= \mathbf{E}\left[W^3\right] - 3\mu\left(\mathbf{E}\left[W^2\right] - \mu^2\right) - \mu^3 \\
&= \mathbf{E}\left[W^3\right] - 3\mu\sigma^2 - \mu^3 \\
&= \mathbf{E}\left[W^3\right] - 3(np)(npq) - (np)^3 \\
&= \mathbf{E}\left[W^3\right] - 3n^2p^2q - (np)^3
\end{aligned}$$

We proceed as we have before for the binomial distribution.

$$\begin{aligned}
&\mathbf{E}\left[W(W-1)(W-2)\right]\\
&=\int_{\Omega_W} w(w-1)(w-2)d\mathbf{P}=\sum_{k=0}^{n}k(k-1)(k-2)\mathbf{P}[W=k]\\
&=\sum_{k=0}^{n}k(k-1)(k-2)\binom{n}{k}p^kq^{n-k}\\
&=\sum_{k=3}^{n}k(k-1)(k-2)\binom{n}{k}p^kq^{n-k}\\
&=\sum_{k=3}^{n}n(n-1)(n-2)\binom{n-3}{k-3}p^kq^{n-k}\\
&=n(n-1)(n-2)p^3\sum_{k=3}^{n}\binom{n-3}{k-3}p^{k-3}q^{n-k}\\
&=n(n-1)(n-2)p^3\sum_{j=0}^{m}\binom{m}{j}p^jq^{m-j}=n(n-1)(n-2)p^3.
\end{aligned}$$

We now write $\mathbf{E}\left[W(W-1)(W-2)\right]=\mathbf{E}\left[W^3\right]-3\mathbf{E}\left[W^2\right]+2\mathbf{E}[W]$, or $\mathbf{E}\left[W^3\right]=\mathbf{E}\left[W(W-1)(W-2)\right]+3\mathbf{E}\left[W^2\right]-2\mathbf{E}[W]$. We can now write

$$\begin{aligned}
\mathbf{E}\left[W^3\right]&=\mathbf{E}\left[W(W-1)(W-2)\right]+3\mathbf{E}\left[W^2\right]-2\mathbf{E}[W]\\
&=n(n-1)(n-2)p^3+3\left(n(n-1)p^2+np\right)-2np.
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{E}\left[(W-\mu)^3\right]=\mathbf{E}\left[W^3\right]-3n^2p^2q-(np)^3\\
&=n(n-1)(n-2)p^3+3\left(n(n-1)p^2+np\right)-2np-3n^2p^2(1-p)-(np)^3\\
&=n^3p^3-3n^2p^3+2np^3+3n^2p^2-3np^2+np-3n^2p^2+3n^2p^3-n^3p^3\\
&=2np^3-3np^2+np\\
&=np\left(-3p^2+2p^3+1\right)=np\left(-3p^2+2p^3+p+q\right)=np\left(-2p^2+2p^3+q\right)\\
&=np\left(-2p(1-p)+q\right)=np\left(-2pq+q\right)=npq(-2p+1)=npq(1-2p)
\end{aligned}$$

Therefore

$$\mathbf{S}(X)=\frac{\mathbf{E}\left[(W-\mu)^3\right]}{\sigma^3}=\frac{npq(1-2p)}{npq\sqrt{npq}}=\frac{(1-2p)}{\sqrt{npq}}$$

## Kurtosis

We compute $\mathbf{K}(X)=\dfrac{\mathbf{E}\left[(W-\mu)^4\right]}{\sigma^4}-3.$ Our computation is aided by observing

$$\begin{aligned}
\mathbf{E}\left[(W-\mu)^4\right]&=\mathbf{E}\left[(W-\mu)(W-\mu)^3\right]\\
&=\mathbf{E}\left[W(W-\mu)^3-\mu(W-\mu)^3\right]
\end{aligned}$$

We can write $\mathbf{E}\left[\mu(W-\mu)^3\right] = \mu\mathbf{E}\left[(W-\mu)^3\right]$, and we know

$\mathbf{E}\left[(W-\mu)^3\right] = npq(1-2p)$. We now need to compute

$$\begin{aligned}
&\mathbf{E}\left[W(W-\mu)^3\right]\\
&= \mathbf{E}\left[W\left(W^3 - 2W^2u - 2W\mu^2 - \mu^3\right)\right]\\
&= \mathbf{E}\left[W^4 - 2\mu W^3 - 2\mu^2 W^2 - \mu^4\right]\\
&= \mathbf{E}\left[W^4\right] - 2\mu\mathbf{E}\left[W^3\right] - 2\mu^2\mathbf{E}\left[W^2\right] - \mu^4
\end{aligned}$$

Since from previous computations, we know that

$$\begin{aligned}
\mathbf{E}\left[W^3\right] &= n(n-1)(n-2)p^3 + 3\left(n(n-1)p^2 + np\right) - 2np\\
&= n^3p^3 - 3n^2p^3 + 2np^3 + 3n^2p^2 - 3np^2 + np.
\end{aligned}$$

and

$$\mathbf{E}\left[W^2\right] = n(n-1)p^2 + np$$

Then

$$\begin{aligned}
&\mathbf{E}\left[W(W-\mu)^3\right] = \mathbf{E}\left[W^4\right] - 2\mu\mathbf{E}\left[W^3\right] - 2\mu^2\mathbf{E}\left[W^2\right] - \mu^4\\
&= \mathbf{E}\left[W^4\right] - 2np\left[n^3p^3 - 3n^2p^3 + 2np^3 + 3n^2p^2 - 3np^2 + np\right]\\
&\quad - 2n^2p^2\left(n(n-1)p^2 + np\right) - n^4p^4\\
&= \mathbf{E}\left[W^4\right] - 2n^4p^4 + 6n^3p^4 - 4n^2p^4 - 6n^3p^3 + 6n^2p^3 - 2n^2p^2\\
&\quad - 2n^4p^4 + 2n^3p^4 - 2n^3p^3 - n^4p^4.
\end{aligned}$$

We proceed as we have before for the binomial distribution, computing

$$
\begin{aligned}
&\mathbf{E}\left[W(W-1)(W-2)(W-3)\ \right] \\
&= \int\limits_{\Omega_W} w(w-1)(w-2)(w-3)\,d\mathbf{P} \\
&= \sum_{k=0}^{n} k(k-1)(k-2)(k-3)\,\mathbf{P}[W=k] \\
&= \sum_{k=0}^{n} k(k-1)(k-2)(k-3)\binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=4}^{n} k(k-1)(k-2)(k-3)\binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=4}^{n} n(n-1)(n-2)(n-3)\binom{n-4}{k-4} p^k q^{n-k} \\
&= n(n-1)(n-2)(n-3)\,p^4 \sum_{k=4}^{n}\binom{n-4}{k-4} p^{k-3} q^{n-k} \\
&= n(n-1)(n-2)(n-3)\,p^4 \sum_{j=0}^{m}\binom{m}{j} p^{j} q^{m-j} \\
&= n(n-1)(n-2)(n-3)\,p^4.
\end{aligned}
$$

We now write

$$
\begin{aligned}
&\mathbf{E}\left[W(W-1)(W-2)(W-3)\right] \\
&= \mathbf{E}\left[W^4\right] - 6\mathbf{E}\left[W^3\right] + 11\mathbf{E}\left[W^2\right] - 6\mathbf{E}[W],
\end{aligned}
$$

Or

$$
\begin{aligned}
&\mathbf{E}\left[W^4\right] \\
&= \mathbf{E}\left[W(W-1)(W-2)(W-3)\right] + 6\mathbf{E}\left[W^3\right] - 11\mathbf{E}\left[W^2\right] + 6\mathbf{E}[W],
\end{aligned}
$$

Continuing,

$$
\begin{aligned}
\mathbf{E}\left[W^4\right] &= \mathbf{E}\left[W(W-1)(W-2)(W-3)\right] \\
&\quad + 6\mathbf{E}\left[W^3\right] - 11\mathbf{E}\left[W^2\right] + 6\mathbf{E}[W], \\
&= n(n-1)(n-2)(n-3)\,p^4 + 6n(n-1)(n-2)\,p^3 \\
&\quad + 18n(n-1)\,p^2 + 6np - 11\left[n(n-1)\,p^2 + np\right] - 6np \\
&= n^4 p^4 - 6n^3 p^4 + 11n^2 p^4 - 6np^4 + 6n^3 p^3 - 6n^2 p^3 \\
&\quad + 7n^2 p^2 - 7np^2 - 11np.
\end{aligned}
$$

We now write

$$
\begin{aligned}
&\mathbf{E}\left[W(W-\mu)^3\right] \\
&= \mathbf{E}\left[W^4\right] - 2\mu\mathbf{E}\left[W^3\right] - 2\mu^2\mathbf{E}\left[W^2\right] - \mu^4 \\
&= n^4 p^4 - 6n^3 p^4 + 11n^2 p^4 - 6np^4 + 6n^3 p^3 - 6n^2 p^3 + 7n^2 p^2 - 7np^2 - 11np \\
&- 2n^4 p^4 + 6n^3 p^4 - 4n^2 p^4 - 6n^3 p^3 + 6n^2 p^3 - 2n^2 p^2 \\
&- 2n^4 p^4 + 2n^3 p^4 - 2n^3 p^3 - n^4 p^4.
\end{aligned}
$$

And

$$\mathbf{E}\left[(W-\mu)^4\right]=\mathbf{E}\left[W(W-\mu)^3-\mu(W-\mu)^3\right]$$

$$= n^4p^4-6n^3p^4+11n^2p^4-6np^4+6n^3p^3-6n^2p^3+7n^2p^2-7np^2-11np$$
$$-2n^4p^4+6n^3p^4-4n^2p^4-6n^3p^3+6n^2p^3-2n^2p^2$$
$$-2n^4p^4+2n^3p^4-2n^3p^3-n^4p^4-np\left(npq\left(1-2p\right)\right)$$
$$= n^4p^4-6n^3p^4+11n^2p^4-6np^4+6n^3p^3-6n^2p^3+7n^2p^2-7np^2-11np$$
$$-2n^4p^4+6n^3p^4-4n^2p^4-6n^3p^3+6n^2p^3-2n^2p^2$$
$$-2n^4p^4+2n^3p^4-2n^3p^3-n^4p^4-n^2p^2+3n^2p^3-2n^2p^4$$

Which simplifies to

$$n^4p^4-6n^3p^4+11n^2p^4-6np^4+6n^3p^3-6n^2p^3+7n^2p^2-7np^2$$
$$-11np-2n^4p^4+6n^3p^4-4n^2p^4-6n^3p^3+6n^2p^3-2n^2p^2$$
$$-2n^4p^4+2n^3p^4-2n^3p^3-n^4p^4-n^2p^2+3n^2p^3-2n^2p^4$$

$$=5n^2p^4-6np^4+4n^2p^2-7np^2-11np-2n^4p^4$$
$$-2n^4p^4+2n^3p^4-2n^3p^3+3n^2p^3$$

Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Multinomial Measure

The multinomial distribution is a generalization of the binomial distribution, an observation permitting us to bring the insight from that probability mass function and measure to this related probability tool

## Prerequisites

The Notion of Random Events
Elementary Set Theory
The Binomial Distribution

## Developing the trinomial distribution

The binomial distribution was based on Bernoulli trials, which themselves were quite intuitive. However a simple adaptation of the Bernoulli trial opens the door to a host of new probability distribution.

## Example: Left ventricular assist devices

The heart is a pump, and when it fails, due to previous heart attacks, infection (myocarditis), or chronic disease (e.g., hypertension or diabetes mellitus), it no longer pumps adequate blood forward to the brains, kidneys, muscles, and other organs, letting blood back up in the heart. This is called heart failure.

Some patients with severe heart failure currently are treated with left ventricular assist devices (LVADs). By carrying out some of the work of the failing heart, they can help to improve the heart's function.

However, the LVAD is only temporary, leaving three possible outcomes of successful LVAD placement. The first is that the patient goes on to die, a death that is postponed but eventually occurs. The second is that the patient is sustained long enough to have a heart transplant (where the old heart and the LVAD are replaced by another human heart). The third is that the patient is able to recover enough heart function with the LVAD and drug therapy that they can have the LVAD removed and perform very well (termed rescue therapy).

Each of these events is mutually exclusive. Let's assume that the probability of rescue is 0.10, the probability of transplant is 0.35, and the probability of death is 0.55. At a major heart failure treatment center, there are 35 patients who will have an LVAD this year. What is the likelihood that 20 die, 10 have heart transplants, and the remaining 5 are rescued.

We will approach this as we did for the binomial solution, breaking the problem into two parts. One is computing the probability that in this sequence of 35 patients, we have 20 deaths, 10 transplants, and 5 rescues. It is reasonable to conclude that one patient's experience is independent of another, we can compute the probability that this occurs as $(0.55)^{20}(0.35)^{10}(0.10)^{5}$. However, there are many sequences of events that will produce this collection of events, and we must find a way to count them.

Fortunately, this is straightforward, and we can rely on our earlier work in counting. Beginning with the deaths, we see that there are 35 patients, and any 20 of them can die. We count the number of ways this can happen as $\binom{35}{20}$. However, once this has occurred, there are $35 - 20 = 15$ patients from whom we select 10 transplants. So the number of ways to choose 20 deaths then 10 transplants is $\binom{35}{20}\binom{15}{10}$.

Continuing, we now have five patients left all of whom are to be rescue patients. This completes the computation, and we have

$$\binom{35}{20}\binom{15}{10}\binom{5}{5} = \left(\frac{35!}{20!15!}\right)\left(\frac{15!}{10!5!}\right)\left(\frac{5!}{5!0!}\right) = \frac{35!}{20!10!5!}.$$

Note that the cancellation that gets us this simple result is not just a property of these particular numbers selected for the problem. Selecting 20 transplants leaves us exactly 15 patients left, and it is precisely these 15 patients from whom the 10 transplants must be selected. We write this result as

$$\frac{35!}{20!10!5!} = \binom{35}{20\,10\,5},$$

and write our final probability as

$$\binom{35}{20\,10\,5}(0.55)^{20}(0.35)^{10}(0.10)^{5} = 0.017$$

We observe that this is distribution that requires more than one random variable. It does not require three since the three classes of patients must sum to a constant. However, there are two random variables whose value is determined by the experiment.

Thus, if we have $n$ objects, and which to choose $k_1, k_2, k_3 \ldots k_n$ objects occurring independently of each other (realizations of the random variables $X_i$, $i = 1,\ldots, m$) with probability $p_1, p_1, p_1, \ldots, p_m$, such that $\sum_{i=1}^{m} k_i = n$, and $\sum_{i=1}^{m} p_i = 1$, then

$$\mathbf{P}\left[k_1, k_1, k_1, \ldots, k_m\right]$$
$$= \binom{n!}{k_1!\quad k_2!\quad k_3!\quad \ldots\quad k_m!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \ldots p_m^{k_m} \mathbf{1}_{\sum_{i=1}^{m} k_i = m}.$$

This is the general form of the multinomial distribution. Its use is based on the multinomial theorem, which states that if there is a collection of $\{x_i\}$ such that $\sum_{i=1}^{m} x_i = n$, then

$$(x_1 + x_2 + x_3 + ... + x_m)^n = \sum_{k_1+k_2+k_3+.+k_m=n} \binom{n}{k_1 \quad k_2 \quad k_3 \quad ... \quad k_m} \prod_{i=1}^{m} x_i^{k_i}.$$

We can see that this is a generalization of the binomial theorem. For the example provided in this chapter, we rely on the trinomial distribution $\sum_{k_1+k_2+k_3=n} \binom{n}{k_1 \quad k_2 \quad k_3} x_1^{k_1} x_2^{k_2} x_3^{k_3}$.

Note that we can collapse the multinomial distribution to the binomial distribution by simply aggregating the objects into two classes. This allows us to see right away that $\mathbf{E}[X_i] = np_i$, and $\mathbf{Var}[X_i] = np_i(1 - p_i)$.

Next Sections
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Hypergeometric Measure

Prerequisites
The Notion of Random Events
Elementary Set Theory
The Binomial Distribution
Counting Events

We are already familiar with hypergeometric measure from working with examples involving sampling without replacement. If we suppose there is a population of *N* objects of which *M* have the trait of interest. We take a sample of size *n* from the population of size *N*. How likely is it that our smaller sample will contain *m* of the objects that have this trait? If *X* is the random variable that is the number of objects with the trait of interest in our smaller sample, then we seek $\mathbf{P}\left[X=m\right]$.

It is typically helpful when working with hypergeometric measure, to start with the sample of interest, drawing its members from the different components of the larger population. In our sample of size *n*, we know that we must have two components; *m* objects should have the trait, and the remaining *n* – *m* should not. Thus, our *m* objects must be selected from all of those with the trait; the number of ways to do this is $\binom{M}{m}$. The remaining *n* – *m* objects in our sample must be selected from the population members without the trait. The number of ways to accomplish this are $\binom{N-M}{n-m}$. The denominator of this probability is simply the number of different possible samples of size *n* from *N*. Compute.

$$\mathbf{P}\left[X=m\right]=\frac{\binom{M}{m}\binom{N-n}{n-m}}{\binom{N}{n}}.$$

When we think of this as merely one of counting, computing the probability is straightforward.

## Example: Diabetes and insulin pumps

Diabetes mellitus is a disease of abnormal glucose metabolism and affects over twenty-six million US subjects. Commonly these patients must progress to insulin injections. In order to help stabilize the amount of insulin the body tissues are exposed to, these subjects requiring insulin can sometimes be provided an insulin pump, which automatically regulates the amount of insulin delivered. However, not every insulin dependent diabetic subject can have a pump inserted.

In a major diabetes treatment center there are 115 insulin dependent diabetics, 57 of which have the insulin pump inserted. Today, 30 patients are scheduled to be seen. The clinic staff only has resources to replenish ten pumps. What is the probability that no more than ten of these thirty patients have an insulin pump?

We first find $\mathbf{P}[X=10]$, and then compute $\mathbf{P}[X\leq 10]$. To find this first probability, we recognize that ten of our thirty patients must have the pump (and must therefore be selected from the 57 in the population who have the insulin pump), and the remaining twenty must not. We compute

$$\mathbf{P}[X=10]=\frac{\binom{57}{10}\binom{58}{20}}{\binom{115}{30}}=0.02.$$

We now compute

$$\mathbf{P}[X\leq 10]=\sum_{m=0}^{10}\frac{\binom{57}{m}\binom{58}{30-m}}{\binom{115}{30}}=0.031.$$

The clinic will very likely run out of resources for these patients today.

The calculation of the mean and variance of hypergeometric measure involves a computation, but we can show that

$$\mathbf{E}[X]=M\frac{n}{N}, \text{ and } \mathbf{Var}[X]=\left(\frac{Mn}{N}\right)\left(\frac{N-n}{N}\right)\left(\frac{N-M}{N-1}\right).$$

### *Example: Population Selection Effects*

Hypergeometric measure has many uses. One of the most interesting is its ability to aid in the detection of selection bias. Selection bias is the process by which the choice of subjects in a study can produce a bias or systematic effect on the research effort's results.

Investigators planning a clinical trial choose the subsets for the study in a sequence of selection steps. They first screen patients (screened population) helping them to decide who is most likely to satisfy the entry requirements of the study. After obtaining consent, they will then test the screened patients to determines to see if these patients meet all of the inclusion and exclusion criteria. These patients who satisfy inclusion/exclusion assessments represent the eligible population. Then after a period of time that can last from minutes to days, patients are randomized to receive therapy. Once the patient is randomized to the study, the investigators are committed to giving the patient the therapy and following them for the duration of the trial. [*]

Thus in our example there are three subsets. The first is the screening subset, the second, the eligible subset, and the third the randomized subset. Each subset is wholly contained in the other (Figure 1).

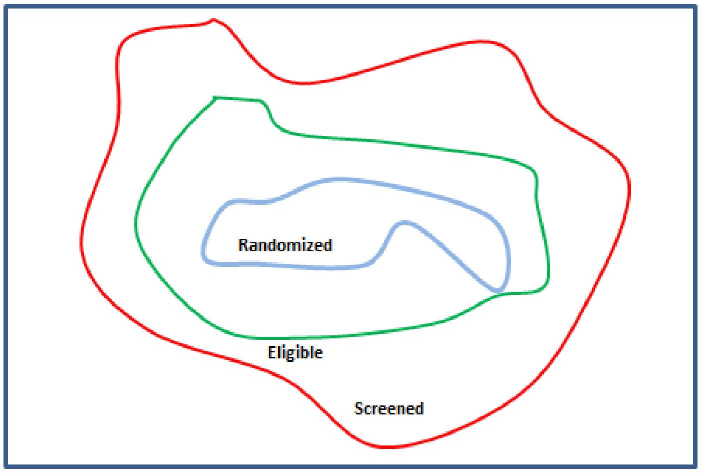


**Figure 1.** Relationship between screened, eligible, and randomized subsets when recruiting patients for a clinical trial.

>

These selections are not race, ethnicity, or gender specific. For example consider demographic breakdowns of these four subsets in a hypothetical clinical trial (Table 1).

<<Table 1 here>

Table 1 depicts the distribution of subjects in the screened, eligible, and randomized subsets of a clinical trial. Of course, the screened subset with its 1010 subjects is the largest of the subsets. There were 400 of these subjects selected as eligible, and 210 of these eligible subjects who were randomized. The numbers below represent the number of subjects that were in each of these subsets

The probabilities in the right two columns of Table 1 provide the likelihood that smaller subsets contained at most as many subjects with the trait given the findings in the screened population. For example, with respect to age, there were 303 older patients in the screened population, 140 in the eligible population, and 63 in the randomized population.

The last two columns represent computations based on hypergeometric measure. If we consider the selection of subjects to be an experiment, then in over 99% of these experiments, there would be at least 140 older subjects in the eligible subset. However, given that there are 140 older subjects in the eligible subset group, it is very unlikely that by chance alone, there would only be 63 older subjects in the randomized group reflected by the probability of 0.018. While 63 could have occurred by chance alone, the likelihood is so remote that we might suspect

[*] This is somewhat simplified. In fact there can be six distinct populations (screened, initial eligibility passed, consented, final eligibility passed, randomized, and treated)

another cause – specifically, that the process of selecting subjects from the eligible to the randomized subsets "selects out" older patients. [*]

A review of the other categories suggests no selection difficulties with Asians, but that African-Americans are also selected out when going from the eligible to the randomized population. Hispanic also appear to be susceptible to the selection process at both stages.

While this process can be illuminating when carried out after the randomization process has been completed for the trial, it is more helpful to execute this analysis during the active screening process, in order to identify and hopefully remove any selection bias issues.

∎

## Example: Assessing clinics

A clinical trial is facing the demise of its cell processing center and has identified two other cell processing facilities. Neither can handle the remaining 80 specimens to be processed (40 to be active and 40 to be placebo), so the decision is made to split the cell processing between them, each agreeing to prepare product from 40 samples. What is the likelihood that one center manufacture all placebo product and the other all active product?

Examine this from the perspective of one of the two cell processing facilities. It will receive 40 samples and their randomization assignments. There are many different configurations of active and placebo assignments of the 40 samples that it can receive. This number is exactly $\binom{80}{40}$. Of the facility's 40 assignments, if all are active, then they must come from the population of active assignments in the 80 specimens. This is reflected by $\binom{40}{40}$. There are no samples that it received from the 40 placebo assignments in the population, represented as $\binom{40}{0}$. Thus the probability that it receives all active samples is $\frac{\binom{40}{40}\binom{40}{0}}{\binom{80}{40}} = \frac{1}{\binom{80}{40}} = 9.3 \text{ x } 10^{-24}$.

Since they may also have received 40 placebo assignments, we double this to find that probability that the center receives 40 of the same assignment is $1.9 \text{ x } 10^{-23}$.

Next sections
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure

[*] It perhaps is helpful to point out that this selection is not deliberately prejudicial, but selects on characteristics that this segment of the population may have. For example, investigators may not randomize patients if they are frail and therefore likely to have more difficulty returning for future required clinic visits. While younger people may also be frail from chronic disease, the elderly are particularly susceptible to this type of selection.

Ordering Random Variables
Asymptotics
Tail Event Measure

# Moments of Hypergometric Measure

The derivation of the moments of hypergeometric measure may seem a daunting task, but we will see it is as easy to manipulate this distribution as we did the binomial distribution.

$$\mathbf{P}\left[X=m\right]=\frac{\binom{M}{m}\binom{N-n}{n-m}}{\binom{N}{n}}1_{m\in I_{[0,n]}}.$$

## Finding the mean

We begin with

$$\mathbf{E}\left[X\right]=\int_{\Omega}xd\mathbf{P}=\sum_{m=0}^{n}m\,\mathbf{P}\left[X=m\right]=\sum_{m=0}^{n}m\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

Continuing

$$=\sum_{m=1}^{n}m\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

$$=\sum_{m=0}^{n}m\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N-1}{n-1}\frac{N}{n}}=\frac{n}{N}\sum_{m=0}^{n}m\frac{\frac{M!}{m!\left(M-m\right)!}\binom{N-M}{n-m}}{\binom{N-1}{n-1}}$$

and we see that

$$= M\frac{n}{N}\sum_{m=0}^{n} m\frac{\frac{(M-1)!}{(m-1)!(M-m)!}\binom{N-M}{n-m}}{\binom{N-1}{n-1}}$$

$$= M\frac{n}{N}\sum_{m=0}^{n}\frac{\binom{M-1}{m-1}\binom{N-M}{n-m}}{\binom{N-1}{n-1}} \quad = M\frac{n}{N}$$

## Variance

To compute $\mathbf{Var}[X]$, we return to the motif that we used in the binomial distribution to first compute $\mathbf{E}[X(X-1)]$, then compute the variance as $\mathbf{E}[X(X-1)]+\mathbf{E}[X]-\mathbf{E}^2[X]$. We compute the factorial moment as

$$\mathbf{E}[X(X-1)] = \int_{\Omega} x(x-1)\,d\mathbf{P} = \sum_{m=0}^{n} m(m-1)\mathbf{P}[X=m]$$

$$= \sum_{m=0}^{n} m(m-1)\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \quad = \quad \sum_{m=2}^{n} m(m-1)\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N-2}{n-2}}$$

$$= \cfrac{\frac{M(M-1)}{N(N-1)}}{n(n-1)}\sum_{m=2}^{n}\frac{\binom{M-2}{m-2}\binom{N-M}{n-m}}{\binom{N-2}{n-2}} \quad = \quad \frac{n(n-1)M(M-1)}{N(N-1)}.$$

We may now write the variance as

$$\mathbf{Var}[X] = \mathbf{E}[X(X-1)]+\mathbf{E}[X]-\mathbf{E}^2[X]$$

$$= \frac{n(n-1)M(M-1)}{N(N-1)}+\frac{Mn}{N}-\frac{M^2n^2}{N^2}$$

$$= \frac{nM(N-M)(N-n)}{N^2(N-1)}.$$

Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure

Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Geometric and Negative Binomial Measures

We will continue our development of probability distributions based on Bernoulli trials by introducing two related distributions; the geometric, and negative binomial measure. Even though these are distributions based on Bernoulli trials, their sample spaces have a new property-the possibility of not just a finite, but an infinite number of values.

## Prerequisites

The Limit Concept
Properties of Probability
Conditional Probability
An Introduction to the Concept of Measure
Elementary Set Theory
Measure and its Properties
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Working with Measure
Properties of Probability
Generating Functions

## Geometric measure

Consider the following scenario. Patients arrive at a clinic independently of each other at a clinic that provides immunizations. It is not uncommon for a patient to require an immunization (tetanus toxoid) for tetanus, an event that occurs with probability *p*. Can we compute the probability that *k* patients are seen before a patient requires the first tetanus shot of the day?

This problem has elements that are familiar to us. The arrival of independent patients, each with the same probability *p* of "success" (in this case, the need for a tetanus shot) is certainly a collection of Bernoulli trials.

However, here is where the familiarity ends. We are accustomed to computing the probability of a great many events around the number of patients who require a tetanus shot (3 patients, at least 5 patients, no more than 2 patients, etc.). However, here we are told that only one patient requires a tetanus shot. We must compute the probability of a number of patients who are seen, when we know the actual number who receive a tetanus shot.

The sample space is altered here. For the Bernoulli and binomial distributions, the number of trials is fixed and we have to compute probabilities of a finite number of successes occurring within those trials. However, in this case, the number of "successes" is fixed, and we must compute the number of trials.

It is the number of trials that is the random variable.

The probabilities are not in and of themselves difficult to compute. However, there is an important new consideration. Let $k$ be the number of subjects who do not require tetanus-toxoid. Then $\mathbf{P}[k=0]=p$ since the only way for $k$ to be zero is for the first patient to require the toxoid. If the second patient is the first to receive toxoid, then we compute $\mathbf{P}[k=2]=qp,$ acknowledging that the first patient did not require toxoid while the second patient did. We can now compute the following sequence of probabilities

$$\begin{aligned}\mathbf{P}[k=0]&=p\\ \mathbf{P}[k=1]&=qp\\ \mathbf{P}[k=2]&=q^2p\\ \mathbf{P}[k=3]&=q^3p\end{aligned}$$

And we easily see that $\mathbf{P}[X=k]=q^k p\mathbf{1}_{k\in\mathrm{I}[0,\infty]}.$[*] However, the indicator presents a potential problem because it signals that we have to sum over all of the nonnegative integers. Formally,

$$\int_{\Omega} d\mathbf{P}=\int q^k p\mathbf{1}_{k\in\mathrm{I}[0,\infty]}=\sum_{k=0}^{\infty}q^k p=1\,?$$

How can an infinite sum of positive numbers be less than infinity since we can never be able to stop counting them? And how can it be equal to one?

It is clear that for many summands, the infinite sum will definitely be infinity. However, must this be the case of all of them? Say that $p=\frac{1}{2}.$ Then $\sum_{k=0}^{\infty}q^k p=\frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k$ . Is possible that $\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k$ is finite?

## Infinite sums as sequences

Let's first take the infinite sum and convert it into an infinite sequence. Define

$$S_k=1+\frac{1}{2}+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^3+\left(\frac{1}{2}\right)^4+\ldots+\left(\frac{1}{2}\right)^k=\sum_{j=0}^{k}\left(\frac{1}{2}\right)^j.$$

Then our infinite sum is a sequence progression, $S_1$, $S_2$, $S_3$, …$S_k$…, each term in the sequence representing one more summand added to the growing sum. Then, the question of whether $\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k$ is finite is converted to the phenomenon of the potential convergence of the sequence $\{S_k\}$. If it converges to some value $L$, then we can say $\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k=L.$

[*] Since there is only one sequence that satisfies the event, namely that the success must occur at the end of the sequence, we do not have to count combinatorics as we did for the binomial distribution.

Examine the following

$$S_k = 1 + \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 + \ldots + \left(\frac{1}{2}\right)^k$$

Then multiply each side by $\frac{1}{2}$ to find

$$\frac{1}{2}S_k = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 + \ldots + \left(\frac{1}{2}\right)^{k+1}$$

Subtracting the second from the first equation above reveals

$$\left(\frac{1}{2}\right)S_k = 1 - \left(\frac{1}{2}\right)^{k+1} \quad \text{or} \quad S_k = \frac{1-\left(\frac{1}{2}\right)^{k+1}}{\left(\frac{1}{2}\right)}.$$

So we must find

$$\lim_{k\to\infty} S_k = \lim_{k\to\infty} \frac{1-\left(\frac{1}{2}\right)^{k+1}}{\left(\frac{1}{2}\right)} = \lim_{k\to\infty} 2\left(1-\left(\frac{1}{2}\right)^{k+1}\right).$$

It is here were our work with limits and continuity pays a handsome reward. We know that the function $2\left(1-\left(\frac{1}{2}\right)^{k+1}\right)$ is continuous.* We also know from our discussion on the limits of continuous functions, that the limit passes through the function. Thus we write

$$\lim_{k\to\infty} S_k = \sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k = \lim_{k\to\infty}\left[2\left(1-\left(\frac{1}{2}\right)^{k+1}\right)\right]$$
$$= 2\left(1-\lim_{k\to\infty}\left(\frac{1}{2}\right)^{k+1}\right) = 2(1-0) = 2.$$

Therefore $\sum_{k=0}^{\infty} q^k p = \frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k = \left(\frac{1}{2}\right)(2) = 1.$

In fact, the probability that there are *k* failures before the first success, when $p = \frac{1}{2}$ and written as $\mathbf{P}[k] = \left(\frac{1}{2}\right)^k \frac{1}{2}\mathbf{1}_{k\in I[0,\infty]} = \left(\frac{1}{2}\right)^{k+1}\mathbf{1}_{k\in I[0,\infty]}$ is a probability measure.

We write now a concise proof that for any probability of success *p*, such that $0 < p < 1$, the $\mathbf{P}[k] = q^k p\mathbf{1}_{k\in[0,\infty]}$ has measure one over the non-negative integers.

* It is not too challenging an exercise to use an $\varepsilon - \delta$ argument to demonstate the continuity of $2\left(1-\left(\frac{1}{2}\right)^{k+1}\right)$.

$$\int_{\Omega} d\mathbf{P} = \int_{\Omega} q^k p \mathbf{1}_{k\in \mathrm{I}[0,\infty]} = \sum_{k=0}^{\infty} q^k p .$$

$$S_k = 1 + q + q^2 + q^3 + q^4 + \ldots + q^k$$

$$qS_k = q + q^2 + q^3 + q^4 + q^5 + \ldots + q^{k+1}$$

$$(1-q) S_k = 1 - q^{k+1}$$

Continuing

$$S_k = \frac{1-q^{k+1}}{(1-q)} : \quad \lim_{k\to\infty} S_k = \lim_{k\to\infty}\left(\frac{1-q^{k+1}}{(1-q)}\right) = \frac{1}{1-q} = \frac{1}{p}$$

$$\int_{\Omega} d\mathbf{P} = \int q^k p \mathbf{1}_{k\in \mathrm{I}[0,\infty]} = \sum_{k=0}^{\infty} q^k p = p\sum_{k=0}^{\infty} q^k = p\left(\frac{1}{p}\right) = 1.$$

Depending on the value of $q$, the probability of $k$ failures before the first success can congregate early or late (Figure 1).

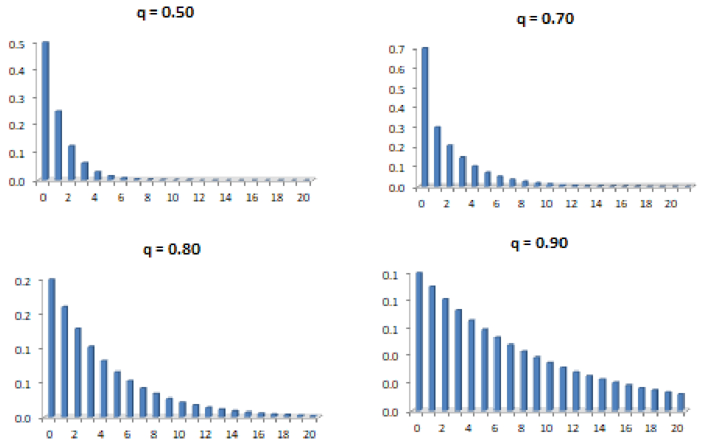


**Figure 1**. Examples of the geometric distribution.

Another parameterization of the geometric distribution is to alter it to find, not just the number of failures before the first success, but the probability that the $k^{\text{th}}$ trial holds the first success. This is just the probability that there are $k-1$ failures before the first success. Defining $W$ as the random variable here, we write this as $\mathbf{P}[W=k] = q^{k-1} p \mathbf{1}_{k\in \mathrm{I}[1,\infty]}$. Note now that the lower bound of $k$ is one. To see that the measure over the entire sample space is equal to one we write

$$\int_{\Omega} d\mathbf{P} = \int q^{k-1} p \mathbf{1}_{k\in \mathrm{I}[1,\infty]} = \sum_{k=1}^{\infty} q^{k-1} p \;\; = p\sum_{k=0}^{\infty} q^k = p\left(\frac{1}{p}\right) = 1$$

## Probability generating function

We can compute the probability generating function of *X*, $\mathbf{E}\left[s^x\right]$ as

$$\mathbf{G}_X(s) = \mathbf{E}\left[s^x\right] = \int_\Omega s^x d\mathbf{P} = \int s^k q^k p \mathbf{1}_{k\in\mathrm{I}[0,\infty]} = \sum_{k=0}^{\infty} p\left(qs\right)^k = \frac{p}{1-qs}.$$

Comparing this to the probability generating function of the negative binomial measure to be developed later in this chapter will demonstrate an interesting relationship between these two probability functions.

### *Example: Conditional negative binomial measure*

Consider the situation when you know there are $k_2$ failures before $r_2$ successes.What is the probabilty that there were *k* failures before *r* successes where $0 \le k \le k_2$ and $0 \le r \le r_2$?

This problem providesa trajectory for the process. Once can identify the probability distribution of interim points in the negative binomial process. What we wish to find is

$\mathbf{P}\left[X_{p,r} = k \,|\, X_{p,r_2} = k_2\right]$ where $X_{p,r}$ is a negarive binomial random variable with parameters *p* and *r*. We begin by writing

$$\mathbf{P}\left[X_{p,r} = k \,|\, X_{p,r_2} = k_2\right] = \frac{\mathbf{P}\left[X_{p,r} = k \cap X_{p,r_2} = k_2\right]}{\mathbf{P}\left[X_{p,r_2} = k_2\right]}$$

Note that
$$\begin{aligned}\mathbf{P}\left[X_{p,r} = k \cap X_{p,r_2} = k_2\right] &= \mathbf{P}\left[X_{p,r} = k \cap X_{p,r_2-r} = k_2 - k\right] \\ &= \mathbf{P}\left[X_{p,r} = k\right]\mathbf{P}\left[X_{p,r_2-r} = k_2 - k\right].\end{aligned}$$

This is the key to the problems solution. Given that there are k failures bdefore the r<sup>th</sup> success, there must be $k_2 - k$ failures. Before another $r_2 - r$ successes. The recognition that these are Bernouli trials justifies taking the product of the joint probablities. Thus,

## Moments of geometric measure

We begin with the expected value *X*, the number of failures before the first success.

$$\mathbf{E}\left[X\right] = \int_\Omega x d\mathbf{P} = \int kq^k p \mathbf{1}_{k\in\mathrm{I}[0,\infty]} = \sum_{k=0}^{\infty} kq^k p = p\sum_{k=0}^{\infty} kq^k$$

Here we can use what we learned from the Using the generating function approach to infinite series. We want to find what function $\mathbf{G} = \sum_{k=0}^{\infty} kq^k$. Recall that taking derivatives of both sides of the geometric series and multiplying both sides by $q$ that we found that $\sum_{k=0}^{\infty} kq^k = \frac{q}{\left(1-q\right)^2} = \frac{q}{p^2}$.

Thus, we have $\mathbf{E}\left[X\right]=p\frac{q}{p^2}=\frac{q}{p}$. This inversion property of many infinite series will be quite helpful as we develop this followed by negative binomial measure.

In order to compute the variance of the geometric disruption, we write

$$\mathbf{E}\left[X^2\right]=\int_\Omega x^2 d\mathbf{P}=\int k^2q^kp\mathbf{1}_{k\in\mathrm{I}[0,\infty]}=\sum_{k=0}^{\infty}k^2q^kp=p\sum_{k=0}^{\infty}k^2q^k,$$

The generating function approach permits us to write that

$\mathbf{G}(s)=\frac{s(s+1)}{(1-s)^3}\rhd\left\{k^2\right\}$, and we compute $\mathbf{E}\left[X^2\right]=p\sum_{k=0}^{\infty}k^2q^k=p\frac{q(q+1)}{p^3}=\frac{q}{p^2}(q+1)$. Thus

$$\begin{aligned}\mathbf{Var}\left[X\right]&=\mathbf{E}\left[X^2\right]-\mathbf{E}^2\left[X\right]\\&=\frac{q}{p^2}(q+1)-\frac{q^2}{p^2}\\&=\frac{q}{p^2}.\end{aligned}$$

## The probability of the second success

Suppose now we asked what is the probability that the $k^{\text{th}}$ trial contains the second success? Let the relevant random variable be $Y$, noting that $Y\geq 2$. Here there are many possible combinations to consider for the occurrence of the second success. In fact for each increase in the value of $Y$, the number of possibilities for the location of the first success increases. Tracking it this way becomes unhelpful.

However, another way to think of this event is that we really don't care when the first success occurred as long as it occurred before the second, and that the second occurred on the last trial. But this is the probability that there is one success in $n-1$ trials, followed by the $n^{\text{th}}$ trial that is itself a success. Since Bernoulli trials are independent we write

$$\mathbf{P}\left[Y=n\right]=\binom{n-1}{1}pq^{n-2}p\mathbf{1}_{n\in\mathrm{I}[2,\infty]}=\binom{n-1}{1}p^2q^{n-2}\mathbf{1}_{n\in\mathrm{I}[2,\infty]}$$

Another way to write this equation is so that the random variable is not the number of trials, but the number of failures before the second success. One advantage of this is that this random variable always has positive probability assigned on each of the nonnegative integers, as opposed to $Y$ above. Then, $n=k+2$, $k=n-2$, and rewrite as

$$\mathbf{P}\left[V=k\right]=\binom{k+1}{1}p^2q^k\mathbf{1}_{k\in\mathrm{I}[0,\infty]}$$

This is an example of a negative binomial random variable. In general, the measure of the number of failures before the $r^{\text{th}}$ success is

$$\mathbf{P}\left[V=k\right]=\binom{k+r-1}{r-1}p^r q^k \mathbf{1}_{k\in\mathrm{I}[0,\infty]}.$$

## Why “negative” binomial measure?

This name, negative binomial, comes from the use of the binomial theorem for negative integers.[*]

$$\binom{k+r-1}{r-1}=\binom{k+r-1}{k}=\left(-1\right)^k\binom{-r}{k}.$$

However, since negative binomial measure is rarely written in this format, the sobriquet “negative” has little meaning for us. If we develop negative binomial measure from a generating function approach, then, again, its name is not particularly illuminating.

We might be better off thinking of this distribution not in terms of its name, but instead what its function does. While the binomial distribution focuses on fixing the number of trials, and computing the distribution of the number of successes within those trials, negative binomial measure fixes the number of successes and computes the distribution of the number of failures before the last success occurs on the last trial.

Use of the generating function approach quickly reveals that the proposed negative binomial measure integrates to one over the entire sample space.

$$\int_{\Omega} d\mathbf{P}=\int\binom{k+r-1}{r-1}p^r q^k \mathbf{1}_{k\in\mathrm{I}[0,\infty]}.=\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}p^r q^k$$

$$=p^r\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}q^k=p^r\left(\frac{1}{\left(1-q\right)^r}\right)=1.$$

We know that $\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}q^k=\left(\frac{1}{\left(1-q\right)^r}\right)$ through the use of our work on infinite series.

Negative binomial measure has a variety of shapes depending on the parameter values (Figure 2).

[*] $\binom{-n}{k}=\left(-1\right)^k\frac{\left(-r\right)\left(-r-1\right)\left(-r-2\right)\left(-r-3\right)...\left(-r-n-1\right)}{k!}.$

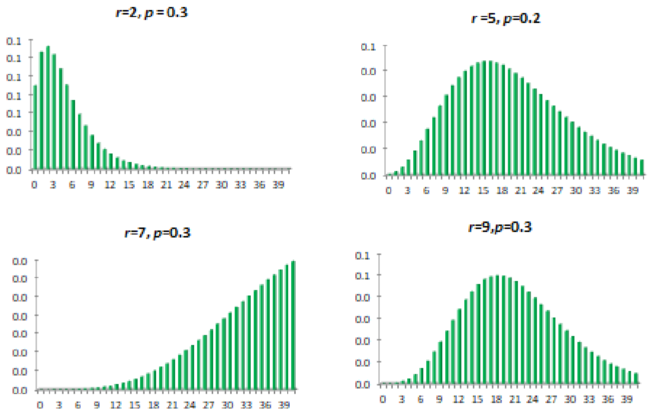


**Figure 2**. Examples of the negative binomial distribution.

## Negative binomial measure moments

It is a worthwhile using both the moment generating function and the combinatoric approaches to compute the mean and variance of the Negative binomial measures. We find that the mean and variance of $Y$, the number of failures before the $r^{\text{th}}$ success is $\frac{rq}{p}$ and $\frac{rq}{p^2}$ respectively.

### *Example: Surveying Pregnant Women*:

A researcher is interested in surveying a community's population of pregnant young women (between 15-30 years of age). From census records, they know the probability a pregnant women lives in any particular domicile is 0.08. What is the probability that at least 800 domiciles must be contacted to reach 75 pregnant women?

Here the probability of a success is 0.08, and $r = 75$. We compute

$$\mathbf{P}\left[V \geq 800\right] = \sum_{k=800}^{\infty} \binom{k+75-1}{75-1} p^{75} q^{k}$$

$$= \sum_{k=800}^{\infty} \binom{k+74}{74} p^{75} q^{k} = 0.72.$$

■

### *Example: Screening subjects for a clinical trial*.

In order to ensure that clinical trial have the best chance of determining the effectiveness of an intervention (commonly termed efficacy) safely, very specific patient populations must be identified. As we have seen in another example, these populations are sometimes described as screened populations, consented populations, and randomized populations.

The randomized subjects are selected from the consented population, and consented subjects are selected from the screen populations. Assume that there is a 75% chance that any consented subjects will be randomized. The sample size goal (i.e., total number of randomized subjects required) is 87. We know the minimum number of consented subjects needed to randomize 87 subjects is 87, What is the number of screened subjects required in order to have a 90% chance that we will be able to select 87 randomized subjects.

Consented subjects can be divided into randomized (success) and nonrandomized (failure subjects). We can compute for any number *n* of consented subjects, the probability that exactly $T_c = n$ consented subjects are required to randomize 87 subjects as

$$\mathbf{P}\left[T_c = n\right] = \binom{n-1}{86}(0.75)^{87}(0.25)^{n-87}.$$

We want the probability that at most *m* subjects are required so that the probability that needing that many consented subjects or less to randomize 87 subjects is 0.90. That is we must find *m* such that

$$\mathbf{P}\left[T_c \le m_c\right] = \sum_{n=87}^{m_c}\binom{n-1}{86}(0.75)^{87}(0.25)^{n-87}.$$

We find that $m_c = 124$.

Now how many subjects must undergo screening in order to have a 90% chance of, from these subjects, identifying 124 screened, consented subjects. In this situation, the probability that a patient is successfully screened and consented is 0.13. Here we compute

$$\mathbf{P}\left[T_s \le m_s\right] = \sum_{n=124}^{m_s}\binom{n-1}{123}(0.13)^{124}(0.87)^{n-123} = 0.90.$$

and we compute $m_s = 1058$. The screening population must be much larger than the consented population because 1) we require more consented subjects than randomized subjects and 2) the probability of successfully consenting a screened patient is much lower than the probability of randomizing a consented subjects.

### *Example: Conditional negative binomial measure*

Consider the situation when you know there are $k_2$ failures before $r_2$ successes.What is the probabilty that there were *k* failures before *r* successes where $0 \le k \le k_2$ and $0 \le r \le r_2$?

This problem providesa trajectory for the process. Once can identify the probability distribution of interim points in the negative binomial process. What we wish to find is

$\mathbf{P}\left[X_{p,r} = k \mid X_{p,r_2} = k_2\right]$ where $X_{p,r}$ is a negarive binomial random variable with parameters *p* and *r*. We begin by writing

$$\mathbf{P}\left[X_{p,r} = k \mid X_{p,r_2} = k_2\right] = \frac{\mathbf{P}\left[X_{p,r} = k \cap X_{p,r_2} = k_2\right]}{\mathbf{P}\left[X_{p,r_2} = k_2\right]}$$

Note that

$$\begin{aligned}\mathbf{P}\left[X_{p,r}=k\cap X_{p,r_2}=k_2\ \right]&=\mathbf{P}\left[X_{p,r}=k\cap X_{p,r_2-r}=k_2-k\ \right]\\&=\mathbf{P}\left[X_{p,r}=k\ \right]\mathbf{P}\left[X_{p,r_2-r}=k_2-k\ \right].\end{aligned}$$

This is the key to the problems solution. Given that there are k failures bdefore the r^th success, there must be $k_2-k$ failures. Before another $r_2-r$ successes. The recognition that these are Bernouli trials justifies taking the product of the joint probablities. Thus,

$$\mathbf{P}\left[X_{p,r}=k\,|\,X_{p,r_2}=k_2\ \right]=\frac{\mathbf{P}\left[X_{p,r}=k\cap X_{p,r_2}=k_2\ \right]}{\mathbf{P}\left[X_{p,r_2}=k_2\ \right]}$$

We now need to simply to expand the negative binomial probabilities nd then the terms involving $p$ to see that

$$=\frac{\mathbf{P}\left[X_{p,r}=k\ \right]\mathbf{P}\left[X_{p,r_2-r}=k_2-k\ \right]}{\mathbf{P}\left[X_{p,r_2}=k_2\ \right]}$$

$$=\frac{\binom{k+r-1}{r-1}p^r(1-p)^k\binom{k_2-k+r_2-r-1}{r_2-r-1}p^{r_2-r}(1-p)^{k_2-k}}{\binom{k_2+r_2-1}{r_2-1}p^{r_2}(1-p)^{k_2}}$$

$$=\frac{\binom{k+r-1}{r-1}\binom{k_2-k+r_2-r-1}{r_2-r-1}}{\binom{k_2+r_2-1}{r_2-1}}$$

Which is similar to a hypergeometric probability. Letting $N=k_2+r_2-1;\ n=r_2-1;\ M=k+r-1$ and $m=r-1,$ and we need to write

$$\binom{k_2-k+r_2-r-1}{r_2-r-1}=\frac{(k_2-k+r_2-r-1)!}{(k_2-k)!(r_2-r-1)!}=\frac{(k_2-k+r_2-r-1)(k_2-k+r_2-r-2)!}{(k_2-k)!(r_2-r-1)(r_2-r-2)!}$$

$$=\frac{(k_2-k+r_2-r-1)}{(r_2-r-1)}\frac{(k_2-k+r_2-r-2)!}{(k_2-k)!(r_2-r-1)(r_2-r-2)!}=\frac{(k_2-k+r_2-r-1)}{(r_2-r-1)}\binom{k_2-k+r_2-r-2}{r_2-r-2}.$$

Therefore

$$\mathbf{P}\left[X_{p,r}=k \mid X_{p,r_2}=k_2\ \right]$$
$$=\frac{\left(r_2-r-1\right)}{\left(k_2-k+r_2-r-1\right)}\frac{\binom{k+r-1}{r-1}\binom{k_2-k+r_2-r-2}{r_2-r-2}}{\binom{k_2+r_2-1}{r_2-1}}$$

Which is a constant times a hypergeometric random variable.

### *Moment analysis of contagion process*

The contagion process is useful for following the growth of pandemics such as the SARS-CoV-22. However, the evaluation of the mean and variance while having a satisfying history of utility may not be completely sufficient to identify how fast the pandemic is growing. Specifically the pandemic expected vale, while it is an exponential growth. may underestimate the growth of the infectiin in the community. Therefore, looking at other moments of the contagion model may provide more helpful predictive solutions. Here we will just examine one such alternative model.

We recall from the previous section that the contagion process results in anegativre binomial distribution, providing $\mathbf{P}_t\left[k\right]$, the probability that there are *k* patients in the system at time *t*. In the case of the contagion process, this is

$$\mathbf{P}_t\left[k\right]=\binom{k+r-1}{r-1}p^r q^k,$$

Where $r=a+\frac{\lambda}{\upsilon}, p=e^{-\upsilon t}$. In this case, *a* is the number of patients in the system at time $t=0$, $\lambda$ is the arrival rate, and $\upsilon$ is the rate of spread of the disease from patient to patient. We know $\mathbf{E}_t\left[k\right]=\frac{rp}{q}$. The new question for us is what is $\mathbf{E}_t\left[k^2\left(k-1\right)\right]$? We can use a generating function argument, but let's first write

$$\mathbf{E}_t\left[k^2\left(k-1\right)\right]=\sum_{k=0}^{\infty}k^2\left(k-1\right)\binom{k+r-1}{r-1}p^r q^k$$
$$=\sum_{k=2}^{\infty}k^2\left(k-1\right)\binom{k+r-1}{r-1}p^r q^k.$$

Let's use a generating approach. We know that $\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}q^k=\frac{1}{\left(1-q\right)^r}$. We have to simply convert this generating function to the generating function involving $k^2\left(k-1\right)$. Begin by taking a derivative with respect to *q* of each side.

$$\sum_{k=0}^{\infty}k\binom{k+r-1}{r-1}q^{k-1}=\frac{r}{\left(1-q\right)^{r+1}}.$$ . Multiply through by *q*

$\sum_{k=0}^{\infty} k \binom{k+r-1}{r-1} q^k = \frac{rq}{(1-q)^{r+1}}$. Another derivative with respect to *q* returns

$$\sum_{k=0}^{\infty} k^2 \binom{k+r-1}{r-1} q^{k-1} = \frac{d}{dq}\left(\frac{rq}{(1-q)^{r+1}}\right) = \frac{r}{(1-q)^{r+1}} + \frac{r(r+1)q}{(1-q)^{r+2}}$$

$$= \frac{r(1-q)+r(r+1)q}{(1-q)^{r+2}}.$$

An addition derivative reveals

$$\sum_{k=0}^{\infty} k^2 k - 1 \binom{k+r-1}{r-1} q^{k-2} = \frac{d}{dq}\left(\frac{r(1-q)+r(r+1)q}{(1-q)^{r+2}}\right)$$

$$= \frac{\left(r(r+1)q\right)(r+2)}{(1-q)^{r+3}}.$$

Multiply by $q^2$ to find

$$\sum_{k=0}^{\infty} k^2 (k-1) \binom{k+r-1}{r-1} q^k = \frac{q^2 \left(r(r+1)q\right)(r+2)}{(1-q)^{r+3}}.$$

Thus

$$\mathbf{E}_t\left[k^2(k-1)\right] = \sum_{k=0}^{\infty} k^2 (k-1) \binom{k+r-1}{r-1} p^r q^k$$

$$= p^r \frac{q^3 r(r+1)(r+3)}{(1-q)^{r+3}} = \left(\frac{q}{p}\right)^3 r(r+1)(r+3).$$

## Generating moments

We can identify the probability generating function $\mathbf{E}\left[s^v\right]$ at once for the negative binomial measure;

$$\mathbf{E}\left[s^v\right]=\int_\Omega s^v d\mathbf{P}=\int s^k\binom{k+r-1}{r-1}p^r q^k \mathbf{1}_{k\in \mathrm{I}[0,\infty]}=\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}p^r(qs)^k$$

$$=p^r\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(qs)^k=\left(\frac{p}{1-qs}\right)^r.$$

The last statement in the argument above $\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(qs)^k=\frac{1}{(1-qs)^r}$ is based on a generating function argument in which derivatives are taken. So, for negative binomial measure, $\mathbf{G}_V(s)=\left(\frac{p}{1-qs}\right)^r$. Recall for geometric measure $\mathbf{G}_X(s)=\frac{p}{1-qs}$, demonstrating that for these two distributions $\mathbf{G}_V(s)=\left(\mathbf{G}_X(s)\right)^r$.

We have seen this type of relationship before, and recall that, if one generating function is the product of another, then the random variable for the first is the sum of the random variables composed from the summands. Thus, the negative binomial random variable is the sum of $r$ i.i.d. geometric random variables each of which has the same probability of success $p$.

We can use this generating function approach to consider another more complicated situation in clinical trial recruitment. Consider another screening problem from clinical trials. Two centers are recruiting subjects for a clinical trial. Let the number of subjects screened from the first clinical trial be denoted as $X_1$ and from the second clinic as $X_2$. If each of $X_1$ and $X_2$ follow negative binomial measure, and are independent of each other, under what circumstances can we manage the distribution of the total number of subjects screened $W=X_1+X_2$?

If we assume that the probability of a screened patient being randomized $p$ is the same for both $X_1$ and $X_2$ then we can use a simple generating function approach to compute the solution. Since we know $G_{X_1}(s)=G_{X_2}(s)=\left(\frac{p}{1-qs}\right)^r$, we write $G_W(s)=G_{X_1}(s)G_{X_2}(s)=\left(\frac{p}{1-qs}\right)^{2r}$ which is just the probability generating function of the negative binomial measure with parameters $p$ and $2r$, i.e., we write

$$\mathbf{P}\left[W=k\right]=\binom{k+2r-1}{2r-1}p^{2r}q^k\mathbf{1}_{k\in \mathrm{I}[0,\infty]}.$$

This result should come as no surprise since we were able to build up the negative binomial measure from the sum of geometric random variables.

If we were to now alter the parameters of the distributions for centers 1 and 2, we can still compute the distribution of $W$ with ease. For example, if each of $X_1$ and $X_2$ continue to have their same probability of success $p$ but different number of randomized subjects $r_1$ and $r_2$, then we can follow the previous development to write $G_W(s)=G_{X_1}(s)G_{X_2}(s)=\left(\frac{p}{1-qs}\right)^{r_1+r_2}$ and

$$\mathbf{P}\left[W=k\right]=\binom{k+r_1+r_2-1}{r_1+r_2-1}p^{r_1+r_2}q^k\mathbf{1}_{k\in \mathrm{I}[0,\infty]}.$$

Matters become more complicated when we consider different probabilities of success $p_1$ and $p_2$. In this case $G_W(s) = G_{X_1}(s)G_{X_2}(s) = \left(\frac{p_1}{1-q_1 s}\right)^{r_1}\left(\frac{p_2}{1-q_2 s}\right)^{r_2}$, and we do not see the easy and natural simplification to which we have become accustomed.

However, the inversion of this product is straightforward and we write for $r_1 = r_2 = 1$

$$\mathbf{P}[W = k] = p_1 p_2 \sum_{m=0}^{k} q_1^m q_2^{k-m}.$$

### *Pulling the neg binomial distribution "out of a hat"*

As an aside, we can commonly identify an inverted generation function as related to the negative binomial distribution. Lets start with a generating function, $\mathbf{G}_t(s) = \frac{a}{b-cs}$. Write this as $\frac{a}{b-cs} = a\frac{1}{b-cs} = \frac{a}{b}\frac{1}{1-\frac{c}{b}}s.$ If $0 < c < b,$ then we might consider $\frac{c}{b}$ a probability. Then write

$\mathbf{G}_t(s) = \frac{a}{b}\frac{b}{b-c}\left(\frac{1-\frac{c}{b}}{1-\frac{c}{b}s}\right) = \frac{a}{b-c}\left(\frac{q}{1-ps}\right).$ Then the coefficient $s^k$ is $\frac{a}{b-c}$ times the probability of a geometric rsndom variable with paramer $p = \frac{c}{b}$.Shou$\mathbf{G}_t(s) = \left(\frac{a}{b-cs}\right)^r$, then we write

$\mathbf{G}_t(s) = \left(\frac{a}{b-c}\right)^r\left(\frac{q}{1-ps}\right)^r$, then the coefficient $s^k$ is $\left(\frac{a}{b-c}\right)^r$ times the probability of a negative binomial random variable with parameters *r* and $p = \frac{c}{b}$. We will use this when modeling the contagion processes.

## Other sections of interest

General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Moments of Negative Binomial Measure

Prerequisites
Negative binomial measure
Generating functions

## Generating function approach

A hint of the power of the generating function approach is revealed in the identification of the moments of the generating function. We wish to find the mean and variance of the negative binomial measure, written as

$$\mathbf{P}[V=k]=\binom{k+r-1}{r-1}p^r q^k \mathbf{1}_{k\in \mathrm{I}[0,\infty]}$$

We begin by writing

$$\mathbf{E}[V]=\int_{\Omega} v\,d\mathbf{P}=\int k\binom{k+r-1}{r-1}p^r q^k \mathbf{1}_{k\in \mathrm{I}[0,\infty]}.=\sum_{k=0}^{\infty}k\binom{k+r-1}{r-1}p^r q^k.$$

However, we simply write $\sum_{k=0}^{\infty}k\binom{k+r-1}{r-1}p^r q^k = p^r\sum_{k=0}^{\infty}k\binom{k+r-1}{r-1}q^k$,

and our  task reduces to inverting $\sum_{k=0}^{\infty}k\binom{k+r-1}{r-1}q^k$. But we from our work in generating functions that $\mathbf{G}(s)=\frac{1}{(1-s)^r}\triangleright\left\{\binom{k+r-1}{r-1}\right\}s^k$.[*] We simply take a derivative of both sides to see that

$$\mathbf{G}(s)=\frac{r}{(1-s)^{r+1}}\triangleright\left\{k\binom{k+r-1}{r-1}\right\}s^{k-1}.$$

and multiply each side by $s$, observing

[*] The $s$ term outside the braces simply signifies that the term included in the braces is the coefficient of $s^k$. Its convenience will be conveyed shortly.

$$\mathbf{G}(s) = \frac{rs}{\left(1-s\right)^{r+1}} \rhd \left\{ k\binom{k+r-1}{r-1} \right\} s^k.$$

Thus, $\sum_{k=0}^{\infty} k\binom{k+r-1}{r-1} q^k = \frac{rq}{\left(1-q\right)^{r+1}}$, and we can finish the expected value computation.

$$\begin{aligned}\mathbf{E}\left[V\right] &= \int_{\Omega} v\, d\mathbf{P} = \int k\binom{k+r-1}{r-1} p^r q^k \mathbf{1}_{k\in \mathrm{I}[0,\infty]\cdot} = \sum_{k=0}^{\infty} k\binom{k+r-1}{r-1} p^r q^k \\ &= p^r \sum_{k=0}^{\infty} k\binom{k+r-1}{r-1} q^k = p^r \frac{rq}{p^{r+1}} = \frac{rq}{p}.\end{aligned}$$

To compute the variance, we need $\mathbf{E}\left[V^2\right]$. We can follow the same process to compute this quantity.

$$\begin{aligned}\mathbf{E}\left[V^2\right] &= \int_{\Omega} v^2\, d\mathbf{P} = \int k^2 \binom{k+r-1}{r-1} p^r q^k \mathbf{1}_{k\in \mathrm{I}[0,\infty]} \\ &= p^r \sum_{k=0}^{\infty} k^2 \binom{k+r-1}{r-1} q^k.\end{aligned}$$

From our work to compute the first moment, we remind ourselves that

$\mathbf{G}(s) = \frac{rs}{\left(1-s\right)^{r+1}} \rhd \left\{ k\binom{k+r-1}{r-1} \right\} s^k$. We then take an additional derivative

$$\mathbf{G}(s) = \frac{\left(r+1\right)rs}{\left(1-s\right)^{r+2}} + \frac{r}{\left(1-s\right)^{r+1}} = \frac{r^2 s + r}{\left(1-s\right)^{r+2}} \rhd \left\{ k^2 \binom{k+r-1}{r-1} \right\} s^{k-1}$$

We now multiply each side by $s$ to find

$$\mathbf{G}(s) = \frac{r^2 s^2 + rs}{\left(1-s\right)^{r+2}} \rhd \left\{ k^2 \binom{k+r-1}{r-1} \right\} s^k$$

And we compute

$$\begin{aligned}\mathbf{E}\left[V^2\right] &= \int_{\Omega} v^2\, d\mathbf{P} = \int k^2 \binom{k+r-1}{r-1} p^r q^k 1_{k\in \mathrm{I}[0,\infty]} \\ &= p^r \sum_{k=0}^{\infty} k^2 \binom{k+r-1}{r-1} q^k = p^r \frac{r^2 q^2 + rq}{\left(1-q\right)^{r+2}} = \frac{r^2 q^2 + rq}{p^2}.\end{aligned}$$

Thus,

$$\mathbf{Var}\left[V\right]=\mathbf{E}\left[V^2\right]-\mathbf{E}^2\left[V\right]$$
$$=\frac{r^2q^2+rq}{p^2}-\frac{r^2q^2}{p^2}$$
$$=\frac{rq}{p^2}.$$

## Without generating functions

To compute the $\mathbf{E}\left[V\right]$, we begin by writing

$$\mathbf{E}\left[V\right]=\sum_{k=0}^{\infty}k\binom{k+r-1}{r-1}p^rq^k$$
$$=\sum_{k=0}^{\infty}k\frac{(k+r-1)!}{k!(r-1)!}p^rq^k=r\sum_{k=0}^{\infty}\frac{(k+r-1)!}{(k-1)!r!}p^rq^k$$

Continuing

$$=r\sum_{k=0}^{\infty}\frac{(k+r-1)!}{(k-1)!r!}p^rq^k$$
$$=\frac{r}{p}\sum_{k=0}^{\infty}\frac{(k+r-1)!}{(k-1)!r!}p^{r+1}q^k=r\frac{q}{p}\sum_{k=0}^{\infty}\binom{k+r-1}{r}p^{r-1}q^{k-1}$$

We recognize $\binom{k+r-1}{r}p^{r-1}q^{k-1}$ as the probability that there are $k$ failures before the $r$+1st success. The sum of this mass function over the non-negative reals is 1, giving us our result that $\mathbf{E}\left[V\right]=\frac{rq}{p}$.

To compute the $\mathbf{Var}\left[V\right]$, we return to the computation of the factorial moment, $\mathbf{E}\left[V(V-1)\right]$. Write

$$\mathbf{E}\left[V(V-1)\right]=\int_{\Omega}v(v-1)\,d\mathbf{P}=\int k(k-1)\binom{k+r-1}{r-1}p^rq^k\mathbf{1}_{k\in \mathrm{I}[0,\infty]}.$$
$$=\sum_{k=0}^{\infty}k(k-1)\binom{k+r-1}{r-1}p^rq^k.$$

And we proceed as before

$$\mathbf{E}\left[V(V-1)\right]=\sum_{k=0}^{\infty}k(k-1)\binom{k+r-1}{r-1}p^rq^k$$
$$=\sum_{k=0}^{\infty}k(k-1)\frac{(k+r-1)!}{k(k-1)(k-2)!(r-1)!}p^rq^k$$
$$=(r+1)r\sum_{k=0}^{\infty}\frac{(k+r-1)!}{(k-2)!(r+1)!}p^rq^k$$

$$
\begin{aligned}
&=(r+1)r\sum_{k=0}^{\infty}\binom{k+r-1}{r+1}p^{r}q^{k}\\
&=(r+1)r\frac{q^2}{p^2}\sum_{k=0}^{\infty}\binom{k+r-1}{r+1}p^{r+2}q^{k-2}=\frac{(r+1)rq^2}{p^2}
\end{aligned}
$$

Now seeing that $\sum_{k=0}^{\infty}\binom{k+r-1}{r+1}p^{r+2}q^{k-2}$ is the sum over all values *k* of the probability of *k* failures before the *r*+2nd success, and must be one. Thus, we now have $\mathbf{E}\left[V(V-1)\right]=\frac{(r+1)rq^2}{p^2}$, and we compute

$$
\begin{aligned}
\mathbf{Var}[V]&=\mathbf{E}\left[V(V-1)\right]+\mathbf{E}[V]-\mathbf{E}^2[V]\\
&=\frac{(r+1)rq^2}{p^2}+\frac{rq}{p}-\frac{r^2q^2}{p^2}\\
&=\frac{r^2q^2+rq^2+rq(1-q)-r^2q^2}{p^2}\\
&=\frac{rq}{p^2}.
\end{aligned}
$$

When the random variable is not *V*, the number of failures before the $r^{\text{th}}$ success but *Y*, the trial on which the $r^{\text{th}}$ success occurs, then $Y=V+r$, therefore $\mathbf{E}[Y]=\frac{rq}{p}+r=\frac{r}{p}$, and

$$\mathbf{Var}[V]=\mathbf{Var}[Y+r]=\mathbf{Var}[Y]=\frac{rq}{p^2}.$$

# Basics of the Poisson Process

Another type of natural experiment on which we can base the construction of probabilities is that of independent arrivals. This is an important departure for us because, prior to this, our work in random variables focused on the 0-1 "success-failure" model of Bernoulli trials.

## Prerequisites

Why Probability
The Random Event
Elementary Set Theory
Properties of Probability
Conditional Probability
Basics of Bernoulli Trials.
Basics of the Binomial Distribution

## Events or Arrivals

Here we retain interest in our concept of independence. However, we disconnect from the notion of an event being either a success or a failure. We are interested now in the concept of an "arrival". An example would be the event that there are exactly seven subjects who arrive to a suburban clinic in an hour.

If you have not seen the probability function for the Poisson process, it can appear to be strange and somewhat "out of the blue". For the Poisson process, the probability that there are $k$ "arrivals" in the time interval $t$ is

$$\mathbf{P}_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

We can compute this for all integer values of $k \geq 0$, which is clearly an infinite number of events. Yet, the sum of all of these probabilities is one. Summing an infinite number of values to obtain a finite value can seem a contradiction. It is the property of some sums, that, when summed to infinity, their sum is not infinity but actually a finite value.

Developed by Siméon Poisson, the Poisson distribution has a wide range of uses and applications. One of the first demonstrations of its ubiquity was the study of the probability distribution of horse kicks in the 19th century Russian army.* It is also very helpful in the consideration of sample spaces composed of rates. More contemporary examples follow.

## Example: Arrivals to a clinic

An emergency center opens for the day. On average, patients arrive at the rate of six per hour. What is the probability that there are seven arrivals in the first hour?

We compute $\mathbf{P}_7\left(1\right)=\frac{\left(6\right)^7}{7!}e^{-7}=0.138.$

A useful feature of the Poisson distribution is that it scales to different time periods based on $\lambda$. If we need to compute the probability for an interval that is different than that from which $\lambda$ represents, we can readily compute the probability. For example if we change the arrival time from one to five hours, and ask what is the probability that 20 subjects arrive in five hours, we compute $\mathbf{P}_{20}\left(5\right)=\frac{\left(\lambda t\right)^k}{k!}e^{-\lambda t}=\frac{\left(\left(6\right)\left(5\right)\right)^{20}}{20!}e^{-(6)(5)}=0.013.$ ■

## Moments of the Poisson process

The mean of the Poisson process can be computed directly as $\mathbf{E}\left[X\right]=\lambda$. It can also be demonstrated that its variance $\mathbf{Var}\left[X\right]=\lambda$. Thus, we have the unusual finding that $\mathbf{E}\left[X\right]=\mathbf{Var}\left[X\right]=\lambda$.

Hypergeometric Measure
Limits and Continuity
Probability as a Continuous Function
Basics of Normal Measure

* This is an amusing concept to us, but at the time it was a serious issue. Horses were a necessary military tool and resource in the 19th century, and horse kicks are violent attacks, commonly fracturing legs, arms and ribs. Add to this the poor standard of care for treatment of these injuries, which commonly led to amputation or death, and one can appreciate why understanding the expected number of these injuries would be of interest.

However, this was not the reason for the application of the probability tool to these accidents. The Czar was actually concerned that the Almighty may not sanction his plan for military campaigns, signally His displeasure by increasing the number of horse kicks. Poisson was asked whether he could discern a change in the horse kick distribution, and therefore God's will.

# General Poisson Measure

Thus far our consideration of different probability distributions has been based on the Bernoulli, itself the results of independent sequences of successes and failures.

However, we will find another type of "natural experiment" generates an entirely new and large collection of probabilities that are described with the sobriquet "the Poisson Process."

## Prerequisites

The Random Event
Elementary Set Theory
Factorials Permutations, and Combinations
The Concept of the Limit
Convergent Series
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Basic Properties of the Lebesgue-Stieltjes Integral
Properties of Probability
Conditional Probability
Advanced Binomial Distribution

## The basic experiment

The basis of the Poisson process is the occurrence of one of a sequence of events that occur at different frequencies but whose frequency or rate of occurrence can be averaged. The arrival of these events are independent of each other.

Classic examples are the arrival of patients to an urgent care clinic, the arrival of phone calls to a police station, the aggregation of red blood corpuscles in a hemocytometer, or the occurrence of misprints on a book page.

In each case, the events occur independently of each other, and can be classified as having arrived at a given average rate. However, the rate need not be time. For example, there can be three arrivals to an urgent care clinic per hour, or 0.07 misprints per page.

What we are interested in is computing the probability of *k* arrivals in a given period of time, for example, what is the probability that there will be seven patients arriving to a suburban clinic in an hour?

For the Poisson process, this measure is

$$\mathbf{P}_k\left(t\right)=\frac{\left(\lambda t\right)^k}{k!}e^{-\lambda t}1_{k\in\left[I_{0,\infty}\right]}.$$

Note, that like geometric measure, each of the nonnegative integers is a possible value of *k*, so that we are relying on the convergence of an infinite sum for this function to have one of the principal features of a probability function i.e., that it integrate to one. In this case, we are relying on the convergent properties of the exponential function, so we may write

$$\int_{\Omega} d\mathbf{P} = \int_{\Omega} \frac{(\lambda t)^k}{k!} e^{-\lambda t} 1_{k\in\left[I_{0,\infty}\right]} = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\lambda t} e^{\lambda t} = 1.$$

As we discussed in the binomial distribution's development the integral sign simply announces our attempt to take a measure over the integrals limits. The integrand is the measuring tool, telling us how to compute the measure.

Developed by Siméon Poisson, the Poisson distribution has a wide range of uses and applications. It has made substantial contributions to computing the likelihood of events involving disease rates.

### *Example: Arrivals to a clinic*

A clinic opens on campus to begin influenza immunizations. Assume subjects arrive at the rate of two per hour. Find the probability that four subjects arrive in the first hour.

We compute $\mathbf{P}_4(1) = \frac{(2)^4}{4!} e^{-2} = 0.092.$

A useful feature of Poisson measure is that it scales $\lambda,$ so that it we need to compute the probability for an interval that is different than that from which the value of $\lambda$ represents, we can readily compute the probability. For example if we change the arrival time from one to six hours, and ask what is the probability that five subjects arrive in six hours, we compute

$$\mathbf{P}_5(6) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \frac{(2(6))^5}{5!} e^{-(2)(6)} = 0.012.$$

■

### *Example: Spect emissions*

Single-photon emission computed tomography (SPECT) imaging with adenosine infusion over four minutes is commonly carried out in order to identify areas of poor heart function during exercise. This procedure requires the use of radioactive particles which are emitted and degrade over time. Assume five particles are emitted each second on average. What is the probability that in six seconds, thirty or more particles are emitted.

Here we assume particles are emitted independently of one another. Then for this example, $\lambda = 5,$ and we compute

$$\mathbf{P}\left[K \geq 30\right] = \sum_{k=30}^{\infty} \frac{(5*6)^k}{k!} e^{-(5*6)} = 0.524.$$

■

### *Example: Cell counts*

Before the advent of technology that automatically counted the number of red cells, a health care worker would place a drop of blood on a slide with grids and then count the number of cells in each grid. Assume that blood does not clump and that there are on average three erythrocytes per grid, then what is the probability that a grid will have no cells? We compute $\mathbf{P}\left[k=0\right]=e^{-3}=0.05$.

This work has modern implications as well. A germane issue in the burgeoning field of cell therapy is whether cells arrive into the tissue where they are required. For example, suppose cells arrive with an average rate of 7 cells per millisecond. What is the likelihood that in four milliseconds, 30 cells will have arrived. We compute

$$\mathbf{P}\left[k=30\right]=\frac{28^{30}}{30!}e^{-28}=0.068.$$

■

***Example: Misprints***

An senior editor is reviewing the draft of a new book in a proteomics series that will ultimately contain ten such books. This book is 495 pages long, and he has been told that a review by the author has identified 75 misprints. How likely is it that ten consecutive pages will have at least one misprint?

Assuming the misprints occur independently of each other, it is reasonable to assume that they "arrive" at the rate of 75/495 or 0.15 misprints per page, or 1.5 misprints per ten pages. The probability that there is at least one misprint in ten pages is

$$\mathbf{P}\left[K\ \geq\ 1\right]=1-\mathbf{P}\left[K\ =\ 0\right]=1-e^{-1.5}=0.77.$$

Note that because they "arrive" independently, the distribution of the pages, (consecutive or not) has no impact on the solution.

■

***Example: Vacancies of the US Supreme Court***.

Emanuel Parzen in his textbook* presented data on vacancies that occurred in the US Supreme Court from 1837 to 1932. He demonstrated that vacancies occurred at a rate of $\lambda=0.50$ vacancies per year. Accepting this, what is the probability that a US president will have to fill at least two vacancies in a four year term?

We compute this simply as $\mathbf{P}\left(4\right)=\sum_{k=2}^{\infty}\frac{\left(\lambda t\right)^{k}}{k!}e^{-\lambda t}$ where $\lambda t=(0.50)(4)=2.$ Thus

$\mathbf{P}\left(4\right)=\sum_{k=2}^{\infty}\frac{(2)^{k}}{k!}e^{-2}=0.59.$ The probability that the president will make at least two appointments in eight years is $\mathbf{P}\left(2\right)=\sum_{k=2}^{\infty}\frac{\left(4\right)^{k}}{k!}e^{-4}=0.91.$

■

It is difficult to overstate the omnipresence of the Poisson process.

## Moments and generating function

The mean of the Poisson process can be computed directly as

* Parzen E. (1960). Modern Probability and its Applications. Wiley, p 256.

$$
\begin{aligned}
\mathbf{E}\left[X\right] &= \int_{\Omega} xd\mathbf{P} = \int_{\Omega} k\frac{\lambda^{k}}{k!}e^{-\lambda}1_{k\in\left[I_{0,\infty}\right]} = \sum_{k=0}^{\infty}k\frac{\lambda^{k}}{k!}e^{-\lambda} = \sum_{k=1}^{\infty}k\frac{\lambda^{k}}{k!}e^{-\lambda} \\
&= \sum_{k=1}^{\infty}\frac{\lambda^{k}}{\left(k-1\right)!}e^{-\lambda} = e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{\left(k-1\right)!} = e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{\left(k-1\right)!} \\
&= e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{\left(k-1\right)!} = e^{-\lambda}\lambda\sum_{k=0}^{\infty}\frac{\lambda^{k}}{k!} = e^{-\lambda}e^{\lambda}\lambda = \lambda
\end{aligned}
$$

To find the variance, we first compute the factorial moment $\mathbf{E}\left[X\left(X-1\right)\right]$.

$$
\begin{aligned}
\mathbf{E}\left[X\left(X-1\right)\right] &= \int_{\Omega} x\left(x-1\right)d\mathbf{P} = \int_{\Omega} k\left(k-1\right)\frac{\lambda^{k}}{k!}e^{-\lambda}1_{k\in\left[I_{0,\infty}\right]} \\
&= \sum_{k=0}^{\infty}k\left(k-1\right)\frac{\lambda^{k}}{k!}e^{-\lambda} = \sum_{k=2}^{\infty}k\left(k-1\right)\frac{\lambda^{k}}{k!}e^{-\lambda} = \sum_{k=2}^{\infty}k\left(k-1\right)\frac{\lambda^{k}}{k!}e^{-\lambda} \\
&= \sum_{k=2}^{\infty}\frac{\lambda^{k}}{\left(k-2\right)!}e^{-\lambda} = \lambda^{2}e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{\left(k-2\right)!} \\
&= \lambda^{2}e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k}}{k!}e^{-\lambda}e^{\lambda}\lambda^{2} = \lambda^{2}.
\end{aligned}
$$

So we compute

$$
\begin{aligned}
\mathbf{Var}\left[X\right] &= \mathbf{E}\left[X^{2}\right]-\mathbf{E}^{2}\left[X\right] = \mathbf{E}\left[X\left(X-1\right)\right]+\mathbf{E}\left[X\right]-\mathbf{E}^{2}\left[X\right] = \\
&= \lambda^{2}+\lambda-\lambda^{2} = \lambda.
\end{aligned}
$$

Thus, we have the unusual finding that $\mathbf{Var}\left[X\right]=\mathbf{E}\left[X\right]=\lambda.$ We can also identify the probability generating function easily,

$$
\begin{aligned}
\mathbf{G}_{s}(t) &= \mathbf{E}\left[s^{x}\right] = \int_{\Omega} s^{x}d\mathbf{P} = \int_{\Omega} s^{k}\frac{\left(\lambda t\right)^{k}}{k!}e^{-\lambda t}1_{k\in\left[I_{0,\infty}\right]} = \sum_{k=0}^{\infty}\frac{\left(\lambda ts\right)^{k}}{k!}e^{-\lambda t} \\
&= e^{-\lambda t}\sum_{k=0}^{\infty}\frac{\left(\lambda ts\right)^{k}}{k!} = e^{\lambda t\left(s-1\right)}.
\end{aligned}
$$

Manipulation of the generating function leads to a particularly succinct identification of the mean and variances.

## Motivation for the Poisson process

Let's now return to the form of the probability function for the Poisson process in order to motivate it. The Poisson distribution function can actually be derived from first principles. We will give just a brief summary of the thought process here, and then refer to the immigration process for a series of detailed derivations.

The process begins with an examination of a system e.g., arrivals of individuals to a community. Since multiple arrivals can occur "on top of each other", we focus on a tiny sliver of time, $\left(t,\ t+\Delta t\right)$ a slice so thin that one and only one arrival can occur, if it occurs at all. For this small time interval $\Delta t,$ $\left(t,\ t+\Delta t\right),$ we write a difference equation for the process in the small time period for $\mathbf{P}_{k}\left(t+\Delta t\right)$ as $\mathbf{P}_{k}\left(t+\Delta t\right)=\lambda\Delta t\mathbf{P}_{k-1}\left(t\right)+\left(1-\lambda\Delta t\right)\mathbf{P}_{k}\left(t\right)$ for all non-negative integers *k*. This is converted to a system of simple differential equations that is in turn converted to one equation using the generating function approach. Inversion produces the Poisson distribution. This entire process can be attributed to Chapman and Kolmogorov.

The characterization of the mean value $\lambda$ can sometimes suggest that the process is well organized and more predictable then is actually the case. The fact that, for example, $\lambda = 4$ patients arriving at an emergency department should not lull us into believing that each hour, plan on four patients arriving.

In fact, the Poisson process is quite chaotic. It is the Poisson process that leads to observations such as patients seem to arrive at a clinic just when it is closing time, or that an emergency room can move from quiescence to pandemonium in the space of few minutes. Descriptions of average event rates can disguise the true haphazardness of the Poisson process

## Sums of independent Poisson processes

Poisson processes have properties that make it one of the easier distributions to work with. In addition to the finding that its mean is also its variance which is simply its parameter $\lambda$, we will now see that sum of Poisson processes is also Poisson.

One quick way to see this is just to carry out the operation on the probability generating functions. Assume that *X* is Poisson($\lambda_1$) and *Y*, independent of *X*, follows a Poisson ($\lambda_2$) process. Then define Z= *X* + *Y*, and write

$$\mathbf{G}_Z(s) = \mathbf{E}\left[s^{(X+Y)}\right] = \mathbf{E}\left[s^X\right]\mathbf{E}\left[s^Y\right] = e^{\lambda_1(s-1)}e^{\lambda_2(s-1)} = e^{(\lambda_1+\lambda_2)(s-1)}.$$

which we recognize as the Poisson process $(\lambda_1 + \lambda_2)$. Note the assumption of independence is key since it permitted us to break the joint expectation of a function of *X* and *Y* into the product of the expectations.

Perhaps another more revealing proof of this assertion is to examine the sum directly. We can use the fact that if *X* + *Y* = *n*, and *X* = *k*, then *Y* must be *n* − *k*. This self-evident observation permits the following;

$$\begin{aligned}\mathbf{P}[Z = n] = \mathbf{P}[X + Y = n] = \mathbf{P}[X = k \cap Y = n-k]\\ = \mathbf{P}[X = k]\mathbf{P}[Y = n-k].\end{aligned}$$

Again, independence permits us to write the joint probability of *X* and *Y* into the product of probabilities, one involving *X*, the other *Y*.

The assertion is true for any value of *k* lying between 0 and *n*, so we now write.

$$\begin{aligned}\mathbf{P}[Z = n] = \sum_{k=0}^{n}\mathbf{P}[X = k]\mathbf{P}[Y = n-k] = \sum_{k=0}^{n}\frac{\lambda_1^k}{k!}e^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}e^{-\lambda_2}\\ = e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^k}{k!}\frac{\lambda_2^{n-k}}{(n-k)!} = e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{1}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k}.\end{aligned}$$

The appearance of the combinatoric in the denominator suggests a familiar pattern. Proceeding

$$e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{1}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k}=\frac{1}{n!}e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k}$$
$$=\frac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\frac{\lambda_1^k}{(\lambda_1+\lambda_2)^k}\frac{\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^{n-k}}$$
$$=\frac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}$$

However, $\sum_{k=0}^{n}\binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}$ is just the sum over all possible values of *k* of a binomial random variable with probability of "success" $\frac{\lambda_1}{\lambda_1+\lambda_2}$. This sum is one, returning our result for the sum of two independent Poisson processes.

It may seem surprising for the Poisson process to emerge "out of the blue" like this. However we will see that these two different probability distributions are similar enough for them to cross each other's path when we are not looking for it.

***Example: Intensive care arrivals***.

A small hospital in an inner city community has one intensive care unit, that accepts patients from both the medical service and the surgical service. Assume that patients are admitted to the intensive care unit from the medical service at the rate of two per day; post-surgical patients arrive at the rate of 3.5 patients per day. What is the probability that 7 patients have arrived in a given day?

Note we are not interested in the distribution of the 7 patients, i.e., how many are medical and how many are surgical, only in the total number of patients. We simply sum the averages and compute;

$$\mathbf{P}(7)=\frac{(5.5)^7}{7!}e^{-5.5}=0.123.$$

However, suppose we were interested in the probability that, given there are seven patients in intensive care, 2 of them were from medical service. In general, we would like $\mathbf{P}\left[X=k\,|\,X+Y=n\right]$. We proceed as follows:

$$\mathbf{P}\left[X=k\,|\,X+Y=n\right]=\frac{\mathbf{P}\left[X=k\cap X+Y=n\right]}{\mathbf{P}\left[X+Y=n\right]}.$$

The numerator is the key. We write this as

$$\mathbf{P}\left[X=k\cap X+Y=n\right]=\mathbf{P}\left[X=k\cap Y=n-k\right]$$
$$=\mathbf{P}\left[X=k\ \right]\mathbf{P}\left[Y=n-k\right].$$

Independence is again critical. You might have a clue what is coming.

$$\mathbf{P}\left[X=k \mid X+Y=n\right]=\frac{\mathbf{P}\left[X=k \cap X+Y=n\right]}{\mathbf{P}\left[X+Y=n\right]}$$
$$=\frac{\mathbf{P}\left[X=k\ \right]\mathbf{P}\left[Y=n-k\right]}{\mathbf{P}\left[X+Y=n\right]}$$
$$=\frac{\frac{\lambda_1^k}{k!}e^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^n}{n!}e^{-\lambda_1+\lambda_2}}=\frac{n!}{k!(n-k)!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n}$$
$$=\binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}.$$

and we see the solution is binomial measure. We compute

$$\mathbf{P}\left[X=2 \mid X+Y=7\right]=\binom{7}{2}\left(\frac{2}{5.5}\right)^2\left(\frac{3.5}{5.5}\right)^5$$
$$=(21)(0.132)(0.104)=0.288.$$

## Expectations and “Layered Cake”

And interesting and useful relationship between the expectation and cumulative distribution function of a nonnegative valued random variable is the relationship

$$\mathbf{E}\left[X\right]=\int_0^\infty\left[1-F_X(x)\right]dx$$

which for discrete measure is simply

$$\mathbf{E}\left[X\right]=\sum_{k=0}^\infty\mathbf{P}\left[X>k\right]$$

The proof is more of a demonstration and is straightforward. Simply write

$$\sum_{k=0}^\infty\mathbf{P}\left[X>k\right]=\mathbf{P}\left[X>0\right]+\mathbf{P}\left[X>1\right]+\mathbf{P}\left[X>2\right]+\mathbf{P}\left[X>3\right]+\mathbf{P}\left[X>4\right]+...$$
$$=\mathbf{P}\left[X>0\right]$$
$$+\mathbf{P}\left[X>1\right]$$
$$+\mathbf{P}\left[X>2\right]$$
$$+\mathbf{P}\left[X>3\right]$$
$$+\mathbf{P}\left[X>4\right]$$
$$+...$$

We simply now expand each of these lines out to write

$$\sum_{k=0}^{\infty}\mathbf{P}[X>k]$$

$$=\mathbf{P}[X=1]+\mathbf{P}[X=2]+\mathbf{P}[X=3]+\mathbf{P}[X=4]+\mathbf{P}[X=5]+\ldots$$

$$+\mathbf{P}[X=2]+\mathbf{P}[X=3]+\mathbf{P}[X=4]+\mathbf{P}[X=5]+\ldots$$

$$+\mathbf{P}[X=3]+\mathbf{P}[X=4]+\mathbf{P}[X=5]+\ldots$$

$$+\mathbf{P}[X=4]+\mathbf{P}[X=5]+\ldots$$

$$+\mathbf{P}[X=5]+\ldots$$

$$+\ldots$$

Note that each succeeding summation is one term shorter than the preceding one, like pieces of cake that are progressively shorter as they get closer and closer to the top (although here, the cake is of infinite height!).

And now we simply count like terms to write

$$\sum_{k=0}^{\infty}\mathbf{P}[X>k]=1\mathbf{P}[X=1]+2\mathbf{P}[X=2]+3\mathbf{P}[X=3]+4\mathbf{P}[X=4]+5\mathbf{P}[X=5]+\ldots$$

$$=\mathbf{E}[X].$$

There is also a continuous analogue for the layered cake equality.

## Queuing Theory

The applicability of the Poisson processes is just an introduction to the rich world of stochastic processes.

A stochastic process is a process that is probabilistic in time. We have motivated and seen how patients arriving to a clinic follows a Poisson process. This required us to model the arrival of patients entering the system.

But suppose that we also watched these patients' departure times as well. That process is also stochastic, and a Poisson process could be built to manage their departure. Furthermore, the simultaneous management of both arrivals and departures (emigration) would allow one to compute the distribution of the number of patients actually in the clinic at a given time.

The management of these complex operations is the introduction to probabilistic systems management  Another use of this system is governing the queuing process. In the 1970's many service delivery systems (e.g., banks, airline ticketing) required waiting on lines (or in the parlance of the field "queuing up") changed from having each server have a line of customers to a system where there was a single line and multiple servers. This change was made because the multi-server, single queue system had a shorter average wait time then the multi-line alternative. This change was motivated by and a consequence of the findings of the Poisson process  applied to queuing theory.

Other applications include epidemiology, in which disease can enter  and then spread through a community.

## Mixing the Poisson with other measures

In another demonstration of the utility of the Poisson distribution, consider the following random variable $W=X+Y$ where $X$ follows a Poisson distribution with mean $\lambda$ and $Y$ follows an independent geometric distribution with parameter *p*, i.e., $\mathbf{P}\left[Y = j\right] = pq^j$. Note that both random variables have positive probability on the nonnegative integers. To find the distribution of *W* we write

$$\mathbf{P}\left[W = k\right] = \mathbf{P}\left[X + Y = k\right] = \sum_{j=0}^{k} \mathbf{P}\left[X = j, Y = k - j\right]$$

$$= \sum_{j=0}^{k} \mathbf{P}\left[X = j\right] \mathbf{P}\left[Y = k - j\right]$$

which continuing is

$$= \sum_{j=0}^{k} \frac{\lambda^j}{j!} e^{-\lambda} pq^{k-j} = pq^k \sum_{j=0}^{k} \frac{\lambda^j}{j!} e^{-\lambda} q^{-j} = pq^k \sum_{j=0}^{k} \frac{\left(\frac{\lambda}{q}\right)^j}{j!} e^{-\lambda}$$

Simplifying reveals

$$= pq^k \sum_{j=0}^{k} \frac{\left(\frac{\lambda}{q}\right)^j}{j!} e^{-\lambda} e^{q} e^{\frac{1}{q}} = pq^k e^{q} \sum_{j=0}^{k} \frac{\left(\frac{\lambda}{q}\right)^j}{j!} e^{-\frac{\lambda}{q}}$$

Which is just the cumulative probability function of a Poisson ($\lambda/q$) scaled by $pq^k e^{-q}$.

## Example: Health care worker census:

Probability problems are commonly best approach by reducing the problem to its simplest elements. When this simplest solution is obtained, then, using the new found intuition gained from providing this initial solution, we can add layer upon layer of required complexity until the final solution is achieved.

Consider a county public health care team required to assess the number of health care workers who report to their clinic on a given day The clinics are of two types. The first reports the actual number of health care workers that it has working that day. The second reports only whether it is fully staffed or not. The management team needs to know the probability distribution of the number of health care workers reporting. Full staffed means there are four health care workers who have reported to work.

Let's approach this based on a set of simple assumptions. Assume there is one clinic of each of these two types. We let *X* be the number of health care workers at the first clinic. This follows a Poisson distribution with parameter $\lambda$. The random variable $X$ can theoretically take on values 0 to infinity, and $\mathbf{P}\left[X = k\right] = \frac{\lambda^k}{k!} e^{-\lambda} 1_{k \subset N}$ where *N* is a natural number.

This would be our solution if we only had to assess the first clinic. However, the second clinic adds 4 staff or no staff. The likelihood that these four staff workers are added or not follows a modified Bernoulli distribution, $\mathbf{P}\left[Y = k\right] = p1_{k \subset 4}$. The possible values of the random variable *Y* are only 0 and 4.[*]

What can we now say about the event $X + Y = k$? It is the event where $Y = 4$ and $X = k - 4$, or $Y = 0$ and $X = k$. This is the key observation. In one circumstance we can use the

[*] This assumes a worst case scenario where if no report is made, then no health care workers come to work. Solutions are available for the other assumpions of 1, 2, or 3 appearing.

Poisson distribution by itself. However, for the event that the second clinic does have four health care workers appear, we only need $k-4$ workers from the Poisson random variable.

Since these events are mutually exclusive and exhaust the space, we can write

$$\mathbf{P}\left[X+Y=k\right]=p\frac{\lambda^{k-4}}{\left(k-4\right)!}e^{-\lambda}1_{k>4}+\left(1-p\right)\frac{\lambda^{k}}{k!}e^{-\lambda}1_{k\subset N}.$$

Now assume that there are three such clinics, the first reporting as a Poisson random variable, the second two reporting as a modified Bernoulli random variable with the same probability *p* of being fully staffed with 4 health care workers.

Before we begin our cerebration, consider that the sum of Bernoulli random variables with the same probability of success is a binomial random variable. If there were standard Bernoulli random variables we would write $\mathbf{P}\left[Y_1+Y_2=k\right]=\binom{2}{k}p^{k}\left(1-p\right)^{n-k}1_{k\subset 0,1,2}.$

However, since the values of modified Bernoulli random variables are 0,1, or 2, the values of the now modified binomial random variables are 0, 4, and 8. We write the modified binomial random variable as

$$\mathbf{P}\left[Y_1+Y_2=k\right]=\binom{2}{\frac{k}{4}}p^{\frac{k}{4}}\left(1-p\right)^{\frac{k}{4}-1}1_{k\subset 0,4,8}.$$ since the function $\frac{k}{4}$ maps 0, 4, 8 to 0, 1, 2.

Thus the probability that $\mathbf{P}\left[X+Y_1+Y_2=m\right]$ is $\sum_{k=0}^{2}\binom{2}{k}p^{k}\left(1-p\right)^{2-k}\frac{\lambda^{m-4k}}{\left(m-4k\right)!}e^{-\lambda}$

Generalizing to *n* such clinics we have

$$\mathbf{P}\left[X+\sum_{j=1}^{n}Y_j=m\right]=\sum_{k=0}^{n}\binom{n}{k}p^{k}\left(1-p\right)^{n-k}\frac{\lambda^{m-4k}}{\left(m-4k\right)!}e^{-\lambda}.$$

Finally, if we have *L* clinics that report the actual number of health care workers who appear, each clinic following its own Poisson process with parameter $\lambda_i, i=1,2,3...L,$ Since, the sum of independent Poisson random variables follows Poisson measure, and its parameters is the sum of the individual parameters, then the probability distribution of the number of health care workers in the system on a given day is

$$\mathbf{P}\left[\sum_{i=1}^{L}X_i+\sum_{j=1}^{n}Y_j=m\right]$$

$$=\sum_{k=0}^{n}\binom{n}{k}p^{k}\left(1-p\right)^{n-k}\frac{\left(\sum_{i=1}^{L}\lambda_i\right)^{m-4k}}{\left(m-4k\right)!}e^{-\sum_{i=1}^{L}\lambda_i}.$$

The solution is available through Lebesgue integration theory as well. Recall that $\int d\mathbf{P}$ requires us to move through the domain of *k*, accumulating measure based on the circumstances of the measure (which is neither binomial nor Poisson). For $k=0,$ we have no events coming from the dichotomous reporting clinics and no events coming from the Poisson reporting clinics.

This occurs with probability $\mathbf{P}[0]=(1-p)^n e^{-\sum_{i=1}^{L}\lambda_i}$. We have one event with probability $\mathbf{P}[1]=(1-p)^n\left[\sum_{i=1}^{L}\lambda_i e^{-\sum_{i=1}^{L}\lambda_i}\right]$. Analogously

$$\mathbf{P}[2]=(1-p)^n\left[\frac{\left(\sum_{i=1}^{L}\lambda_i\right)^2}{2!}e^{-\sum_{i=1}^{L}\lambda_i}\right]$$

$$\mathbf{P}[3]=(1-p)^n\left[\frac{\left(\sum_{i=1}^{L}\lambda_i\right)^3}{3!}e^{-\sum_{i=1}^{L}\lambda_i}\right].$$

$\mathbf{P}[4]$ presents a complication. One of the dichotomously reporting clinics reports that they are fully staffed, or none of them report a full staff and all health care workers are reported from the Poisson reporting clinics. We write this as

$$\mathbf{P}[4]=\binom{n}{1}p(1-p)^{n-1}e^{-\sum_{i=1}^{L}\lambda_i}+(1-p)^n\left[\frac{\left(\sum_{i=1}^{L}\lambda_i\right)^4}{4!}e^{-\sum_{i=1}^{L}\lambda_i}\right].$$

And we see that we can accumulate probability as we move through the values of $k$. It can sometimes be easier to take this approach rather than our first heuristic one, and in the process develop a new measure.

■

## Other uses of Poisson measure

The Poisson can sometimes be applied to processes that we know ultimately are not random at all. One example, is the arrival and departure of planes in the airspace around an airport. At first blush this seems like a natural application of the Poisson process, because planes appear to appear randomly in the airspace, with others departing randomly. We envision the construction of an average number of arrivals and departures in a unit time and compute probabilities in accordance with the Poisson process.

However, isn't airspace tightly controlled? The terminal radar approach control (TRACON) facilities are the controllers who tightly monitored what occurs in the airspace surrounding major airports. Airplanes file flight plans, and most commonly arrive in accordance with those flight plans. Similarly, planes take off according to tight schedules. In fact, there is very little random about planes entering and leaving airspace. Who among us would fly if it was random?

What has happened here is the combination of events each of which is strictly controlled. Every small aspect of the management of airplane movement (either on the ground or in the air) is tightly controlled and nonrandom, yet the sum or gestalt of all of these features produces a process which has the appearance of randomness.

Similarly for the appearance of hurricanes and typhoons. We understand now that these storms are complex interactions between water, and atmosphere, where the right combination of heat, wind, water, and barometric pressure can produce killer typhoons and super hurricanes. There is nothing random about the meteorology that produces these storms, the sum of innumerable effects. Yet the Poisson (and negative binomial) processes do quite well in predicting these storms. The complex ensemble of deterministic events commonly appears to be random.

In addition, the Poisson distribution can be used to approximate both the binomial distribution and the negative binomial measure. In turn, the Poisson distribution can be approximated by normal measure. These and other useful approximations will be covered in the asymptotics chapter.

Alternative Derivation of Poisson Moments

Immigration-Emigration Modeling
Contagion
Death Process
The Emigration-Death Process
Immigration-Death Process
Continuous Probability Measure
Variable Transformations
Uniform and Beta Measure
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# Alternative Derivation of Poisson Moments

A quicker way to identify the moments of the Poisson distribution requires the use of derivative and uniform convergence

We require $\mathbf{E}\left[X\right]=\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}$. This is a power function for $\lambda$. We know $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=e^{\lambda}$. The uniform convergence of the exponential series of $\mathbf{G}_s(t)$ permits us to take a derivative with respect to $\lambda$ of each side of this equation and equate them. Thus. $\sum_{k=0}^{\infty}k\frac{\lambda^{k-1}}{k!}=e^{\lambda}$, and multiplying both sizes by $\lambda$ reveals $\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}=\lambda e^{\lambda}$. Returning to our original statement we have

$$\mathbf{E}\left[X\right]=\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}=e^{-\lambda}\lambda e^{\lambda}=\lambda.$$

To find the variance we can avoid the factorial argument used in the derivation referred to in the Poisson process by writing $\mathbf{E}\left[X^2\right]=\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}$. We know from above that $\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}=\lambda e^{\lambda}$, so we take another derivative with respect to $\lambda$ in order to write $\sum_{k=0}^{\infty}k^2\frac{\lambda^{k-1}}{k!}=\lambda e^{\lambda}+e^{\lambda}$, and multiply each side by $\lambda$ to produce $\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}=\lambda^2e^{\lambda}+\lambda e^{\lambda}$. Thus

$$\begin{aligned}\mathbf{E}\left[X^2\right]&=\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}=e^{-\lambda}\left(\lambda^2e^{\lambda}+\lambda e^{\lambda}\right)\\&=\lambda^2+\lambda,\end{aligned}$$

And the variance is simply $\lambda^2 + \lambda - \lambda^2 = \lambda$.

Immigration-Emigration Modeling
Contagion
Death Process
The Emigration-Death Process
Immigration-Death Process
Continuous Probability Measure
Variable Transformations
Uniform and Beta Measure
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# Poisson Arrival Departure Models

The Poisson process is commonly referred to as an immigration process. We will now demonstrate this from first principles using a line of reasoning that will be useful for expanding this derivation to processes of epidemiologic significance.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions

## Process of immigration

Consider an emergency department (ED) that accepts the random arrival of patients. Let's assume that at time $t = 0$, there are no patients. As time moves ahead, patients arrive at the ED independently of each other. We will assume for this example that patients do not leave the ED (we will relax this assumption in a later discussion). Then as time goes on, the number $k$ of patients in the ED can either stay the same, or increase. We ultimately want to find the probability that there are $k$ systems in the emergency department at time $t$, a probability that we will denote at $\mathbf{P}_k(t)$.

Let us assume that the average arrival rate of patients over time is $\lambda t$. It may be for example six patients per hour. Then how many patients are in the system at time $t + \Delta t$?

In order to examine this, let's conduct a "thought experiment". Suppose we can actually slow the passage of time down, to the point where minutes seems to us like days. This permits us to observe the arrival of patients in a very small fragment of time, $\Delta t$. In this experiment, we can squeeze $\Delta t$, allowing it to be so small that we can avoid the circumstance where more than one patient arrives in the time interval $t + \Delta t$.

In this small time interval, there are only two possibilities; either a patient arrives, or no patient arrives. The probability that a patient arrives in this small time interval is $\lambda \Delta t$ The probability that a patient does not arrive is $1 - \lambda \Delta t$.

We can use this to compute the probability of several important events in this time interval $\Delta t$. For example, what is the probability that there are no patients in the system at time $t + \Delta t$ i.e., $\mathbf{P}_0(t + \Delta t)$? Since we assume patients can only arrive, we cannot decrease the number of patients to zero in the $\Delta t$ interval. The only way that we can have no patients in the ED at time $t + \Delta t$ is to have no patients at time *t* (an event that occurs with probability $\mathbf{P}_0(t)$ and there are no arrivals in time $t + \Delta t$ Thus, for this small time interval we can write

$$\mathbf{P}_0(t + \Delta t) = \mathbf{P}_0(t)(1 - \lambda \Delta t).$$

Now, how can we compute the probability that there is one patient in the emergency department at time $t + \Delta t$ i.e., $\mathbf{P}_1(t + \Delta t)$? In this case, there are two paths to get to one patient in the system at time *t*. We can have no patients in the system at time *t*, plus an arrival during the $\Delta t$ time interval. Alternatively, we can have one patient in the system at time *t*, and then are no arrivals in the $\Delta t$ interval. We are now ready to write an equation for $\mathbf{P}_1(t + \Delta t)$

$$\mathbf{P}_1(t + \Delta t) = \mathbf{P}_0(t)\lambda \Delta t + \mathbf{P}_1(t)(1 - \lambda \Delta t).$$

Similarly, for $\mathbf{P}_2(t + \Delta t)$ we can write

$$\mathbf{P}_2(t + \Delta t) = \mathbf{P}_1(t)\lambda \Delta t + \mathbf{P}_2(t)(1 - \lambda \Delta t).$$

In general, for $k > 0$

$$\mathbf{P}_k(t + \Delta t) = \mathbf{P}_{k-1}(t)\lambda \Delta t + \mathbf{P}_k(t)(1 - \lambda \Delta t).$$

Thus, the collection of equations of interest is

$$\begin{array}{ll} \mathbf{P}_0(t + \Delta t) = \mathbf{P}_0(t)(1 - \lambda \Delta t) & k = 0 \\ \mathbf{P}_k(t + \Delta t) = \mathbf{P}_{k-1}(t)\lambda \Delta t + \mathbf{P}_k(t)(1 - \lambda \Delta t) & k > 0. \end{array}$$

This collection of equations describe in a recursive fashion, the probabilities for all possible values of *k* that are of interest. They represent the set of relationships commonly known as the Chapman-Kolmogorov equations named for Sydney Chapman and Andrey Kolmogorov.

Beginning with $\mathbf{P}_0(t + \Delta t) = \mathbf{P}_0(t)(1 - \lambda \Delta t)$, we can write

$$\begin{gathered}\mathbf{P}_0\left(t+\Delta t\right)=\mathbf{P}_0\left(t\right)\left(1-\lambda\Delta t\right)\\ \mathbf{P}_0\left(t+\Delta t\right)=\mathbf{P}_0\left(t\right)-\mathbf{P}_0\left(t\right)\lambda\Delta t\\ \frac{\mathbf{P}_0\left(t+\Delta t\right)\text{-}\mathbf{P}_0\left(t\right)}{\Delta t}=-\lambda\mathbf{P}_0\left(t\right)\end{gathered}$$

We can now take the limit as $\Delta t\to 0$ to find

$$\lim_{\Delta t\to 0}\frac{\mathbf{P}_0\left(t+\Delta t\right)\text{-}\mathbf{P}_0\left(t\right)}{\Delta t}=\frac{d\mathbf{P}_0\left(t\right)}{dt}-\lambda\mathbf{P}_0\left(t\right)$$

This is a first order differential equation for $\mathbf{P}_0\left(t\right)$.

We know return to the equation for $\mathbf{P}_k\left(t\right)$ and follow the same approach

$$\begin{aligned}&\mathbf{P}_k\left(t+\Delta t\right)=\mathbf{P}_{k-1}\left(t\right)\lambda\Delta t+\mathbf{P}_k\left(t\right)\left(1-\lambda\Delta t\right)\\ &\mathbf{P}_k\left(t+\Delta t\right)-\mathbf{P}_k\left(t\right)=\mathbf{P}_{k-1}\left(t\right)\lambda\Delta t-\mathbf{P}_k\left(t\right)\lambda\Delta t\\ &\frac{\mathbf{P}_k\left(t+\Delta t\right)-\mathbf{P}_k\left(t\right)}{\Delta t}=\lambda\mathbf{P}_{k-1}\left(t\right)-\lambda\mathbf{P}_k\left(t\right)\\ &\lim_{\Delta t\to 0}\frac{\mathbf{P}_k\left(t+\Delta t\right)-\mathbf{P}_k\left(t\right)}{\Delta t}=\frac{d\mathbf{P}_k\left(t\right)}{dt}=\lambda\mathbf{P}_{k-1}\left(t\right)-\lambda\mathbf{P}_k\left(t\right)\end{aligned}$$

We now have two differential equations, one for $k = 0$, the other for $k > 0$.

$$\begin{aligned}&\frac{d\mathbf{P}_0\left(t\right)}{dt}-\lambda\mathbf{P}_0\left(t\right) && k=0\\ &\frac{d\mathbf{P}_k\left(t\right)}{dt}=\lambda\mathbf{P}_{k-1}\left(t\right)-\lambda\mathbf{P}_k\left(t\right) && k>0\end{aligned}$$

## The generating function argument

Note that this system consists of an infinite number of equations. Our approach will be to collapse this infinite set of equations into a single equation involving a probability generating function, solve for the generating function, and then finally invert it.

Multiplying each side of both equation sets by the appropriate value of $s^k$,

$$\begin{aligned}&s^0\frac{d\mathbf{P}_0\left(t\right)}{dt}-\lambda s^0\mathbf{P}_0\left(t\right)\\ &s^k\frac{d\mathbf{P}_k\left(t\right)}{dt}=\lambda s^k\mathbf{P}_{k-1}\left(t\right)-\lambda s^k\mathbf{P}_k\left(t\right)\end{aligned}$$

and we can now add these equations to see

$$\sum_{k=0}^{\infty}s^k\frac{d\mathbf{P}_k\left(t\right)}{dt}=\lambda\sum_{k=1}^{\infty}s^k\mathbf{P}_{k-1}\left(t\right)-\lambda\sum_{k=0}^{\infty}s^k\mathbf{P}_k\left(t\right)\text{ for }k\geq 0.$$

Notice the starting value of the indices differ from summand to summand.

We have collapsed an infinite set of equations into one equation, setting up a generating function argument. Define $\mathbf{G}_s(t) = \sum_{k=0}^{\infty} s^k \mathbf{P}_k(t)$. The goal is to solve for $\mathbf{G}_s(t)$, and then invert it, in order to identify $\mathbf{P}_k(t)$.

We take these terms one at a time. $\sum_{k=0}^{\infty} s^k \frac{d\mathbf{P}_k(t)}{dt}$ can be reduced as follows;

$$\sum_{k=0}^{\infty} s^k \frac{d\mathbf{P}_k(t)}{dt} = \sum_{k=0}^{\infty} \frac{ds^k\mathbf{P}_k(t)}{dt} = \frac{d\sum_{k=0}^{\infty} s^k\mathbf{P}_k(t)}{dt} = \frac{d\mathbf{G}_s(t)}{dt}.$$

Note how we switched the derivative and summation sign in the second equality sign. This is justified only if the infinite sum satisfies the property of uniform convergence, a property that we have demonstrated for the probability generating function. Managing the remaining terms are well within our skills. We write

$$\lambda\sum_{k=1}^{\infty} s^k \mathbf{P}_{k-1}(t) = s\lambda\sum_{k=1}^{\infty} s^{k-1}\mathbf{P}_{k-1}(t) = s\lambda\sum_{k=0}^{\infty} s^k\mathbf{P}_k(t) = s\lambda\mathbf{G}_s(t),$$

Our equation is now

$$\frac{d\mathbf{G}_s(t)}{dt} = s\lambda\mathbf{G}_s(t) - \lambda\mathbf{G}_s(t) = \lambda\mathbf{G}_s(t)(s-1).$$

which is a very simple first order differential equation that we can readily solve.

$$\frac{d\mathbf{G}_s(t)}{dt} = \lambda\mathbf{G}_s(t)(s-1)$$
$$\frac{d\mathbf{G}_s(t)}{\mathbf{G}_s(t)} = \lambda(s-1)dt$$
$$\int\frac{d\mathbf{G}_s(t)}{\mathbf{G}_s(t)} = \int\lambda(s-1)dt$$
$$\ln(\mathbf{G}_s(t)) = \lambda t(s-1) + \mathbf{C}$$

Where **C** is the constant of integration. We use the state of the system at the beginning to determine its value. At $t = 0$, the only nonzero probability is $\mathbf{P}_0(t) = 1$. Thus $\mathbf{G}_s(0) = s^0\mathbf{P}_0(0) = 1$. Thus, $\mathbf{C} = 0$ and we can conclude $\ln(\mathbf{G}_s(t)) = \lambda t(s-1)$ or $\mathbf{G}_s(t) = e^{\lambda t(s-1)}$. The inversion from our study of the Poisson process generating function reveals $\mathbf{P}_k(t) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$.

How does the solution change if we have not zero but a number of patients, say *a*, in the system at time $t = 0$. The guiding Chapman-Kolmogorov equations don't change (although the interval of interest for *k* does. However, the boundary condition changes for which we find the constant of integration **C**. Thus, at $t = 0$, $\mathbf{G}_s(0) = s^a$. since $\mathbf{P}_a(0) = 1$. Thus $\mathbf{C} = s^a$ and we can conclude $\ln(\mathbf{G}_s(t)) = \lambda t(s-1) + s^a$ or $\mathbf{G}_s(t) = s^a e^{\lambda t(s-1)}$. The inversion requires us to draw only coefficients of $s^{k-a}$ from the $e^{\lambda t(s-1)}$ component $\mathbf{G}_s(t)$. Thus $\mathbf{P}_k(t)=\frac{(\lambda t)^{k-a}}{(k-a)!}e^{-\lambda t}$. In order to have *k* patients in the department at time *t*, we only need to experience $k-a$ arrivals.

## Emigration model

Of course patients when they complete their treatment, leave the emergency department. Let us assume no new arrivals are accepted after midnight. What is the probability that there are *k* patients in the emergency room at time *t* after midnight, given the rate at which they leave or emigrate is $\mu$.

Some simple observations speed the solution. If there are *a* patients in the emergency department at midnight, then no more than *a* patients can emigrate; thus $0 \le k \le a.$ Also, if you divide the system into emergency department and elsewhere, then departures from the emergency department at rate $\mu$ actually represent arrivals elsewhere with an arrival rate $\mu$. Thus, the probability that there are *k* patients in the emergency room at time *t* is the probability of $a-k$ arrivals elsewhere or

$$\mathbf{P}_k(t)=\frac{(\mu t)^{a-k}}{(a-k)!}e^{-\mu t}1_{0\le k\le a.}$$

The generating function approach will support this conclusion, although it is more complicated.

We begin with the Chapman-Kolmogorov equations. As before we focus on the time interval from *t* to $t+\Delta t.$ allowing $\Delta t$ to get smaller and smaller so that only one departure can take place in that interval. Then beginning with the probability that, if there are *a* patients in the system at time *t*, than there are *a* patients at the end of the $t+\Delta t$ is

$$\mathbf{P}_a(t+\Delta t)=\mathbf{P}_a(t)(1-\mu\Delta t).$$

Following our previous work on the immigration model, we can write

$$\begin{aligned}
&\mathbf{P}_a(t+\Delta t)=\mathbf{P}_a(t)(1-\mu\Delta t)\\
&\frac{\mathbf{P}_a(t+\Delta t)-\mathbf{P}_a(t)}{\Delta t}=-\mu\mathbf{P}_a(t)\\
&\lim_{\Delta t\to\infty}\frac{\mathbf{P}_a(t+\Delta t)-\mathbf{P}_a(t)}{\Delta t}=\frac{d\mathbf{P}_a(t)}{dt}=-\mu\mathbf{P}_a(t)
\end{aligned}$$

For $0\le k<a$ we have

$$\begin{aligned}
&\mathbf{P}_k(t+\Delta t)=\mathbf{P}_{k+1}(t)\mu\Delta t+\mathbf{P}_k(t)(1-\mu\Delta t)\\
&\frac{\mathbf{P}_k(t+\Delta t)-\mathbf{P}_k(t)}{\Delta t}=\mu\mathbf{P}_{k+1}(t)-\mu\mathbf{P}_k(t)\\
&\lim_{\Delta t\to\infty}\frac{\mathbf{P}_k(t+\Delta t)-\mathbf{P}_k(t)}{\Delta t}=\frac{d\mathbf{P}_k(t)}{dt}=\mu\mathbf{P}_{k+1}(t)-\mu\mathbf{P}_k(t)
\end{aligned}$$

Therefore

$$\frac{d\mathbf{P}_k(t)}{dt} = \mu\mathbf{P}_{k+1}(t) - \mu\mathbf{P}_k(t) \text{ for } 0 \le k < a$$
$$\frac{d\mathbf{P}_a(t)}{dt} = -\mu\mathbf{P}_a(t) \quad \text{for } k = a$$

Multiplying the first equation by $s^k$ and the second by $s^a$ produces

$$s^k\frac{d\mathbf{P}_k(t)}{dt} = \mu s^k\mathbf{P}_{k+1}(t) - \mu s^k\mathbf{P}_k(t) \text{ for } 0 \le k < a$$
$$s^a\frac{d\mathbf{P}_a(t)}{dt} = -\mu s^a\mathbf{P}_a(t) \quad \text{for } k = a$$

and finally, summing these $a$ + 1 equations produces

$$\sum_{k=0}^{a} s^k \frac{d\mathbf{P}_k(t)}{dt} = \mu\sum_{k=0}^{a-1} s^k\mathbf{P}_{k+1}(t) - \mu\sum_{k=0}^{a} s^k\mathbf{P}_k(t).$$

Defining $\mathbf{G}_s(t) = \sum_{k=0}^{a} s^k\mathbf{P}_k(t)$, we, being guided by our experience with the immigration process, can write

$$\sum_{k=0}^{a} s^k \frac{d\mathbf{P}_k(t)}{dt} = \sum_{k=0}^{a}\frac{d\left(s^k\mathbf{P}_k(t)\right)}{dt} = \frac{d\sum_{k=0}^{a} s^k\mathbf{P}_k(t)}{dt} = \frac{d\mathbf{G}_s(t)}{dt}$$
$$\mu\sum_{k=0}^{a-1} s^k\mathbf{P}_{k+1}(t) = \mu s^{-1}\sum_{k=0}^{a-1} s^{k+1}\mathbf{P}_{k+1}(t) = \mu s^{-1}\sum_{k=1}^{a} s^k\mathbf{P}_k(t)$$
$$= \mu s^{-1}\left(\sum_{k=0}^{a} s^k\mathbf{P}_k(t) - \mathbf{P}_0(t)\right)$$
$$= \mu s^{-1}\left(\mathbf{G}_s(t) - \mathbf{P}_0(t)\right)$$
$$\mu\sum_{k=0}^{a} s^k\mathbf{P}_k(t) = \mu\mathbf{G}_s(t).$$

And the $a+1$ equations collapse to

$$\frac{d\mathbf{G}_s(t)}{dt} = \mu s^{-1}\left(\mathbf{G}_s(t) - \mathbf{P}_0(t)\right) - \mu\mathbf{G}_s(t)$$
$$\frac{d\mathbf{G}_s(t)}{dt} = \mu\mathbf{G}_s(t)\left(s^{-1} - 1\right) - \mu s^{-1}\mathbf{P}_0(t)$$

Let’s assume that the ED departure rate is relatively small compared to the number of individuals in the system so that the probability we actually get down to no individuals left, $\mathbf{P}_0(t)$ is zero. Thus, $\frac{d\mathbf{G}_s(t)}{dt} = \mu\mathbf{G}_s(t)\left(s^{-1} - 1\right)$ and we proceed as we did for the immigration process.

$$\frac{d\mathbf{G}_s(t)}{dt} = \mu\mathbf{G}_s(t)\left(s^{-1}-1\right)$$
$$\frac{d\mathbf{G}_s(t)}{\mathbf{G}_s(t)} = \mu\left(s^{-1}-1\right)dt$$
$$\int\frac{d\mathbf{G}_s(t)}{\mathbf{G}_s(t)} = \int\mu\left(s^{-1}-1\right)dt$$
$$\ln\mathbf{G}_s(t) = \mu t\left(s^{-1}-1\right)+\mathbf{C}.$$

Examination of the boundary conditions reveals that when $t = 0$, $k = a$. Thus $\mathbf{G}_s(0) = s^a$, and $\mathbf{C} = \ln s^a$. Therefore

$$\ln\mathbf{G}_s(t) = \mu t\left(s^{-1}-1\right)+\ln s^a$$
$$\mathbf{G}_s(t) = e^{\mu t\left(s^{-1}-1\right)}s^a.$$

## Emigration process inversion

Let's focus on the generating function $\mathbf{G}_s^*(t) = e^{\mu t\left(s^{-1}-1\right)}$. Suppose we have a probability function identical to the Poisson distribution, but defined on the negative integers, i.e., $\sum_{k=-\infty}^{0}\frac{(\mu t)^k}{|k|!}e^{-\mu t} = \sum_{k=0}^{\infty}\frac{(\mu t)^{-k}}{|-k|!}e^{-\mu t}$. The generating function for this probability function is

$$\mathbf{G}_s^*(t) = \sum_{k=0}^{\infty}s^{-k}\frac{(\mu t)^{-k}}{|-k|!}e^{-\mu t} = \sum_{k=0}^{\infty}s^{-k}\frac{(\mu t)^{k}}{k!}e^{-\mu t}$$
$$= e^{-\mu t}\sum_{k=0}^{\infty}\frac{\left(s^{-1}\mu t\right)^k}{k!} = e^{\mu t\left(s^{-1}-1\right)}.$$

We might define the probability generating function with generating function $\mathbf{G}_s^*(t)$ as the negative Poisson distribution, i.e., $\mathbf{P}\left[Y=-k\right] = \frac{(\mu t)^k}{k!}e^{-ut}$ for $0 < k < -\infty$.

Now returning to $\mathbf{G}_s(t)$, the probability that $k$ subjects have left the department at time $t$ means that we want coefficient of $s$ to the power $a-k$ or $\frac{(\mu t)^{a-k}}{(a-k)!}e^{-ut}$.[*]

## Immigration-emigration

Finally, we will manage both processes simultaneously with arrivals to the Emergency Department occurring as a Poisson process with parameter $\lambda$ and departures with parameter $\mu$ per

[*] If the generating function was $e^{\mu t(s-1)}s^a$, the coefficient of $s^k$ requires the coefficient of $s^{k-a}$ from the Poisson distribution. However, since the distribution is negative valued, we need the coefficient of not $s^{k-a}$ but $s^{a-k}$.

unit time. Let's also assume that there are a patients in the emergency room at time $t = 0$. Our goal is to find $\mathbf{P}_k(t)$, the probability that there are $k$ patients in the ED at time $t$.

Before we compute this probability formally using the Chapman-Kolmogorov equations, we might try a heuristic approach. Let's begin with finding $\mathbf{P}_a(t)$, the probability that no change has occurred over time $t$. This event would only take place if the number of arrivals is equal to the number of departures. The probability that there were exactly $m$ arrivals and $m$ departures is simply $\frac{(\lambda t)^m}{m!}e^{-\lambda t}\frac{(\mu t)^m}{m!}e^{-\mu t}$. We only need sum this probability for all values of $m$ to find

$$\mathbf{P}_a(t) = \sum_{m=0}^{\infty}\frac{(\lambda t)^m}{m!}e^{-\lambda t}\frac{(\mu t)^m}{m!}e^{-\mu t}$$

which can also be written as

This last result, while not being easy to calculate, does give some insight into the event. Here, we have the probability of $2m$ Poisson "events" with rate $\lambda t + \mu t$, multiplied by the binomial probability that exactly $m$ are arrivals and $m$ are departures. Let's now generalize to the case where $k > a$, i.e., there have been more arrivals than departures by time $t$. Then regardless of how many departures there were, there had to be that many arrivals, plus an additional $k - a$ arrivals . Thus we may write

$$\mathbf{P}_k(t) = \sum_{m=0}^{\infty}\frac{(\lambda t)^{m+k-a}}{(m+k-a)!}e^{-\lambda t}\frac{(\mu t)^m}{m!}e^{-\mu t}\mathbf{1}_{k>a,}$$

If $k < a$, then departures exceed arrivals and, if there were $m$ arrivals, there must be $m + a - k$ departures. We conclude

$$\mathbf{P}_k(t) = \sum_{m=0}^{\infty}\frac{(\lambda t)^m}{m!}e^{-\lambda t}\frac{(\mu t)^{m+a-k}}{(m+a-k)!}e^{-\mu t}\mathbf{1}_{0\le k<a}.$$

So, we have three formulas for each of three different scenarios. This presages our work with the Chapman Kolmogorov equations.

Using this approach we, as before, focus on what occurs in the time from $t$ to $t + \Delta t$ where $\Delta t$ is so small that only one arrival, one departure, or no change in the number of patients can take place. This permits us to write an equation for $\mathbf{P}_k(t + \Delta t)$ as

$$\mathbf{P}_k(t+\Delta t) = \mathbf{P}_{k-1}(t)\lambda\Delta t + \mathbf{P}_{k+1}(t)\mu\Delta t + \mathbf{P}_k(t)(1-\lambda\Delta t - \mu\Delta t).$$

We can rearrange terms to write

$$\frac{\mathbf{P}_k(t+\Delta t) - \mathbf{P}_k(t)}{\Delta t} = \lambda\mathbf{P}_{k-1}(t) - \lambda\mathbf{P}_k(t) + \mu\mathbf{P}_{k+1}(t) - \mu\mathbf{P}_k(t).$$

We then take limits to see

$$\lim_{\Delta t \to 0} \frac{\mathbf{P}_k\left(t+\Delta t\right)-\mathbf{P}_k\left(t\right)}{\Delta t} = \frac{d\mathbf{P}_k\left(t\right)}{dt}$$
$$= \lambda \mathbf{P}_{k-1}\left(t\right) - \lambda \mathbf{P}_k\left(t\right) + \mu \mathbf{P}_{k+1}\left(t\right) - \mu \mathbf{P}_k\left(t\right).$$

Define $\mathbf{G}_s(t) = \sum_{k=0}^{\infty} s^k \mathbf{P}_k\left(t\right)$ and multiply each term in the above equation by $s^k$ to find

$$s^k \frac{d\mathbf{P}_k\left(t\right)}{dt} = \lambda s^k \mathbf{P}_{k-1}\left(t\right) - \lambda s^k \mathbf{P}_k\left(t\right) + \mu s^k \mathbf{P}_{k+1}\left(t\right) - \mu s^k \mathbf{P}_k\left(t\right).$$

and summing over all *k*, we compute

$$\sum_{k=0}^{\infty} s^k \frac{d\mathbf{P}_k\left(t\right)}{dt}$$
$$= \lambda \sum_{k=0}^{\infty} s^k \mathbf{P}_{k-1}\left(t\right) - \lambda \sum_{k=0}^{\infty} s^k \mathbf{P}_k\left(t\right) + \mu \sum_{k=0}^{\infty} s^k \mathbf{P}_{k+1}\left(t\right) - \mu \sum_{k=0}^{\infty} s^k \mathbf{P}_k\left(t\right).$$

We recognize components of both the immigration and emigration process here, and taking from those two developments we find

$$\frac{d\mathbf{G}_s(t)}{dt} = \lambda\left(s-1\right) + \mu\left(s^{-1}-1\right).$$

Solving this first order differential equation produces

$$\ln \mathbf{G}_s(t) = \lambda t\left(s-1\right) + \mu t\left(s^{-1}-1\right) + \mathbf{C}$$

The boundary condition $\mathbf{G}_s(0) = s^a$, and we conclude

$$\mathbf{G}_s(t) = s^a e^{\lambda t(s-1)} e^{\mu t\left(s^{-1}-1\right)}$$

Setting aside the boundary condition for a moment, we see that the probability generating function for the immigration-emigration model is the product of two generating function producing $a_k = \frac{\left(\lambda t\right)^k}{k!} e^{-\lambda t}$, and $b_k = \frac{\left(\mu t\right)^k}{k!} e^{-\mu t}$. Some examples are provided (Figure 1).

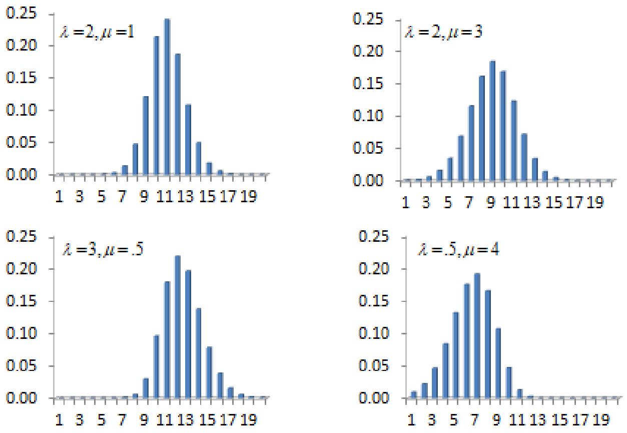


**Figure 1.** Effect on the probability of *k* subjects in the emergency department as a function Of the arrival parameter λ and the departure parameter μ

Contagion

Death Process

The Emigration-Death Process

Variable Transformations

Uniform and Beta Measure

Normal Measure

Compounding

F and T Measure

Ordering Random Variables

Asymptotics

Tail Event Measure

# Birth Process

Development of the study of probabilities that change over time (stochastic processes) has opened a wide field of applications. One of the fields of investigation has been tracking the spread of a contagious disease, such as SARS-COV-2where individuals bring a disease into a community, spreading the disease to its susceptible members.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions
Negative binomial measure
Binomial distribution

## Introduction to the birth process

We begin with a community that is exposed to an infectious disease. At the beginning of the observation time period, there are already *a* subjects in the community with the disease. The disease spreads from individual to individual, its rate of spread governed by the parameter $\upsilon$, i.e., the rate of new infections generated by those already infected is $\upsilon\Delta t$. Clearly, the greater this rate, the more rapidly the disease spreads through the community.

However, another source of disease is individuals arriving into the community, by airplane, ship, train, bus, or car. These individuals arrive at the rate $\lambda\Delta t$ in time period $\Delta t$. Once in the community, they spread the infection at the same rate as those with the disease who are already in the community. It is our goal to identify the probability $\mathbf{P}_k\left[t\right]$ of the number of individuals in the community with the disease at time *t*.

For this simple model, we will assume that there are no deaths, no cures, and no people who leave the community (or emigrate). With our simplifying assumptions, the proportion of patients with the disease will only increase,[*] an increase based solely on the immigration and birth rates.

However, even in this relatively straightforward process, computing $\mathbf{P}_k\left[t\right]$ can seem like a daunting task with the disease being regularly transmitted from one community member to another (who themselves can spread the disease deeper into the populace) in addition to new arrivals who are equally capable of spreading disease. To help manage this process, we will look at what can happen to the disease in a small sliver of time that we denote $\Delta t$.

In our analysis, we will drive the $\Delta t$ to be so small that only one of two events can happen, either 1) one and only one new case of disease appears or 2) no new cases appears.

Continuing with this scenario a new case occurs from the cases already present in the population. In this small interval of time, the probability of the disease being spread to a new individual is $\upsilon\Delta t$.

With this as background, we can now address the question, if there are *k* affected subjects in the community at time *t*, what is the probability that exactly one of them spreads the disease to a susceptible individual and hence generate a new case? If we assume that each affected individual spreads the disease independently of the other, we can use the binominal distribution to find.

$$\mathbf{P}\left[\text{exactly 1 additional case}\right] = \binom{k}{1}\left(\upsilon\Delta t\right)\left(1-\upsilon\Delta t\right)^{k-1}.$$

We may invoke the binomial theorem on the term $\left(1-\upsilon\Delta t\right)^{k-1}$ to write

$$\begin{aligned}\left(1-\upsilon\Delta t\right)^{k-1} &= \sum_{j=0}^{k-1}\binom{k-1}{j}\left(-1\right)^j\left(\upsilon\Delta t\right)^j \\ &= 1-\binom{k-1}{1}\upsilon\Delta t+\binom{k-1}{2}\left(\upsilon\Delta t\right)^2-\binom{k-1}{3}\left(\upsilon\Delta t\right)^3+....\end{aligned}$$

Therefore

$$\mathbf{P}\left[\text{exactly 1 additional case}\right] = \binom{k}{1}\left(\upsilon\Delta t\right)\left(1-\upsilon\Delta t\right)^{k-1}$$

$$= k\left(\upsilon\Delta t\right)\left(1-\binom{k-1}{1}\upsilon\Delta t+\binom{k-1}{2}\left(\upsilon\Delta t\right)^2-\binom{k-1}{3}\left(\upsilon\Delta t\right)^3+..\right).$$

However, since $\upsilon\Delta t$ is so small, we can safely assume that higher order terms are negligible and can be ignored. We therefore write $\mathbf{P}\left[\text{exactly 1 additional case}\right] \sim k\left(\upsilon\Delta t\right)$. Thus the more subjects with the disease, the greater the likelihood that an additional patient will contract the illness. This is the "birth" part of the immigration-birth process.

The driving force for spreading the number of cases in the community is really powered by the number of cases already there. This "birth force" grows ever stronger as the number of cases in the system increases.

Immigration, on the other hand is an influence that is not proportional to the number of cases outside the community. This leads us to believe that, even if $\lambda$ and $\upsilon$ are close, the major source of the new cases will still come from within the community, since the driving force is not $\upsilon$ but $k\upsilon$.

With this result, we can ask, how can we progress to their being *k* affected individuals or cases in the system at time $t+\Delta t$ if the time interval $\Delta t$ is so small that we only permit one event to occur in this interval? Since we have cases in the system at time 0, there are only two ways for this to happen; 1) there are $k-1$ in the system at time *t* and a new case occurred in interval

[*] More realistic discussions are referred to at the end of this section.

$t+\Delta t$, by an arrival, or by its spread from an infected person to an uninfected patient, or 2) there were already $k$ patients in the system and no new cases occurred during this time interval. We formally write this as.

$$\mathbf{P}_k\left[t+\Delta t\right]=\mathbf{P}_{k-1}\left[t\right]\left(k-1\right)\upsilon\Delta t+\mathbf{P}_k\left[t\right]\left(1-kv\Delta t\right).$$

This holds for $k\geq a$ where $a$ is the number of patients in the community at time $t=0$. We can reformulate the previous equation as

$$\frac{\mathbf{P}_k\left[t+\Delta t\right]-\mathbf{P}_k\left[t\right]}{\Delta t}=\left(k-1\right)\upsilon\mathbf{P}_{k-1}\left[t\right]-k\upsilon\mathbf{P}_k\left[t\right].$$

It we let $\Delta t$ get smaller and smaller, taking the limit as $\Delta t$ goes to zero, we write

$$\lim_{\Delta t\to 0}\frac{\mathbf{P}_k\left[t+\Delta t\right]-\mathbf{P}_k\left[t\right]}{\Delta t}=\frac{d\mathbf{P}_k\left[t\right]}{dt}=\left(k-1\right)\upsilon\mathbf{P}_{k-1}\left[t\right]-k\upsilon\mathbf{P}_k\left[t\right]$$

that is

$$\frac{d\mathbf{P}_k\left[t\right]}{dt}=\left(k-1\right)\upsilon\mathbf{P}_{k-1}\left[t\right]-k\upsilon\mathbf{P}_k\left[t\right]\qquad k\geq a$$

We will use the generating function approach to collapse this infinite collection of equations into one equation. Let's define $\mathbf{G}_s(t)=\sum_{k=a}^{\infty}s^k\mathbf{P}_k\left[t\right].$ Ultimately we want to find the value of $\mathbf{P}_k\left[t\right]$ for $k\geq a$. Multiplying each side of the previous equation by $s^k$ produces

$$s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}=\left(k-1\right)s^k\upsilon\mathbf{P}_{k-1}\left[t\right]\text{-}ks^k\upsilon\mathbf{P}_k\left[t\right]\quad k\geq a.$$

From our work on the immigration process, we see that this produces.

$$\frac{\partial\mathbf{G}_s(t)}{\partial t}=\upsilon s\left(s-1\right)\frac{\partial\mathbf{G}_s(t)}{\partial s}.$$

Thus we have collapsed an infinite number of equations into one. While in general, these equations can be a challenge to solve, we will see that we can solve this equation without too much difficulty.

This type of partial differential equation has been discussed. A useful tool for the solution of differential equations is that, it one is written as

$$P\frac{\partial F(x,y)}{\partial x}+Q\frac{\partial F(x,y)}{\partial y}=R,$$

Then a useful collection of equalities is

$$\frac{dx}{P}=\frac{dy}{Q}=\frac{dF}{R}.$$

Our equation, written as

$$\frac{\partial \mathbf{G}_s(t)}{\partial t} - \upsilon s\left(s-1\right)\frac{\partial \mathbf{G}_s(t)}{\partial s} = 0,$$

is of this form and therefore we may write

$$\frac{dt}{1} = \frac{ds}{-\upsilon s\left(s-1\right)} = 0.$$

We will use these equalities in combination with the state of the system at time $t$ =0 to identify the generating function $\mathbf{G}_s(t)$ and then invert. Begin with

$$\frac{dt}{1} = \frac{d\mathbf{G}_s(t)}{0}.$$

Integrating, $d\mathbf{G}_s\left(t\right) = 0$ or $\mathbf{G}_s(t) = \Phi\left(C\right)$.

Next we work with $\frac{dt}{1} = \frac{ds}{-\upsilon s(s-1)}$;

$$\frac{dt}{1} = \frac{ds}{-\upsilon s(s-1)}$$

$$\frac{-\upsilon dt}{1} = \frac{ds}{s(s-1)} = \frac{ds}{s-1} - \frac{ds}{s}$$

$$\int -\upsilon dt = \int \frac{ds}{s-1} - \int \frac{ds}{s}$$

$$-\upsilon t = \ln\left(s-1\right) - \ln\left(s\right) = \ln\left(\frac{s-1}{s}\right)$$

Now we just manipulate the constant

$$-\upsilon t = \ln\left(\frac{s-1}{s}\right) + C_1$$

$$C_1 = -\upsilon t \ln\left(\frac{s}{s-1}\right)$$

$$C_2 = \frac{s}{s-1}e^{-\upsilon t}$$

We can now write $\mathbf{G}_s(t) = \Phi\left(c_1\right) = \Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right)$,

Thus our work reveals $\mathbf{G}_s(t) = \Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right)$.

We can now invoke our boundary conditions to identify the final form of $\mathbf{G}_s(t)$. At $t = 0$, $k = a$, leading to

$\mathbf{G}_s(0) = s^a = \Phi\left(e^{-\upsilon(0)}\frac{s}{s-1}\right) = \Phi\left(\frac{s}{s-1}\right)$, or $\Phi\left(\frac{s}{s-1}\right) = s^a$. Assigning $z = \frac{s}{s-1}$ produces $s = \frac{z}{z-1}$, so $\Phi(z) = \left(\frac{z}{z-1}\right)^{a+\frac{\lambda}{\upsilon}}$. Returning to nonzero values of *t* reveals

$$\mathbf{G}_s(t) = \Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right) = \left[\frac{e^{-\upsilon t}\frac{s}{s-1}}{e^{-\upsilon t}\frac{s}{s-1} - 1}\right]^a$$

Simplifying, we find.

$$\left[\frac{e^{-\upsilon t}\frac{s}{s-1}}{e^{-\upsilon t}\frac{s}{s-1} - 1}\right]^a = \left[\frac{se^{-\upsilon t}}{se^{-\upsilon t} - (s-1)}\right]^a = \left[\frac{se^{-\upsilon t}}{1 - s\left(1 - e^{-\upsilon t}\right)}\right]^a$$

Thus

$$\mathbf{G}_s(t) = \left[\frac{se^{-\upsilon t}}{1 - s\left(1 - e^{-\upsilon t}\right)}\right]^a = \left[\frac{e^{-\upsilon t}}{1 - s\left(1 - e^{-\upsilon t}\right)}\right]^a s^a$$

But this we recognize as the probability generating function for negative binomial measure.scaledby $s^a$. Thus,

$$\mathbf{P}_k[t] = \binom{k-1}{a-1} e^{-a\upsilon t}\left(1 - e^{-\upsilon t}\right)^k \mathbf{1}_{k \geq a}.$$

We have seen that the mean and variance of negative binomial measure are $\frac{rq}{p}$ and variance $\frac{rq}{p^2}$. In the case of the birth process, this translates to

$$\mathbf{E}[k] \quad = \left(a + \frac{\lambda}{\upsilon}\right) e^{\left(a+\frac{\lambda}{\upsilon}\right)\upsilon t}\left(1 - e^{-\upsilon t}\right) - \sum_{k=0}^{a} \binom{k + \frac{\lambda}{\upsilon} - 1}{a + \frac{\lambda}{\upsilon} - 1} e^{-\left(a+\frac{\lambda}{\upsilon}\right)\upsilon t}\left(1 - e^{-\upsilon t}\right)^k$$

# Contagion: Immigration-Birth Process

Development of the study of probabilities that change over time (stochastic processes) has opened a wide field of applications. One of the fields of investigation has been tracking the spread of a contagious disease, such as SARS-COV-2where individuals bring a disease into a community, spreading the disease to its susceptible members.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions
Negative binomial measure
Binomial distribution

## Current issues

There is no more contemporary, incisive example than the coronavirus pandemic of 2020-21, caused by SARS-CoV-2. In that case, once the virus enters the community, it spreads from infected individual to non-infected individual over time. Mathematics and epidemiology were ready to apply quantitative tools to predict the spread and growth of epidemics and pandemics.

The COVID-19 pandemic has driven these complex models to the public's awareness. Populations around the world that in 2019 knew little to nothing about epidemiology now learn about (or at least tolerate) discussions about infectivity coefficients and case fatality rates, engaging in discussions about "flattening the curve".

The discussion that follows generates one of the simplest curves, that of the immigration-birth or contagion model. Models discussed in public are more advanced generalizations of this approach, but they all have the same weakness. Despite their mathematical elegance, each is only as good as the data used to drive the model.

They are quite elegant examples of the spread of contagion in the past, describing influenza outbreaks as well as a burst of suicides within a community [1]. In these circumstances, precise retrospective estimates of the necessary parameters were available, permitting the model to optimally project..

The corona virus experience of 2020-21 was different because it required that models perform with real time, dynamic parameter estimates. The public demanded the best information for planning, yet the best information required by the models was commonly not available. Like driving a car in the dark with dim headlights, it's not surprising that "finding the road" with initial model predictions was difficult. However, these models improved over time as their input data improved.

## Introduction to the contagion process

We begin with a community that is exposed to an infectious disease. At the beginning of the observation time period, there are already *a* subjects in the community with the disease. The disease spreads from individual to individual, its rate of spread governed by the parameter $\upsilon$, i.e., the rate of new infections generated by those already infected is $\upsilon\Delta t$. Clearly, the greater this rate, the more rapidly the disease spreads through the community.

However, another source of disease is individuals arriving into the community, by airplane, ship, train, bus, or car. These individuals arrive at the rate $\lambda\Delta t$ in time period $\Delta t$. Once in the community, they spread the infection at the same rate as those with the disease who are already in the community. It is our goal to identify the probability $\mathbf{P}_k\left[t\right]$ of the number of individuals in the community with the disease at time *t*.

For this simple model, we will assume that there are no deaths, no cures, and no people who leave the community (or emigrate). With our simplifying assumptions, the proportion of patients with the disease will only increase,[*] an increase based solely on the immigration and birth rates.

However, even in this relatively straightforward process, computing $\mathbf{P}_k\left[t\right]$ can seem like a daunting task with the disease being regularly transmitted from one community member to another (who themselves can spread the disease deeper into the populace) in addition to new arrivals who are equally capable of spreading disease. To help manage this process, we will look at what can happen to the disease in a small sliver of time that we denote $\Delta t$.

In our analysis, we will drive the $\Delta t$ to be so small that only one of two events can happen, either 1) one and only one new case of disease appears or 2) no new cases appears.

Continuing with this scenario, there are two processes that could produce a single new case in time $\Delta t$. One way would be as a new arrival to the community. The probability of this event in the small time interval $\Delta t$ is $\lambda\Delta t$. However, it is also possible that a new case occurs from the cases already present in the population. In this small interval of time, the probability of the disease being spread to a new individual is $\upsilon\Delta t$.

With this as background, we can now address the question, if there are *k* affected subjects in the community at time *t*, what is the probability that exactly one of them spreads the disease to a susceptible individual and hence generate a new case? If we assume that each affected individual spreads the disease independently of the other, we can use the binominal distribution to find.

$$\mathbf{P}\left[\text{exactly 1 additional case}\right] = \binom{k}{1}\left(\upsilon\Delta t\right)\left(1-\upsilon\Delta t\right)^{k-1}.$$

[*] More realistic discussions are referred to at the end of this section.

We may invoke the binomial theorem on the term $(1-\upsilon\Delta t)^{k-1}$ to write

$$\begin{aligned}(1-\upsilon\Delta t)^{k-1} &= \sum_{j=0}^{k-1}\binom{k-1}{j}(-1)^j(\upsilon\Delta t)^j \\ &= 1-\binom{k-1}{1}\upsilon\Delta t+\binom{k-1}{2}(\upsilon\Delta t)^2-\binom{k-1}{3}(\upsilon\Delta t)^3+....\end{aligned}$$

Therefore

$$\mathbf{P}[\text{exactly 1 additional case}] = \binom{k}{1}(\upsilon\Delta t)(1-\upsilon\Delta t)^{k-1}$$

$$= k(\upsilon\Delta t)\left(1-\binom{k-1}{1}\upsilon\Delta t+\binom{k-1}{2}(\upsilon\Delta t)^2-\binom{k-1}{3}(\upsilon\Delta t)^3+..\right).$$

However, since $\upsilon\Delta t$ is so small, we can safely assume that higher order terms are negligible and can be ignored. We therefore write $\mathbf{P}[\text{exactly 1 additional case}] \sim k(\upsilon\Delta t)$. Thus the more subjects with the disease, the greater the likelihood that an additional patient will contract the illness. This is the “birth” part of the immigration-birth process.

It is essential to understand the distinction between these two ways to produce a case. If the newly appearing case is due to arrival, then the probability of one additional case is $\lambda\Delta t$. if by the “birth” or spread of the infection when there are *k* cases already in the community, then the probability is $k\upsilon\Delta t$.

The driving force for spreading the number of cases in the community is really powered by the number of cases already there. This “birth force” grows ever stronger as the number of cases in the system increases.

Immigration, on the other hand is an influence that is not proportional to the number of cases outside the community. This leads us to believe that, even if $\lambda$ and $\upsilon$ are close, the major source of the new cases will still come from within the community, since the driving force is not $\upsilon$ but $k\upsilon$.

With this result, we can ask, how can we progress to their being *k* affected individuals or cases in the system at time $t+\Delta t$ if the time interval $\Delta t$ is so small that we only permit one event to occur in this interval? Since we have cases in the system at time 0, there are only two ways for this to happen; 1) there are *k* − 1 in the system at time *t* and a new case occurred in interval $t+\Delta t$, either by an arrival, or by its spread from an infected person to an uninfected patient, or 2) there were already *k* patients in the system and no new cases occurred during this time interval. We formally write this as.

$$\mathbf{P}_k[t+\Delta t] = \mathbf{P}_{k-1}[t]\lambda\Delta t+\mathbf{P}_{k-1}[t](k-1)\upsilon\Delta t+\mathbf{P}_k[t](1-k\upsilon\Delta t-\lambda\Delta t).$$

This holds for $k \geq a$ where *a* is the number of patients in the community at time $t=0$. We can reformulate the previous equation as

$$\frac{\mathbf{P}_k[t+\Delta t]-\mathbf{P}_k[t]}{\Delta t} = \lambda\mathbf{P}_{k-1}[t]-\lambda\mathbf{P}_k[t]+(k-1)\upsilon\mathbf{P}_{k-1}[t]-k\upsilon\mathbf{P}_k[t].$$

It we let $\Delta t$ get smaller and smaller, taking the limit as $\Delta t$ goes to zero, we write

$$\lim_{\Delta t\to 0}\frac{\mathbf{P}_k\left[t+\Delta t\right]-\mathbf{P}_k\left[t\right]}{\Delta t}$$
$$=\frac{d\mathbf{P}_k\left[t\right]}{dt}=\lambda\mathbf{P}_{k-1}\left[t\right]-\lambda\mathbf{P}_k\left[t\right]+\left(k-1\right)\upsilon\mathbf{P}_{k-1}\left[t\right]-k\upsilon\mathbf{P}_k\left[t\right]$$

that is

$$\frac{d\mathbf{P}_k\left[t\right]}{dt}=\lambda\mathbf{P}_{k-1}\left[t\right]-\lambda\mathbf{P}_k\left[t\right]+\left(k-1\right)\upsilon\mathbf{P}_{k-1}\left[t\right]-k\upsilon\mathbf{P}_k\left[t\right]\qquad k\geq a$$

We will use the generating function approach to collapse this infinite collection of equations into one equation. Let's define $\mathbf{G}_s(t)=\sum_{k=a}^{\infty}s^k\mathbf{P}_k\left[t\right]$. Ultimately we want to find the value of $\mathbf{P}_k\left[t\right]$ for $k\geq a$. Multiplying each side of the previous equation by $s^k$ produces

$$s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}=\lambda s^k\mathbf{P}_{k-1}\left[t\right]-\lambda s^k\mathbf{P}_k\left[t\right]+\left(k-1\right)s^k\upsilon\mathbf{P}_{k-1}\left[t\right]\text{-}ks^k\upsilon\mathbf{P}_k\left[t\right]\quad k\geq a.$$

From our work on the immigration process, we see that this produces.

$$\frac{\partial\mathbf{G}_s(t)}{\partial t}=\lambda\left(s-1\right)\mathbf{G}_s(t)+\upsilon s\left(s-1\right)\frac{\partial\mathbf{G}_s(t)}{\partial s}.$$

Thus we have collapsed an infinite number of equations into one. While in general, these equations can be a challenge to solve, we will see that we can solve this equation without too much difficulty.

This type of partial differential equation has been discussed. A useful tool for the solution of differential equations is that, it one is written as

$$P\frac{\partial F(x,y)}{\partial x}+Q\frac{\partial F(x,y)}{\partial y}=R,$$

Then a useful collection of equalities is

$$\frac{dx}{P}=\frac{dy}{Q}=\frac{dF}{R}.$$

Our equation, written as

$$\frac{\partial\mathbf{G}_s(t)}{\partial t}-\upsilon s\left(s-1\right)\frac{\partial\mathbf{G}_s(t)}{\partial s}=\lambda\left(s-1\right)\mathbf{G}_s(t),$$

is of this form and therefore we may write

$$\frac{dt}{1}=\frac{ds}{-\upsilon s\left(s-1\right)}=\frac{d\mathbf{G}_s(t)}{\lambda\left(s-1\right)\mathbf{G}_s(t)}.$$

We will use these equalities in combination with the state of the system at time $t$ =0 to identify the generating function $\mathbf{G}_s(t)$ and then invert. Begin with

$$\frac{ds}{-\upsilon s\left(s-1\right)}=\frac{d\mathbf{G}_s(t)}{\lambda\left(s-1\right)\mathbf{G}_s(t)}$$
$$\frac{ds}{-\upsilon s}=\frac{d\mathbf{G}_s(t)}{\lambda\mathbf{G}_s(t)}$$
$$-\frac{\lambda}{\upsilon}\frac{ds}{s}=\frac{d\mathbf{G}_s(t)}{\mathbf{G}_s(t)}$$
$$-\frac{\lambda}{\upsilon}\int\frac{ds}{s} = \int\frac{d\mathbf{G}_s(t)}{\mathbf{G}_s(t)}.$$

Integrating, we have $\ln\left(\mathbf{G}_s(t)\right)=-\frac{\lambda}{\upsilon}\ln\left(s\right)+C,$ or $\mathbf{G}_s(t)=s^{-\frac{\lambda}{\upsilon}}\Phi\left(C\right).$

Next we work with $\frac{dt}{1}=\frac{ds}{-\upsilon s(s-1)}$;

$$\frac{dt}{1}=\frac{ds}{-\upsilon s(s-1)}$$
$$\frac{-\upsilon dt}{1}=\frac{ds}{s(s-1)}=\frac{ds}{s-1}-\frac{ds}{s}$$
$$\int-\upsilon dt=\int\frac{ds}{s-1}-\int\frac{ds}{s}$$
$$-\upsilon t=\ln\left(s-1\right)-\ln\left(s\right)=\ln\left(\frac{s-1}{s}\right)$$

Now we just manipulate the constant

$$-\upsilon t=\ln\left(\frac{s-1}{s}\right)+C_1$$
$$C_1=-\upsilon t\ln\left(\frac{s}{s-1}\right)$$
$$C_2=\frac{s}{s-1}e^{-\upsilon t}$$

We can now write $\mathbf{G}_s(t)=\Phi\left(c_1\right)=\Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right),$

Thus our work reveals $\mathbf{G}_s(t)=s^{-\frac{\lambda}{\upsilon}}\Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right).$

We can now invoke our boundary conditions to identify the final form of $\mathbf{G}_s(t)$. At $t = 0$, $k = a$, leading to

$\mathbf{G}_s(0) = s^a = s^{-\frac{\lambda}{\upsilon}}\Phi\left(e^{-\upsilon(0)}\frac{s}{s-1}\right) = s^{-\frac{\lambda}{\upsilon}}\Phi\left(\frac{s}{s-1}\right)$, or $\Phi\left(\frac{s}{s-1}\right) = s^{a+\frac{\lambda}{\upsilon}}$. Assigning $z = \frac{s}{s-1}$ produces $s = \frac{z}{z-1}$, so $\Phi(z) = \left(\frac{z}{z-1}\right)^{a+\frac{\lambda}{\upsilon}}$. Returning to nonzero values of *t* reveals

$$\mathbf{G}_s(t) = \Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right) = \left[\frac{e^{-\upsilon t}\frac{s}{s-1}}{e^{-\upsilon t}\frac{s}{s-1}-1}\right]^{a+\frac{\lambda}{\upsilon}}$$

Simplifying, we find.

$$\left[\frac{e^{-\upsilon t}\frac{s}{s-1}}{e^{-\upsilon t}\frac{s}{s-1}-1}\right]^{a+\frac{\lambda}{\upsilon}} = \left[\frac{se^{-\upsilon t}}{se^{-\upsilon t}-(s-1)}\right]^{a+\frac{\lambda}{\upsilon}} = \left[\frac{se^{-\upsilon t}}{1-s\left(1-e^{-\upsilon t}\right)}\right]^{a+\frac{\lambda}{\upsilon}}$$

Thus

$$\mathbf{G}_s(t) = \left[\frac{se^{-\upsilon t}}{1-s\left(1-e^{-\upsilon t}\right)}\right]^{a+\frac{\lambda}{\upsilon}} s^{-\frac{\lambda}{\upsilon}} = \left[\frac{e^{-\upsilon t}}{1-s\left(1-e^{-\upsilon t}\right)}\right]^{a+\frac{\lambda}{\upsilon}} s^{a+\frac{\lambda}{\upsilon}-\frac{\lambda}{\upsilon}}$$

$$= \left[\frac{e^{-\upsilon t}}{1-s\left(1-e^{-\upsilon t}\right)}\right]^{a+\frac{\lambda}{\upsilon}} s^a.$$

But this we recognize as the probability generating function for negative binomial measure.scaledby $s^a$. Thus,

$$\mathbf{P}_k[t] = \binom{k+\frac{\lambda}{\upsilon}-1}{a+\frac{\lambda}{\upsilon}-1} e^{-\left(a+\frac{\lambda}{\upsilon}\right)\upsilon t}\left(1-e^{-\upsilon t}\right)^k \mathbf{1}_{k\geq a}.$$

Note that the influence of the immigration process is determined not just by $\lambda$, but the relative values of $\lambda$ and $\upsilon$ as determined by $\frac{\lambda}{\upsilon}$. If this ratio is large, then the contribution of the immigration parameter is very much like the number of individuals in the population at time *t* = 0. Small values of the $\frac{\lambda}{\upsilon}$ ratio reduce the impact of the immigration process. We have seen that the mean and variance of negative binomial measure are $\frac{rq}{p}$ and variance $\frac{rq}{p^2}$. In the case of the birth process, this translates to

$$\mathbf{E}[k] \quad = \left(a+\frac{\lambda}{\upsilon}\right)e^{\left(a+\frac{\lambda}{\upsilon}\right)\upsilon t}\left(1-e^{-\upsilon t}\right) - \sum_{k=0}^{a}\binom{k+\frac{\lambda}{\upsilon}-1}{a+\frac{\lambda}{\upsilon}-1}e^{-\left(a+\frac{\lambda}{\upsilon}\right)\upsilon t}\left(1-e^{-\upsilon t}\right)^k$$

An evaluation of the mean value as a function of *k* provides the expected result. The exponential growth of this mean value as a function of time that was the basis of the commonly used 2020 expression "flatten the curve." More realistic scenarios handle the arrival and departure of individuals from the community with the disease, and the occurrence of deaths among the diseased.

### *Negative binomial bridge*

Seeing that the contagion process is a negative binomial distribution, we can take advantage of everything we know about the negative binomial distribution are to apply to contagions e.e, their Bernoulli trial properties.

One interesting notion is to use the contagion process to look backward.

For example if we know that there were are patients suffering from Covid-19 at the beginning of the time process $t = 0$. We now know how to compute the probability of the number of Covid-19 patients in the community at time *t*.

But how about if we try to run the process backwards.

Suppose we know that while there are $k_2$ patients who have COVI-19 at time $t = t_2$, what is the distributionof patients who have the disease at time $t$ whete $0 \le t \le t_2$?

Let $\mathbf{P}\left[X_t = k\right]$ be the probability of *k* cases in the system at time *t*. We need to compute $\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right]$ where $0 \le t \le t_2$ and $0 \le k \le k_2$

$$\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right] = \frac{\mathbf{P}\left[X_t = k \cap X_{t_2} = k_2\right]}{\mathbf{P}\left[X_{t_2} = k_2\right]}.$$

Using the Bernoulli trial propeert we write

$$\mathbf{P}\left[X_t = k \cap X_{t_2} = k_2\right] = \mathbf{P}\left[X_t = k \cap X_{t_2 - t} = k_2 - k\right]$$

Let's first write $\mathbf{P}\left[X_t = k\right] = \binom{k + r - 1}{r - 1} p^r q^k$. We will first assume that *p* is time invariant.

Proceding

$$\mathbf{P}\left[X_t = k \cap X_{t_2 - t} = k_2 - k\right] = \mathbf{P}\left[X_t = k\right]\mathbf{P}\left[X_{t_2 - t} = k_2 - k\right], \text{ and}$$

$$\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right] = \frac{\mathbf{P}\left[X_t = k\right]\mathbf{P}\left[X_{t_2 - t} = k_2 - k\right]}{\mathbf{P}\left[X_{t_2} = k_2\right]}.$$

The three probabilities on the right are negative binomial probabilities. Thus

$$\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right] = \frac{\binom{k+r_1-1}{r_1-1} p^{r_1} q^k \binom{k_2-k+r_2-r_1-1}{r-1} p^{r_2-r_1} q^{k_2-k}}{\binom{k_2+r_2-1}{r_2-1} p^{r_2} q^{k_2}}.$$

In this circumstance, this is a simple counting problem.

$$\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right] = \frac{\binom{k+r_1-1}{r_1-1}\binom{k_2-k+r_2-r_1-1}{r-1}}{\binom{k_2+r_2-1}{r_2-1}}.$$

The first quotient is a constant times a hypergeometric probability hypergeometric probability, for $0 \le k \le k_2$, $0 \le r_1 \le r_2$. Examing the quotient of exponents, we see for the contagion model,

$$\frac{p^{r_1} q^k p^{r_2-r_1} q^{k_2-k}}{p^{r_2} q^{k_2}} = \frac{e^{-\upsilon r_1 t}\left(1-e^{-\upsilon t}\right)^k e^{-\upsilon(r_2-r_1)(t_2-t)}\left(1-e^{-(t_2-t)}\right)^{k_2-k}}{e^{-\upsilon r_2 t_2}\left(1-e^{-\upsilon t_2}\right)^{k_2}}$$

$$= \frac{e^{-\upsilon r_1 t} e^{-\upsilon(r_2-r_1)(t_2-t)}}{e^{-\upsilon r_2 t_2}} \frac{\left(1-e^{-\upsilon t}\right)^k \left(1-e^{-(t_2-t)}\right)^{k_2-k}}{\left(1-e^{-\upsilon t_2}\right)^{k_2}}$$

$$= e^{\upsilon r_2 t_2 - \upsilon r_1 t - \upsilon r_2 t_2 + \upsilon r_2 t + \upsilon r_1 t_2 - \upsilon r_1 t}\left[\frac{\left(1-e^{-\upsilon t}\right)}{\left(1-e^{-(t_2-t)}\right)}\right]^k \left[\frac{\left(1-e^{-(t_2-t)}\right)}{\left(1-e^{-\upsilon t_2}\right)}\right]^{k_2}$$

$$= e^{-\upsilon(r_2-r_1)(t_2-t)}\left[\frac{\left(1-e^{-\upsilon t}\right)}{\left(1-e^{-(t_2-t)}\right)}\right]^k \left[\frac{\left(1-e^{-(t_2-t)}\right)}{\left(1-e^{-\upsilon t_2}\right)}\right]^{k_2}$$

## Other contagion findings

### *Difference in cases between clinics*

Let's assume that we have two hospital. Each of which are accepting patients with COVID-19 infections. What's the probability that at any given point in time, the number of infections admitted at one hospital hospital *X* is greater than the number admitted by hospital *Y*.

Our intuition tells us that the answer is going to be related to both the. Infectivity rates $\upsilon$, the arrival rate $\lambda$, and the number of cases each has had at the beginning of the pandemic *a*.

We have several ways to examine this question. One way to simply produce the expected value and variance of the difference, $X - Y$. We know that if X follows a negative binomial distribution with parameters $r_X$ and variance $p_X$, then $\mathbf{E}\left[X\right] = \frac{r_X q_X}{p_X}$ and $\mathbf{Var}\left[X\right] = \frac{r_X q_X}{p_X^2}$. Then

$$\mathbf{E}\left[X-Y\right]=\mathbf{E}\left[X\right]-\mathbf{E}\left[Y\right]\;=\;\frac{r_Xq_X}{p_X}-\frac{r_Yq_Y}{p_Y}$$
$$=\left(a_X+\frac{\lambda_X}{\upsilon_X}\right)e^{\upsilon_X t}\left(1-e^{-\upsilon_X t}\right)\;-\left(a_Y+\frac{\lambda_Y}{\upsilon_Y}\right)e^{\upsilon_Y t}\left(1-e^{-\upsilon_Y t}\right).$$

If wr assume that the contagin parameter is the same across the two hospitals then $\upsilon_X=\upsilon_Y=\upsilon$, then

$$\mathbf{E}\left[X-Y\right]=\left[\left(a_X+\frac{\lambda_X}{\upsilon}\right)-\left(a_Y+\frac{\lambda_Y}{\upsilon}\right)\right]e^{\upsilon t}\left(1-e^{-\upsilon t}\right).$$

The last line produced under the assumption that $\upsilon_X=\upsilon_Y=\upsilon$.

We can also approach this problem in the difference in hospital rates by computing directly $\mathbf{P}\left[X>Y\right]$. Lets assume that the parameters are the same for the two distributions. Then, $\mathbf{P}\left[X>Y\right]=\mathbf{P}\left[Y>X\right]$. We can write

$$1=\mathbf{P}\left[X>Y\right]+\mathbf{P}\left[Y>X\right]+\mathbf{P}\left[X=Y\right]$$
$$=2\mathbf{P}\left[X>Y\right]+\mathbf{P}\left[X=Y\right]$$

and $\mathbf{P}\left[X>Y\right]=\dfrac{1-\mathbf{P}\left[X=Y\right]}{2}$. We can write $\mathbf{P}\left[X=Y\right]$ as

$$\mathbf{P}\left[X=Y\right]=\sum_{k=0}^{\infty}\mathbf{P}\left[X=k\right]\mathbf{P}\left[Y=k\right]$$
$$=\sum_{k=0}^{\infty}\left(\begin{pmatrix}k+a+\frac{\lambda}{\upsilon}-1\\ a+\frac{\lambda}{\upsilon}-1\end{pmatrix}e^{-\left(a+\frac{\lambda}{\upsilon}\right)\upsilon t}\left(1-e^{-\upsilon t}\right)^k\right)^2.$$

If the parameters of the two distributions are not equal, then we try to compute $\mathbf{P}\left[X>Y\right]$ direct. Whatever value $X$ takes then $Y$ must be nonnegative and less than that value. Thus

$$\mathbf{P}\left[X>Y\right]=\sum_{k=1}^{\infty}\mathbf{P}\left[X=k\right]\sum_{j=0}^{k-1}\mathbf{P}\left[Y=j\right]$$
$$=\sum_{k=0}^{\infty}\begin{pmatrix}k+a_X+\frac{\lambda_X}{\upsilon_X}-1\\ a_X+\frac{\lambda_X}{\upsilon_X}-1\end{pmatrix}e^{-\left(a_X+\frac{\lambda_X}{\upsilon_X}\right)\upsilon_X t}\left(1-e^{-\upsilon_X t}\right)^k\sum_{j=0}^{k-1}\begin{pmatrix}j+a_Y+\frac{\lambda_Y}{\upsilon_Y}-1\\ a_Y+\frac{\lambda_Y}{\upsilon_Y}-1\end{pmatrix}e^{-\left(a_Y+\frac{\lambda_Y}{\upsilon_Y}\right)\upsilon_Y t}\left(1-e^{-\upsilon_Y t}\right)^j.$$

### *Conditional negative binomial*

Assume two states *X* and *Y*, report a total of *n* patients with Covid-19.what is the probability that *k* of them are from state *X*? This is "looking backward" in the negative binomial process.

We begin the computation of this conditional probability $\mathbf{P}\left[X=k \mid X+Y=n\right]$.

$$\mathbf{P}\left[X=k \mid X+Y=n\right]=\frac{\mathbf{P}\left[X=k\cap X+Y=n\right]}{\mathbf{P}\left[X+Y=n\right]}.$$

We write the numerator as

$$\mathbf{P}\left[X=k\cap X+Y=n\right]=\mathbf{P}\left[X=k\cap Y=n-k\right]=\mathbf{P}\left[X=k\right]\mathbf{P}\left[Y=n-k\right]$$
$$=\binom{k+a_X+\frac{\lambda_X}{\upsilon_X}-1}{a_X+\frac{\lambda_X}{\upsilon_X}-1}e^{-\left(a_X+\frac{\lambda_X}{\upsilon_X}\right)\upsilon_X t}\left(1-e^{-\upsilon_X t}\right)^k\binom{n-k+a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}{a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}e^{-\left(a_Y+\frac{\lambda_Y}{\upsilon_Y}\right)\upsilon_Y t}\left(1-e^{-\upsilon_Y t}\right)^{n-k}$$

We compute $\mathbf{P}\left[X+Y=n\right]$ as $\sum_{k=0}^{n}\mathbf{P}\left[X=k\right]\mathbf{P}\left[Y=n-k\right]$. We write this as

$$=\sum_{k=0}^{n}\binom{k+a_X+\frac{\lambda_X}{\upsilon_X}-1}{a_X+\frac{\lambda_X}{\upsilon_X}-1}e^{-\left(a_X+\frac{\lambda_X}{\upsilon_X}\right)\upsilon_X t}\left(1-e^{-\upsilon_X t}\right)^k\binom{n-k+a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}{a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}e^{-\left(a_Y+\frac{\lambda_Y}{\upsilon_Y}\right)\upsilon_Y t}\left(1-e^{-\upsilon_Y t}\right)^{n-k}$$

And our solution is

$$\mathbf{P}\left[X=k \mid X+Y=n\right]$$
$$=\frac{\binom{k+a_X+\frac{\lambda_X}{\upsilon_X}-1}{a_X+\frac{\lambda_X}{\upsilon_X}-1}e^{-\left(a_X+\frac{\lambda_X}{\upsilon_X}\right)\upsilon_X t}\left(1-e^{-\upsilon_X t}\right)^k\binom{n-k+a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}{a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}e^{-\left(a_Y+\frac{\lambda_Y}{\upsilon_Y}\right)\upsilon_Y t}\left(1-e^{-\upsilon_Y t}\right)^{n-k}}{=\sum_{k=0}^{n}\binom{k+a_X+\frac{\lambda_X}{\upsilon_X}-1}{a_X+\frac{\lambda_X}{\upsilon_X}-1}e^{-\left(a_X+\frac{\lambda_X}{\upsilon_X}\right)\upsilon_X t}\left(1-e^{-\upsilon_X t}\right)^k\binom{n-k+a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}{a_Y+\frac{\lambda_Y}{\upsilon_Y}-1}e^{-\left(a_Y+\frac{\lambda_Y}{\upsilon_Y}\right)\upsilon_Y t}\left(1-e^{-\upsilon_Y t}\right)^{n-k}}$$

### *Conditioning on the future – the memoryless process*

One interesting conditional probability problem involving the negative binomial distribution is conditioning on the future.

Let's suppose that we have at time point $t_1$, $n_1$ cases in a county. We eould like to know the probability that at some time in the future $t_2$ of having $n_2$ cases in the county.

Answering this question will bring us face to face to one of the most important characteristics of bernoulli trials – the markovian property. The markovian property essentially says that in order to condition on the future one simply needs to know the current state of the process and not the entire history of the process.

Let's assume that there are 63 Covid-19 cases in our county at mid month and we want to know how likely it is that we will have 200 cases by the end of the month. This problem is indeed very complicated if we have to keep track of not just how many midmonth cases there are, but also how many patients there were at 10 days into the month or five days into the month or one day end of the month. The problem is more difficult because the probability condition is more complicated.

However if the probability of having $n_2$ patients at time $t_2$ is based simply on the most recent observation (63 patients at midmonth) the problem simplifies at once. This property is the memoryless property. It plays a central role in many stochastic processes such as Brownian motion.

We presume we are at some time $t_1 > 0,$ there are $n_1$ patients with Covid-19 at that time. Then $X(t_1) = n_1$. What we need to calculate is $\mathbf{P}\left[X(t_2) = n_2 \mid X(t_1) = n_1\right]$. We begin with

$$\mathbf{P}\left[X(t_2) = n_2 \mid X(t_1) = n_1\right] = \frac{\mathbf{P}\left[X(t_2) = n_2 \cap X(t_1) = n_1\right]}{\mathbf{P}\left[X(t_1) = n_1\right]}.$$

Examining the numerator, $\mathbf{P}\left[X(t_2) = n_2 \cap X(t_1) = n_1\right]$we ask how do we get to $X(t_2) = n_2$? There is only one way, and that is to use the remaining time $t_2 - t_1$ to obtain $n_2 - n_1$ cases. Thus

$$\mathbf{P}\left[X(t_2) = n_2 \cap X(t_1) = n_1\right] = \mathbf{P}\left[X(t_2 - t_1) = n_2 - n_1 \cap X(t_1) = n_1\right].$$

However we assume Covid-19 cases arrive independently, so the arrival pattern of cases after $t_1$ is independent of the arrival before $t_1$. Similarly, the arrival pattern of cases up to some time prior to $t_{1,}$ sat $t_a$ such that $0 < t_a < t_1$ is independent of the cases that occur after time$t_a$. Thus, $\mathbf{P}\left[X(t_1) = n_1 \mid X(t_a) = n_a\right]$ is independent of the the pattern of prior arrival of cases before time $t_a$. This is the heart of the Markov process.

Acknowledging this property affords us a great simplification. We can now write

$$\mathbf{P}\left[X(t_2) = n_2 \mid X(t_1) = n_1\right] = \frac{\mathbf{P}\left[X(t_2) = n_2 \cap X(t_1) = n_1\right]}{\mathbf{P}\left[X(t_1) = n_1\right]}$$

$$= \frac{\mathbf{P}\left[X(t_2 - t_1) = n_2 - n_1 \ \cap \ X(t_1) = n_1\right]}{\mathbf{P}\left[X(t_1) = n_1\right]}$$

$$= \frac{\mathbf{P}\left[X(t_2 - t_1) = n_2 - n_1\right]\mathbf{P}\left[X(t_1) = n_1\right]}{\mathbf{P}\left[X(t_1) = n_1\right]}$$

$$= \mathbf{P}\left[X(t_2 - t_1) = n_2 - n_1\right].$$

But this last expression is simply a negative binomial probabiliy with probability running from time $t = 0$ to time $t = t_2 - t_1$. Note that we have essentially restarted the process. At this new $t = 0,$ we have $n_1$ cases in the system. Assuming our usual parametetization of the death ptocess, we find

$$\mathbf{P}\left[X\left(t_2\right)=n_2 \mid X\left(t_1\right)=n_1\right]=\mathbf{P}\left[X\left(t_2-t_1\right)=n_2-n_1\right]$$

$$=\binom{n_2-n_1+n_1+\dfrac{\lambda}{\upsilon}-1}{n_1+\dfrac{\lambda}{\upsilon}-1} e^{-\upsilon\left(n_1+\frac{\lambda}{\upsilon}-1\right)\left(t_2-t_1\right)}\left(1-e^{-\upsilon\left(t_2-t_1\right)}\right)^{\left(n_2-n_1\right)}$$

$$=\binom{n_2+\dfrac{\lambda}{\upsilon}-1}{n_1+\dfrac{\lambda}{\upsilon}-1} e^{-\upsilon\left(n_1+\frac{\lambda}{\upsilon}-1\right)\left(t_2-t_1\right)}\left(1-e^{-\upsilon\left(t_2-t_1\right)}\right)^{\left(n_2-n_1\right)}.$$

Death Process
The Emigration-Death Process
Immigration-Death Process
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

References

1. Davis BR, Hardy RJ. A suicide epidemic model. *Soc Biol*. 1986;33(3-4):291–300. doi:10.1080/19485565.1986.9988646

# The Death Process

We have described how the binomial distribution arises from the consideration of the sums of independently and identically distributed Bernoulli trials. It's use can commonly be predicted from consideration of the construction of the event.

However, an intriguing observation of intrinsic processes is that from these systems can spring a probability distribution that did not suggest its appearance earlier in the problem's consideration. Such is the case with the death process.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions
Negative binomial measure
Binomial distribution
Binomial probability generating function

## Introduction to the death process

The death process describes a system in which events occur which decrease the number of individuals or objects over time. However, so did the emigration process. What is the difference?

## Distinguishing emigration from death

The death process, like the emigration process requires no arrivals into the community, only removals. However, the difference between the emigration process and the death process is profound.

In the emigration process, arrivals of individuals with the disease occur at a constant average rate, regardless of the number of diseased individuals in the community. Whether there are 10 individuals, or 1010 individuals in the community, the emigration rate is the same.

The death process (which might actually be unhelpfully named) describes a process in which removals occur, but their rate of removal is dependent upon the number of diseased individuals in the community. Removals occur more quickly when there are many more diseased subjects in the community then not.

Thus the death process's rate dependent removal is more like the birth process whose disease dissemination rate is proportional to the number of individuals in the community.

From a public health perspective, both the emigration process and the "death" process can describe the occurrence of deaths. In a large hospice facility that cares for patients at the end of their lives, an emigration process may aptly predict the number of deaths in a given time. Deaths are not related to the number in the hospice system. As long as there are patients in the hospice unit, deaths occur at a constant rate.

However consider a circumstance in an isolated Emergency Department in an underserved community, as the SARS-CoV-2 virus spreads through the community. The number of critically ill subjects increases, rapidly consuming the resources required for their care. In this circumstance, the larger the number of these patients in the system, the fewer resources there are to care for them, and the "departure" rate climbs (either through actual deaths or transfers).[*] This is a process where the departures are actually proportional to the number of subjects in the clinic.

Another example of emigration rates that are accelerated by the number in the system would be a queuing arrangement when the servers work faster when the number in the system is greater, and decrease their performance rate (throughput) as the number in the system decline. A "death process" characterizes this system as well.

## Underlying assumptions

Assume that we are in a facility that is experiencing departures/removals in accordance with a death process. The departure/removal rate in at this facility is given by $\omega$ for a period of time $t$. For example $\omega$ may be 3 departures per week. Our goal is to find the probability distribution for the number of remaining patients in the facility at time $t$, $\mathbf{P}_k\left[t\right]$. We assume that at the start of the observation process, that there are $a$ subjects in the hospice unit.

## Developing the difference equations

Our approach will be to observe the possible changes over time when the time interval that has passed is very small. This small change in time constrains the number of possible events that can take place to a mathematically manageable level. As before, we will develop a collection of difference-differential equations, collapse them to one equation that is based on a probability generating function, solve it, and then invert the generating function.

Let's begin by watching this process transpire over a very small period of time. In fact, this will be a vanishing small period of time, that we will call $\Delta t$. We demand that $\Delta t$ be so small, that two patients cannot die within it. Therefore, only two events can occur. Either a single patient dies, or they do not. In this small slice of time, the probability of departure for each of the patients is $\omega\Delta t$. What is the probability of a single departure in this small time interval?

If we assume that there are $k$ patients at risk of dying at time $t$, what is the probability that exactly one of them dies in time period $t+\Delta t$? Assuming that departures are independently of each other, we can use the binomial distribution to compute.

[*] This is another reason why the descriptor "death" is not apropos. Perhaps a better descriptor would be "powered emigration" versus "static emigration".

$$\mathbf{P}\left[\text{exactly 1 death}\right] = \binom{k}{1}(\omega\Delta t)(1-\omega\Delta t)^{k-1}.$$

Using the binomial theorem, we write

$$(1-\omega\Delta t)^{k-1} = \sum_{j=0}^{k-1}\binom{k-1}{j}(-\omega\Delta t)^{j}$$
$$= 1 - \binom{k-1}{1}\omega\Delta t + \binom{k-1}{2}(\omega\Delta t)^{2} - \binom{k-1}{3}(\omega\Delta t)^{3} + \ldots.$$

Therefore

$$\mathbf{P}\left[\text{exactly 1 death}\right] = \binom{k}{1}(\omega\Delta t)(1-\omega\Delta t)^{k-1}$$
$$= k(\omega\Delta t)\left(1 - \binom{k-1}{1}\omega\Delta t + \binom{k-1}{2}(\omega\Delta t)^{2} - \binom{k-1}{3}(\omega\Delta t)^{3} + ..\right).$$

However, given that $\omega\Delta t$ is so small, we can safely assume that higher order terms are negligible and can be safely ignored. We therefore write

$$\mathbf{P}\left[\text{exactly 1 death}\right] \sim k(\omega\Delta t).$$

With this result, we can ask, how can we progress to there being *k* patients in the system at time $t+\Delta t$ if the time interval $\Delta t$ is so small that we only permit one event to occur in this interval. For $0 \le k \le a-1$, there are only two ways for this to happen; 1) there are *k*+1 patients in the system at time *t* and a departure occurred in interval $t+\Delta t$, or 2) there were already *k* patients in the system and no departure occurred during this time interval. We formally write this as.

$$\mathbf{P}_k\left[t+\Delta t\right] = \mathbf{P}_{k+1}\left[t\right](k+1)\omega\Delta t + \mathbf{P}_k\left[t\right](1-k\omega\Delta t).$$

This holds for $0 \le k \le a-1$. Note that this is different than the statement for the emigration model in which the force of emigration was not dependent on the current number in the system. In the death model, the greater the number of objects in the system, the greater the likelihood that one will depart.

We can reformulate the previous equation as

$$\frac{\mathbf{P}_k\left[t+\Delta t\right]-\mathbf{P}_k\left[t\right]}{\Delta t} = (k+1)\omega\,\mathbf{P}_{k+1}\left[t\right] - k\omega\,\mathbf{P}_k\left[t\right].$$

It we let $\Delta t$ get smaller and smaller, taking the limit as $\Delta t$ goes to zero, we write

$$\lim_{\Delta t\to 0}\frac{\mathbf{P}_k\left[t+\Delta t\right]-\mathbf{P}_k\left[t\right]}{\Delta t} = \frac{d\mathbf{P}_k\left[t\right]}{dt} = (k+1)\omega\mathbf{P}_{k+1}\left[t\right] - k\omega\mathbf{P}_k\left[t\right]$$

that is

$$\frac{d\mathbf{P}_k\left[t\right]}{dt} = (k+1)\omega\mathbf{P}_{k+1}\left[t\right] - k\omega\mathbf{P}_k\left[t\right] \quad 0 \le k \le a-1$$

For $\mathbf{P}_a\left[t+\Delta t\right]$ we have

$$\mathbf{P}_a\left[t+\Delta t\right]=\mathbf{P}_a\left[t\right]\left(1-a\omega\Delta t\right).$$

or

$$\lim_{\Delta t\to 0}\frac{\mathbf{P}_a\left[t+\Delta t\right]-\mathbf{P}_a\left[t\right]}{\Delta t}=-a\omega\mathbf{P}_a\left[t\right]$$

$$\frac{d\mathbf{P}_a\left[t\right]}{dt}=-a\omega\mathbf{P}_a\left[t\right].$$

Thus, our system of difference-differential equations is

$$\frac{d\mathbf{P}_k\left[t\right]}{dt}=\left(k+1\right)\omega\mathbf{P}_{k+1}\left[t\right]\text{-}k\omega\mathbf{P}_k\left[t\right]\quad 0\le k\le a-1$$

$$\frac{d\mathbf{P}_a\left[t\right]}{dt}=-a\omega\mathbf{P}_a\left[t\right].$$

We will use the generating function approach to collapse this system of a equations into one equation.

Let's define $\mathbf{G}_s(t)=\sum_{k=0}^{a}s^k\mathbf{P}_k\left[t\right]$. Ultimately we want to find the value of $\mathbf{P}_k\left(t\right)$ for $0\le k\le a;$ Multiplying each side of the above equations by $s^k$ produces

$$s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}=\left(k+1\right)s^k\omega\mathbf{P}_{k+1}\left[t\right]\text{-}ks^k\omega\mathbf{P}_k\left[t\right]\quad 0\le k\le a-1$$

and summing over the relevant values of *k* gives

$$\sum_{k=-0}^{a-1}s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}=\omega\sum_{k=0}^{a-1}\left(k+1\right)s^k\mathbf{P}_{k+1}\left[t\right]\text{-}\omega\sum_{k=0}^{a-1}ks^k\mathbf{P}_k\left[t\right].$$

The next equation simply becomes

$$s^a\frac{d\mathbf{P}_a\left[t\right]}{dt}=-\omega a\,s^a\mathbf{P}_a\left[t\right].$$

Adding these two produces

$$\sum_{k=0}^{a}s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}=\omega\sum_{k=0}^{a-1}\left(k+1\right)s^k\mathbf{P}_{k+1}\left[t\right]\text{-}\omega\sum_{k=0}^{a}ks^k\mathbf{P}_k\left[t\right]$$

We will simplify each of these, term by term. For $\sum_{k=0}^{a}s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}$ we have

$$\sum_{k=0}^{a}s^k\frac{d\mathbf{P}_k\left[t\right]}{dt}=\sum_{k=0}^{a}\frac{ds^k\mathbf{P}_k\left[t\right]}{dt}=\frac{d}{dt}\sum_{k=0}^{a}s^k\mathbf{P}_k\left[t\right]=\frac{d\mathbf{G}_s(t)}{dt}$$

$$\omega\sum_{k=0}^{a-1}\left(k+1\right)s^k\mathbf{P}_{k+1}\left[t\right]=\omega\sum_{k=0}^{a-1}\frac{ds^{k+1}}{ds}\mathbf{P}_{k+1}\left[t\right]$$

$$=\omega\sum_{k=1}^{a}\frac{ds^k}{ds}\mathbf{P}_k\left[t\right]=\omega\sum_{k=0}^{a}\frac{ds^k}{ds}\mathbf{P}_k\left[t\right]-\frac{ds^0}{ds}\mathbf{P}_0\left[t\right]$$

We will assume that $\mathbf{P}_0\left(t\right)=0.$[*] proceeding;

$$=\omega\sum_{k=0}^{a}\frac{ds^k}{ds}\mathbf{P}_k\left[t\right]=\omega\sum_{k=0}^{a}\frac{ds^k\mathbf{P}_k\left[t\right]}{ds}=\omega\frac{d\sum_{k=0}^{a}s^k\mathbf{P}_k\left[t\right]}{ds}$$

$$=\omega\frac{d\mathbf{G}_s(t)}{ds}$$

And $\omega\sum_{k=0}^{a}ks^k\mathbf{P}_k\left[t\right]$is simplified as follows.

$$\omega\sum_{k=0}^{a}ks^k\mathbf{P}_k\left[t\right]=\omega s\sum_{k=0}^{a}ks^{k-1}\mathbf{P}_k\left[t\right]=\omega s\sum_{k=0}^{a}\frac{ds^k}{ds}\mathbf{P}_k\left[t\right]$$

$$=\omega s\sum_{k=0}^{a}\frac{ds^k\mathbf{P}_k\left[t\right]}{ds}\quad=\omega s\frac{d\sum_{k=0}^{a}s^k\mathbf{P}_k\left[t\right]}{ds}=\ \omega s\frac{d\mathbf{G}_s(t)}{ds}.$$

Thus, we have

$$\frac{d\mathbf{G}_s(t)}{dt}=\ \omega\frac{d\mathbf{G}_s(t)}{ds}-\omega s\frac{d\mathbf{G}_s(t)}{ds}=\omega\left(1-s\right)\frac{d\mathbf{G}_s(t)}{ds}$$

But since the derivatives are with respect to two different variables (*s* and *t*) we write

$$\frac{\partial\mathbf{G}_s(t)}{\partial t}=\ \omega\left(1-s\right)\frac{\partial\mathbf{G}_s(t)}{\partial s}.$$

This is the difference-differential equation that we must solve.

We have collapsed *a* equations into one partial differential equation. While in general, these equations can be a challenge to solve, this one is relatively easy.

Recall that if a differential equation is written as

$$P\frac{\partial F(x,y)}{\partial x}+Q\frac{\partial F(x,y)}{\partial y}=R,$$

then

$$\frac{dx}{P}=\frac{dy}{Q}=\frac{dF}{R}.$$

Our equation, written as $\frac{\partial\mathbf{G}_s(t)}{\partial t}-\ \omega\left(1-s\right)\frac{\partial\mathbf{G}_s(t)}{\partial s}=0$ is of this form which implies that

[*] This is an interesting assumption. Even though only deaths can be produced, $\mathbf{P}_0\left(t\right)=0$ means one never gets to the condition that the all patients die. In animal population studies this is called the extinction probability.

$$\frac{dt}{1} = \frac{ds}{-\omega(1-s)} = \frac{d\mathbf{G}_s(t)}{0}.$$

We will use these equalities and the state of the system at time $t = 0$ to identify the generating function $\mathbf{G}_s(t)$ and then invert. Begin with $\frac{dt}{1} = \frac{d\mathbf{G}_s(t)}{0}$, or $d\mathbf{G}_s(t) = 0$, which implies $\mathbf{G}_s(t)$ is a constant. We write this constant $\Phi(c_1)$. Next we work with $\frac{dt}{1} = \frac{ds}{-\omega(1-s)}$;

$$\begin{aligned}
&\frac{dt}{1} = \frac{ds}{-\omega(1-s)} \\
&-\omega dt = \frac{ds}{(1-s)} \\
&\int -\omega dt = \int \frac{ds}{(1-s)} \\
&-\omega t = -\ln(1-s) + C \\
&-\omega t + \ln(1-s) = C \\
&C = c_2 = (1-s)e^{-\omega t}.
\end{aligned}$$

We can now write $\mathbf{G}_s(t) = \Phi(c_1) = \Phi\left((1-s)e^{-\omega t}\right)$, and examine the boundary condition for clarification of the form of the function $\Phi$. We know that there are $a$ subjects in the hospice at time $t = 0$. We can therefore write

$$\mathbf{G}_s(0) = s^a = (1-s)e^{-\omega(0)} = \Phi(1-s)$$

Let $z = 1-s$ or $s = 1-z$, and we find $\Phi(z) = (1-z)^a$. This is for $t = 0$. For any other value of $t$ we have

$$\mathbf{G}_s(t) = \Phi\left((1-s)e^{-\omega t}\right) = \left[1-(1-s)e^{-\omega t}\right]^a.$$

However, we observe $1-(1-s)e^{-\omega t} = 1-e^{-\omega t} + se^{-\omega t}$ and

$$\mathbf{G}_s(t) = \left(1-e^{-\omega t} + se^{-\omega t}\right)^a.$$

This is the generating function for the binomial distribution with probability of “success” $p = e^{-\omega t}$. We can write that if there are a patients in the hospice at time t and they die at rate $\omega$ for unit time, then the probability there are $k$ patients alive at time $t$ is

$$\mathbf{P}_k(t) = \binom{a}{k} e^{-k\omega t}\left(1-e^{-\omega t}\right)^{a-k} \mathbf{1}_{0\le k\le a}.$$

and we know from the moments of the binomial distribution that

$$\mathbf{E}[k] = ae^{-k\omega t};\ \mathbf{Var}[k] = ae^{-k\omega t}\left(1-e^{-\omega t}\right).$$

Figure 1 provides an example of the distribution of the number in the system at different times for a single value of $\omega = 0.20$.

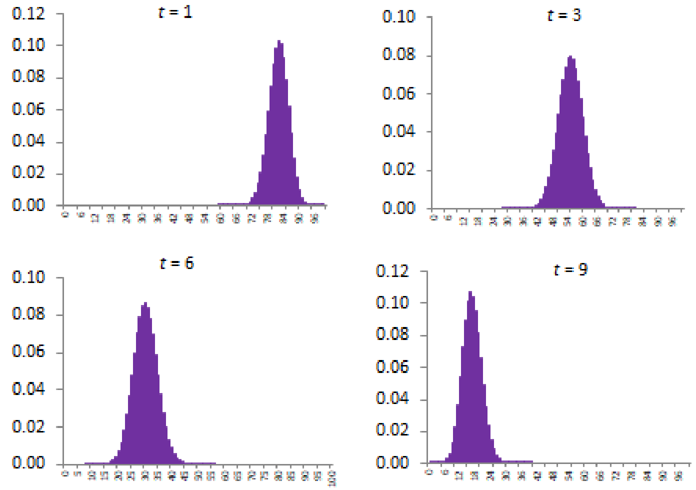


Figure 1. Distribution of the number alive in the death process as a function of time; $\omega$=0.2.

### *Binomial bridge*

Seeing that the death process is a binomial distribution, we can take advantage of everything we know about the binomial distribution are to apply to this death process. One of them is the notion of Bernoulli trials.

One interesting notion is to use the death process to look backward.

For example if we know that there were are patients suffering from Covid-19 at the beginning of the time process *t* = 0. We now know how to compute the probability of the number of Covid-19 patients who die by time *t*.

But how about if we try to run the process backwards.

Suppose we know that while there are $n_2$ patients at time $t=0$ sand that there are $k_2$ patients who have died at time $t=t_2$, what is the distributionof patients who died at $t=t_1$, whete $0 \le t_1 \le t_2$?

We would expect that the as *t* move from 0 to $t_1$, the expected number of deaths would increase; however what do we know about the variance? Estimates at the beginning of $t=0$ should be close to zero and therefore the variance should be small. However when we move away from time 0, then the variance is large as time increases. However, when we get closer and closer to time $t_2$ then, the time where the number of deaths is equal to be our estimates should get closer to be and therefore the variance should decrease.

Let $\mathbf{P}\left[X_t = k\right]$ be the probability of *k* deaths in the system at time *t*. We assume that there are $n_2$ subjects in the sysytem at time $t=0$. We need to compute $\mathbf{P}\left[X_t\left(n_1\right)=k \mid X_{t_2}\left(n_2\right)=k_2\right]$ where $0 \le t \le t_1$ and $0 \le k \le k_1$ and $0 \le n_1 \le n_2$.

$$\mathbf{P}\left[X_{t_1}\left(n_1\right)=k \mid X_{t_2}\left(n_2\right)=k_2\right]=\frac{\mathbf{P}\left[X_{t_1}\left(n_1\right)=k \cap X_{t_2}\left(n_2\right)=k_2\right]}{\mathbf{P}\left[X_{t_2}\left(n_2\right)=k_2\right]}.$$

Using the Bernoulli trial property we write

$$\mathbf{P}\left[X_t\left(n_1\right)=k \cap X_{t_2}\left(n_2\right)=k_2\right]$$
$$=\mathbf{P}\left[X_t\left(n_1\right)=k \cap X_{t_2-t}\left(n_2-n_1\right)=k_2-k\right]$$

Proceding,

$$\frac{\mathbf{P}\left[X_t=k\right]\mathbf{P}\left[X_{t_2-t}\left(n_2-n_1\right)=k_2-k\right]}{\mathbf{P}\left[X_{t_2}=k_2\right]}$$
$$=\frac{\binom{n_1}{k}e^{-\omega kt}\left(1-e^{-\omega t}\right)^{n_1-k}\binom{n_2-n_1}{k_2-k}e^{-\omega(k_2-k)(t_2-t)}\left(1-e^{-\omega(t_2-t)}\right)^{n_2-n_1-(k_2-k)}}{\binom{n_2}{k_2}e^{-\omega k_2t_2}\left(1-e^{-\omega t_2}\right)^{n_2-k_2}}$$
$$=\frac{\binom{n_1}{k}\binom{n_2-n_1}{k_2-k}}{\binom{n_2}{k_2}}\frac{e^{-\omega kt}\left(1-e^{-\omega t}\right)^{n_1-k}e^{-\omega(k_2-k)(t_2-t)}\left(1-e^{-\omega(t_2-t)}\right)^{n_2-n_1-(k_2-k)}}{e^{-\omega k_2t_2}\left(1-e^{-wt_2}\right)^{n_2-k_2}}.$$

The first quotient is a hypergeometric probability. Examing the quotient of exponents, we see

$$\mathbf{P}\left[X_t=k \mid X_{t_2}=k_2\right]$$
$$=\frac{\binom{n_1}{k}\binom{n_2-n_1}{k_2-k}}{\binom{n_2}{k_2}}\frac{e^{-\omega kt}\left(1-e^{-\omega t}\right)^{n_1-k}e^{-\omega(k_2-k)(t_2-t)}\left(1-e^{-\omega(t_2-t)}\right)^{n_2-n_1-(k_2-k)}}{e^{-\omega k_2t_2}\left(1-e^{-\omega t_2}\right)^{n_2-k_2}}$$
$$=\frac{\binom{n_1}{k}\binom{n_2-n_1}{k_2-k}}{\binom{n_2}{k_2}}\left[\frac{e^{-\omega kt}e^{-\omega(k_2-k)(t_2-t)}}{e^{-wk_2t_2}}\right]\left[\frac{\left(1-e^{-\omega t}\right)^{n_1-k}\left(1-e^{-\omega(t_2-t)}\right)^{n_2-n_1-(k_2-k)}}{\left(1-e^{-wt_2}\right)^{n_2-k_2}}\right]$$

Observe

$$\frac{e^{-\omega kt}e^{-\omega\left(k_2-k\right)\left(t_2-t\right)}}{e^{-wk_2t_2}}=e^{wk_2t_2-\omega kt-wk_2t_2+wk_2t+wkt_2-wkt}$$
$$=e^{-\omega kt+wk_2t+wkt_2-wkt}=e^{-w\left(k_2-k\right)t-wk\left(t_2-t\right)}.$$

$$\left[\frac{\left(1-e^{-\omega t}\right)^{n_1-k}\left(1-e^{-\omega(t_2-t)}\right)^{n_2-n_1-(k_2-k)}}{\left(1-e^{-wt_2}\right)^{n_2-k_2}}\right]$$
$$=\left[\frac{\left(1-e^{-\omega t}\right)}{\left(1-e^{-\omega(t_2-t)}\right)}\right]^{n_1-k}\left[\frac{\left(1-e^{-\omega(t_2-t)}\right)}{\left(1-e^{-wt_2}\right)}\right]^{n_2-k_2}.$$

And the solution is

$$\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right]$$

$$= \frac{\binom{n_1}{k}\binom{n_2 - n_1}{k_2 - k}}{\binom{n_2}{k_2}}$$

$$\mathbf{P}\left[X_t = k \mid X_{t_2} = k_2\right]$$

$$= \frac{\binom{n_1}{k}\binom{n_2 - n_1}{k_2 - k}}{\binom{n_2}{k_2}} e^{-w(k_2 - k)t - wk(t_2 - t)} \left[\frac{\left(1 - e^{-\omega t}\right)}{\left(1 - e^{-\omega(t_2 - t)}\right)}\right]^{n_1 - k} \left[\frac{\left(1 - e^{-\omega(t_2 - t)}\right)}{\left(1 - e^{-wt_2}\right)}\right]^{n_2 - k_2}$$

$$\left[\frac{\left(1 - e^{-\omega t}\right)}{\left(1 - e^{-\omega(t_2 - t)}\right)}\right]^{n_1 - k} \left[\frac{\left(1 - e^{-\omega(t_2 - t)}\right)}{\left(1 - e^{-wt_2}\right)}\right]^{n_2 - k_2} .$$

Contagion
Death Process
The Emigration-Death Process
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Birth-Death Model

## Introduction

In this section, the concept of death to the birth model is added. Each of these two processes has a force proportional to the number of cases in the population.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions
Negative binomial measure
Binomial distribution

## Motivation

The only new cases are due to births in the existing population and the only exits are due to death. The disease spreads in the population according to the birth process. However, the death process leads to a moderation in its effect since its force decreases the number of cases

## The Chapman-Kolmogorov equations

The process for developing the Chapman-Kolmogorov forward equations will be exactly analogous to their development thus far in this chapter. Begin with an enumeration of the three ways the population can have $n$ diseased patients at time. Let's allow $\Delta t$ to be so small that only one event can occur in this short period of time. Then what types of events can occur? A first would be that the number of cases in the population is $n-1$ at time $t$ and there is a "birth" or spread of the disease from time $t$ to time $t+\Delta t,$ an event occurs with probability $(n-1)\upsilon\Delta t.$[*] Alternatively, there may be $n+1$ cases of disease in the population at time $t$ and a death occurs in the population, an event which occurs with probability $(n+1)\omega\Delta t.$ Finally, there may be $n$ individuals at time $t$ and neither a birth nor a death occurs. The probability of the absence of these events is $1-n\upsilon\Delta t-n\omega\Delta t.$ We assume that $\upsilon >> \omega$ (i.e. much greater) such that at no time in the system are there are no individuals ( $\mathbf{P}_0[t]=0 \text{ for } t\geq 0.$ )

Recall that "death" does not mean death of the subject, but death of the infection. While this can mean the death of the patient it can also mean that the patient has been cured by therapy.

With this as background, the Chapman-Kolmogorov forward equations may be written as

$$\begin{aligned}&\mathbf{P}_n[t+\Delta t]\\ &=(n-1)\upsilon\,\Delta t\,\mathbf{P}_{n\text{-}1}[t]+(n+1)\omega\,\Delta t\mathbf{P}_{n+1}[t]+(1-n\,\upsilon\Delta t-n\,\omega\,\Delta t)\mathbf{P}_n[t]\end{aligned}$$

proceeding to convert equations represented by to the anticipated difference-differential equation.

$$\begin{aligned}&\mathbf{P}_n[t+\Delta t]\ -\mathbf{P}_n[t]\\ &=\ (n-1)\upsilon\,\Delta t\,\mathbf{P}_{n\text{-}1}[t]+(n+1)\omega\,\Delta t\mathbf{P}_{n+1}[t]-n\,\upsilon\Delta t-n\,\omega\,\Delta t\end{aligned}$$

Dividing by $\Delta t$ then taking a limit reveals

$$\begin{aligned}&\frac{\mathbf{P}_n[t+\Delta t]\ -\mathbf{P}_n[t]}{\Delta t}\\ &=(n-1)\upsilon\,\mathbf{P}_{n\text{-}1}[t]+(n+1)\omega\,\mathbf{P}_{n+1}[t]-n\,\upsilon\mathbf{P}_n[t]-n\,\omega\mathbf{P}_n[t]\end{aligned}$$

$$\begin{aligned}&\lim_{\Delta t\to 0}\frac{\mathbf{P}_n[t+\Delta t]\ -\mathbf{P}_n[t]}{\Delta t}\\ &=(n-1)\upsilon\,\mathbf{P}_{n\text{-}1}[t]+(n+1)\omega\,\mathbf{P}_{n+1}[t]-n\,\upsilon\mathbf{P}_n[t]-n\,\omega\,\mathbf{P}_n[t]\end{aligned}$$

Usig the definition of a derivative shows

$$\frac{\mathrm{d}\mathbf{P}_n[t]}{\mathrm{d}t}=(n-1)\upsilon\,\mathbf{P}_{n\text{-}1}[t]+(n+1)\omega\,\mathbf{P}_{n+1}[t]-n\,\upsilon\mathbf{P}_n[t]-n\,\omega\,\mathbf{P}_n[t].$$

We can now proceed with the application of the generating function.

[*] The probability of at least one new case in time $\Delta t$ is $\sum_{k=1}^{n}\binom{n}{k}(\upsilon\Delta t)^k(1-\upsilon\Delta t)^{n-k}$ which when expanded is $n\upsilon\Delta t$ plus other terms involving highrr powers of $\Delta t$ that are much smaller.

## Gen function for the birth-death model

Define the generating function as $\mathbf{G}_s\left[t\right] = \sum_{n=0}^{\infty} s^n \mathbf{P}_n\left[t\right]$ and multiplying our difference equation by $s^n$ reveals

$$s^n \frac{\mathrm{d}\mathbf{P}_n\left[t\right]}{\mathrm{d}t} = s^n\left(n-1\right)\upsilon\mathbf{P}_{n-1}\left[t\right] + s^n\left(n+1\right)\omega\mathbf{P}_{n+1}\left[t\right] - s^n n\upsilon\mathbf{P}_n\left[t\right] - s^n n\omega\mathbf{P}_n\left[t\right]$$

Taking summations over the range of *n* demonstrates

$$\begin{aligned}\sum_{n=0}^{\infty} s^n \frac{d\mathbf{P}_n\left[t\right]}{dt} &= \sum_{n=0}^{\infty}\left(n-1\right)s^n\upsilon\,\mathbf{P}_{n-1}\left[t\right] \;-\sum_{n=0}^{\infty} ns^n\,\upsilon\mathbf{P}_n\left[t\right]\\ &+\sum_{n=0}^{\infty}\left(n+1\right)s^n\omega\,\mathbf{P}_{n+1}\left[t\right] - \sum_{n=0}^{\infty} ns^n\omega\,\mathbf{P}_n\left[t\right]\end{aligned}$$

$$\begin{aligned}\sum_{n=0}^{\infty} s^n \frac{d\mathbf{P}_n\left[t\right]}{dt} &= \upsilon\sum_{n=0}^{\infty}\left(n-1\right)s^n\,\mathbf{P}_{n-1}\left[t\right] \;-\upsilon\sum_{n=0}^{\infty} ns^n\,\mathbf{P}_n\left[t\right]\\ &+\omega\sum_{n=0}^{\infty}\left(n+1\right)s^n\,\mathbf{P}_{n+1}\left[t\right] - \omega\sum_{n=0}^{\infty} ns^n\,\mathbf{P}_n\left[t\right]\end{aligned}$$

We can take these terms one at a time. Note that for each demonstation, we choose a small enough value *s* such that $\sum_{n=0}^{\infty} s^n\mathbf{P}_n\left[t\right]$ is uniformly convergent. This allows us to exchange the derivative and the infinite sum.

$$\sum_{n=0}^{\infty} s^n \frac{d\mathbf{P}_n\left[t\right]}{dt} = \frac{d\sum_{n=0}^{\infty} s^n\,\mathbf{P}_n\left[t\right]}{dt} = \frac{d\mathbf{G}_t\left[s\right]}{dt}.$$

$$\begin{aligned}\upsilon\sum_{n=0}^{\infty}\left(n-1\right)s^n\,\mathbf{P}_{n-1}\left[t\right] &= \upsilon^2\sum_{n=0}^{\infty}\left(n-1\right)s^{n-2}\,\mathbf{P}_{n-1}\left[t\right] = \upsilon^2\sum_{n=0}^{\infty}\frac{ds^{n-1}\,\mathbf{P}_{n-1}\left[t\right]}{ds}\\ &= \upsilon^2\sum_{n=-1}^{\infty}\frac{ds^n\,\mathbf{P}_n\left[t\right]}{ds} = \upsilon^2\sum_{n=0}^{\infty}\frac{ds^n\,\mathbf{P}_n\left[t\right]}{ds}\\ &= \upsilon^2\frac{d}{ds}\sum_{n=0}^{\infty}s^n\,\mathbf{P}_n\left[t\right] = \upsilon^2\frac{d\,\mathbf{G}_t\left[s\right]}{ds}\end{aligned}$$

$$\begin{aligned}\upsilon\sum_{n=0}^{\infty} ns^n\,\mathbf{P}_n\left[t\right] &= \upsilon s\sum_{n=0}^{\infty} ns^{n-1}\,\mathbf{P}_n\left[t\right] = \upsilon s\sum_{n=0}^{\infty}\frac{ds^n\,\mathbf{P}_n\left[t\right]}{ds}\\ &= \upsilon s\frac{d}{ds}\sum_{n=0}^{\infty}s^n\,\mathbf{P}_n\left[t\right] = \upsilon s\frac{d\,\mathbf{G}_t\left[s\right]}{ds}\end{aligned}$$

$$\omega\sum_{n=0}^{\infty}(n+1)s^{n}\,\mathbf{P}_{n+1}[t]=\omega\sum_{n=0}^{\infty}\frac{ds^{n+1}\,\mathbf{P}_{n+1}[t]}{ds}=\omega\sum_{n=1}^{\infty}\frac{ds^{n}\,\mathbf{P}_{n}[t]}{ds}$$

$$=\omega\sum_{n=0}^{\infty}\frac{ds^{n}\,\mathbf{P}_{n}[t]}{ds}=\omega\frac{d}{ds}\sum_{n=0}^{\infty}s^{n}\,\mathbf{P}_{n}[t]=\omega\frac{d\,\mathbf{G}_{t}[s]}{ds}$$

$$\omega\sum_{n=0}^{\infty}ns^{n}\,\mathbf{P}_{n}[t]=\omega s\sum_{n=0}^{\infty}ns^{n-1}\,\mathbf{P}_{n}[t]=\omega s\sum_{n=0}^{\infty}\frac{ds^{n}\,\mathbf{P}_{n}[t]}{ds}$$

$$=\omega s\frac{d}{ds}\sum_{n=0}^{\infty}s^{n}\,\mathbf{P}_{n}[t]=\omega s\frac{d\,\mathbf{G}_{t}[s]}{ds}$$

We may now write

$$\frac{\partial\mathbf{G}_{t}[s]}{\partial t}=\nu s^{2}\frac{\partial\mathbf{G}_{t}[s]}{\partial s}-\nu s\frac{\partial\mathbf{G}_{t}[s]}{\partial s}+\omega\frac{\partial\mathbf{G}_{t}[s]}{\partial s}-\omega s\frac{\partial\mathbf{G}_{t}[s]}{\partial s}$$

$$\frac{\mathbf{G}_{t}[s]}{\partial t}=\nu s(s-1)\frac{\mathbf{G}_{t}[s]}{\partial s}+\omega(1-s)\frac{\mathbf{G}_{t}[s]}{\partial s}$$

$$\frac{\partial\mathbf{G}_{t}[s]}{\partial t}=\left(\nu s(s-1)+\omega(1-s)\right)\frac{\partial\mathbf{G}_{t}[s]}{\partial s}$$

$$\frac{\partial\mathbf{G}_{t}[s]}{\partial t}-\left(\nu s(s-1)+\omega(1-s)\right)\frac{\partial\mathbf{G}_{t}[s]}{\partial s}=0$$

which, when simplified, reveals

$$\frac{\partial\mathbf{G}_{t}[s]}{\partial t}-\left((\nu s-w)(s-1)\right)\frac{\partial\mathbf{G}_{t}[s]}{\partial s}=0.$$

This conforms to the form of the partial differential equation to which we are accustomed, and can now proceed with outlining first its general solution, then its particular solution.

## Solution for birth-death partial differential equation

The solution is a multistep process. First we identify the general solution for $\mathbf{G}_{t}[s]$, then using the boundary condition, we find a specific solution. We proceed to finding the general solution by writing the subsidiary equations from the partial differential equation above as

$$\frac{dt}{1}=\frac{ds}{-(\nu s-\omega)(s-1)}=\frac{d\,\mathbf{G}_{t}[s]}{0}$$

Using the first and third relationships, we see that

$\frac{dt}{1}=\frac{d\,\mathbf{G}_{t}[s]}{0}$ or $d\,\mathbf{G}_{t}[s]=(0)dt$ which means $\mathbf{G}_{t}[s]$ is a constant. We write this as $\mathbf{G}_{t}[s]=\Phi[c]$. To find what this constant is in its general form, we evaluate

$\frac{dt}{1} = \frac{ds}{-(\nu s-\omega)(s-1)}$ as follows

$$\frac{dt}{1} = \frac{ds}{-(\nu s-\omega)(s-1)}$$
$$-\int dt = \int \frac{ds}{(\nu s-\omega)(s-1)}.$$

Using partial fractions $\frac{1}{(\nu s-\omega)(s-1)} = \frac{A}{\nu s-\omega} + \frac{B}{s-1}$ or $1 = A(s-1) + B(\nu s-\omega)$. Setting $s=1$ reveals that $B = \frac{1}{\upsilon-\omega}$. Similarly letting $s = \frac{\omega}{\upsilon}$ shows $A = \frac{1}{\frac{\omega}{\upsilon}-1} = \frac{\upsilon}{\omega-\upsilon}$. Thus

$$\begin{aligned}
-\int dt = \int \frac{ds}{(\nu s-\omega)(s-1)} &= \int \frac{\frac{\upsilon}{\omega-\upsilon}}{\nu s-\omega} ds \quad + \int \frac{\frac{1}{\upsilon-\omega}}{s-1} \\
&= \frac{1}{\upsilon-\omega}\ln(s-1) - \frac{1}{\upsilon-\omega}\ln(\upsilon s-w) \\
&= -\left(\frac{1}{\upsilon-\omega}\ln(\upsilon s-w) - \frac{1}{\upsilon-\omega}\ln(s-1)\right) \\
&= \frac{-1}{\upsilon-\omega}\ln\left(\frac{\upsilon s-w}{s-1}\right).
\end{aligned}$$

This reveals that $(\upsilon-\omega)t + c = \ln\left(\frac{\upsilon s-w}{s-1}\right)$ or $c = -(\upsilon-\omega)t\ln\left(\frac{\upsilon s-w}{s-1}\right)$. Exponentiating reveals

$c_2 = e^{-(\upsilon-\omega)t}\frac{\upsilon s-\omega}{s-1}$ and $\mathbf{G}_t[s] = \Phi\left[e^{-(\upsilon-\omega)t}\frac{\upsilon s-\omega}{s-1}\right]$.

This is the general solution to the partial differental equation for $\mathbf{G}_t[s]$. We use the boundary solution to find the particular solution. Beginning with the observation that there are *a* cases of disease at time $t=0$, we find

$$\mathbf{G}_0[s] = s^a = \Phi\left[e^{-(\upsilon-\omega)0}\frac{\upsilon s-\omega}{s-1}\right] = \Phi\left[\frac{\upsilon s-\omega}{s-1}\right].$$

Letting $z = \frac{\upsilon s-\omega}{s-1}$ gives $s = \frac{z-\omega}{z-\upsilon}$ allows $\Phi[z] = \frac{z-\omega}{z-\upsilon}$.

$$\Phi(z) = \left[\frac{z-\omega}{z-\nu}\right]^{a_0}$$

(1.1)

We can now write

$$\mathbf{G}_t[s]=\Phi\left[e^{-(\upsilon-\omega)t}\frac{\upsilon s-\omega}{s-1}\right]^a=\left[\frac{e^{-(\upsilon-\omega)t}\dfrac{\upsilon s-\omega}{s-1}-\omega}{e^{-(\upsilon-\omega)t}\dfrac{\upsilon s-\omega}{s-1}-\upsilon}\right]^a$$

$$=\left[\frac{e^{-(\upsilon-\omega)t}(\upsilon s-\omega)-\omega(s-1)}{e^{-(\upsilon-\omega)t}(\upsilon s-\omega)-\upsilon(s-1)}\right]^a$$

$$=\left[\frac{\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)s+\omega\left(1-e^{-(\upsilon-\omega)t}\right)}{\left(\upsilon-\omega e^{-(\upsilon-\omega)t}\right)-\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)s}\right]^a$$

## Inversion of $\mathbf{G}_t[s]$

W can begin he inversion of $\mathbf{G}_t[s]$ be first letting $a=1,$ and then rewrite this as

$$\mathbf{G}_t[s]=\left[\frac{\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)s+\omega\left(1-e^{-(\upsilon-\omega)t}\right)}{\left(\upsilon-\omega e^{-(\upsilon-\omega)t}\right)-\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)s}\right]$$

$$=\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)s+\omega\left(1-e^{-(\upsilon-\omega)t}\right)\frac{1}{\left(\upsilon-\omega e^{-(\upsilon-\omega)t}\right)-\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)s}$$

This reminds us of the convolution of binomial (for $a=1,$ the Bernoulli distribution) and negative binomial (in its simplest form, the geometric distribution). However we have to identify their distribtional probabilities, $p_B$ and $p_N$ respectively.

For the Bernoulli distribution, we use a prior result.we have seen that $\mathbf{G}[s]=p_Bs+(1-p_B).$ Yet, here $\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)+\omega\left(1-e^{-(\upsilon-\omega)t}\right)\neq 1.$ In fact $\upsilon e^{-(\upsilon-\omega)t}-\omega+\omega\left(1-e^{-(\upsilon-\omega)t}\right)=(\upsilon-w)e^{-(\upsilon-\omega)t}.$ Thus

$$\left[\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)s+\omega\left(1-e^{-(\upsilon-\omega)t}\right)\right]$$

$$=(\upsilon-w)e^{-(\upsilon-\omega)t}\left[\frac{\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)}{(\upsilon-w)e^{-(\upsilon-\omega)t}}s+\frac{\omega\left(1-e^{-(\upsilon-\omega)t}\right)}{(\upsilon-w)e^{-(\upsilon-\omega)t}}\right]$$

and $p_B=\dfrac{\upsilon e^{-(\upsilon-\omega)t}-\omega}{(\upsilon-w)e^{-(\upsilon-\omega)t}}.$

The denominator of $\mathbf{G}_t[s],\ \dfrac{1}{\left(\upsilon-\omega e^{-(\upsilon-\omega)t}\right)-\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)s}$

can be converted to that of a geometric randon variable $\dfrac{q_N}{1-p_Ns}.$ We write

$$\frac{1}{\left(\upsilon-\omega e^{-(\upsilon-\omega)t}\right)-\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)s}=\frac{1}{\upsilon-\omega e^{-(\upsilon-\omega)t}}\frac{1}{1-\dfrac{\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)}{\upsilon-\omega e^{-(\upsilon-\omega)t}}s}$$

We compute

$$q_N=1-\frac{\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)}{\upsilon-\omega e^{-(\upsilon-\omega)t}}=\frac{\upsilon-\omega e^{-(\upsilon-\omega)t}-\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)}{\upsilon-\omega e^{-(\upsilon-\omega)t}}$$
$$=\frac{(1-\omega)e^{-(\upsilon-\omega)t}}{\upsilon-\omega e^{-(\upsilon-\omega)t}}.$$

And the denominator of $\mathbf{G}_t[s]$ becomes

$$\frac{1}{\upsilon-\omega e^{-(\upsilon-\omega)t}}\frac{\dfrac{(1-\omega)e^{-(\upsilon-\omega)t}}{\upsilon-\omega e^{-(\upsilon-\omega)t}}}{1\ -\ \dfrac{\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)}{\upsilon-\omega e^{-(\upsilon-\omega)t}}s}$$

Thus, for $a>0$,

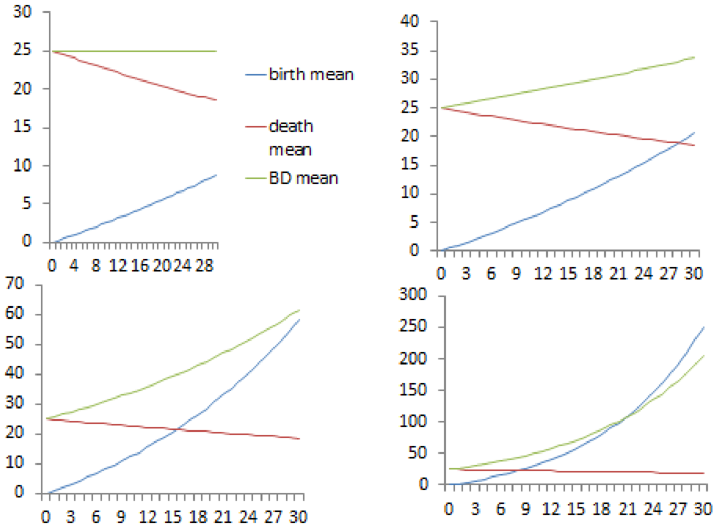
$$\mathbf{G}_t[s]=\left(\frac{(\upsilon-w)e^{-(\upsilon-\omega)t}}{\upsilon-\omega e^{-(\upsilon-\omega)t}}\frac{\left[\dfrac{\left(\upsilon e^{-(\upsilon-\omega)t}-\omega\right)}{(\upsilon-w)e^{-(\upsilon-\omega)t}}s+\dfrac{\omega\left(1-e^{-(\upsilon-\omega)t}\right)}{(\upsilon-w)e^{-(\upsilon-\omega)t}}\right]}{\left[\dfrac{\dfrac{(1-\omega)e^{-(\upsilon-\omega)t}}{\upsilon-\omega e^{-(\upsilon-\omega)t}}}{1\ -\ \dfrac{\upsilon\left(1-e^{-(\upsilon-\omega)t}\right)}{\upsilon-\omega e^{-(\upsilon-\omega)t}}s}\right]}\right)^a$$

The mean and variance are available from Bailey[*] as

$$\mathbf{E}\left[X_t(n)\right]=ae^{(\upsilon-\omega)t}$$
$$\mathbf{Var}\left[X_t(n)\right]=\frac{a(\upsilon+\omega)}{\upsilon-\omega}e^{(\upsilon-\omega)t}\left(e^{(\upsilon-\omega)t}-1\right).$$

Graphs of the mean values of birth, deatth, and the birth death process follows

[*] Bailey NTJ. 1964. The Elements of Stochastic Processes. New York. John Wiley and Sons.



Comparison of mean number of cases for birth, death, and birth-death processes

Contagion
Death Process
The Immigration-Death Process
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Emigration-Death Model

The model in this section considers only the possibility of a decrease of the occurrence of disease in the population, considering two forces which reduce the size of the population. One is independent of the population size. The other is proportional to that size. As before, we expect in the derivation of the generating function $\mathbf{G}_s(t)$ the appearance of a partial differential equation in both $t$ and $s$. Its solution will be straightforward.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions
Negative binomial measure
Binomial distribution

## Elaboration

In this circumstance, there are two processes working simultaneously. The first is the force which extinguishes the disease in an affected population at a rate that is proportional to the population size. This is either an overwhelming force, or a meager one based on the number of individuals in the population.

The second force is the emigration force, which removes patients from the system at a rate independent of the population size. Each of these two forces acts independently of the other.

An example of this circumstances would be post COVID-19 care. Recovery from this disease can be prolonged with ongoing respiratory and neurologic sequela. A community has a relatively large number of these impacted patients. A rehabilitative center is available in another community, and patients can emigrate there.

However, within the community there is a growing number of rehabilitative specialists who can successfully treat this disease locally. This force of treatment, more proportional to the number of these patients in the system, is better described by a "death process".

Of course, there will be the unfortunate victims of COVID-19 patients who die. The fate of these individuals is handled by a death process as well.*

## Developing the equations

In establishing the emigration-death differential equations, we return to our mechanism of slowing time down sufficiently so that one and only one event can occur in the time period $\left(t, t+\Delta t\right)$. Specifically, how can there be *n* patients in the system at time $t+\Delta t$? Assume that the emigration parameter is $\mu$ and the death parameter is $\omega$.

Then, this circumstance may occur in one of two ways. The first is that there are $n+1$ patients in the population at time *t*, and a death occurs, an event which occurs with probability $\left(n+1\right)\omega t$. In addition, again with $n+1$ patients at time *t*, a patient could leave the system with probability $\mu\Delta t$. . Finally, there could be *n* patients in the system with neither a death, nor an emigration occurring in time Δt. Assume that $\mathbf{P}_0\left(t\right)=0$, and $\mathbf{P}_a\left(0\right)=1$. The Chapman-Kolmogorov equations for this system are.

$$\mathbf{P}_{\mathrm{n}}(t+\Delta t)=\mathbf{P}_{n+1}(t)(n+1)\omega\Delta t \;+\; \mathbf{P}_{n+1}(t)\mu\Delta t \\ +\mathbf{P}_n(t)\left(1-n\omega\Delta t-\mu\Delta t\right)$$

Collecting all of the terms involving Δt on the left hand side of the equation and take a limit as Δt decreases in size to zero.

$$\frac{\mathbf{P}_{\mathrm{n}}(t+\Delta t)\;-\mathbf{P}_{\mathrm{n}}(t)}{\Delta t}\;=(n+1)\omega\mathbf{P}_{n+1}(t)+\mu\mathbf{P}_{n+1}(t) \\ -n\omega\mathbf{P}_{\mathrm{n}}(t)-\mu\mathbf{P}_n(t)$$

And taking a limit we get

$$\lim_{\Delta t\rightarrow 0}\frac{\mathbf{P}_{\mathrm{n}}(t+\Delta t)\;-\mathbf{P}_{\mathrm{n}}(t)}{\Delta t}\;=(n+1)\omega\mathbf{P}_{n+1}(t)+\mu\mathbf{P}_{n+1}(t) \\ -n\omega\mathbf{P}_{\mathrm{n}}(t)-\mu\mathbf{P}_n(t)$$

leading to

$$\frac{\mathrm{d}\mathbf{P}_{\mathrm{n}}(t)}{\mathrm{d}t}=(n+1)\omega\mathbf{P}_{n+1}(t)+\mu\mathbf{P}_{n+1}(t)-n\omega\mathbf{P}_{\mathrm{n}}(t)-\mu\mathbf{P}_n(t)$$

## Introducing the generating function

Begin as always by first defining the generating function

$$\mathbf{G}_s(t)=\sum_{\mathrm{n}=0}^{\mathrm{a}}s^{\mathrm{n}}\mathbf{P}_{\mathrm{n}}(t)$$

[*] The removal rate for those who die will likely be different the rate for those patients who are treated.

and move forward with the conversion and consolidation of the equation:

$$\frac{d\mathbf{P}_{\mathrm{n}}(t)}{dt}=(n+1)\omega\mathbf{P}_{n+1}(t)+\mu\mathbf{P}_{n+1}(t)-n\omega\mathbf{P}_{\mathrm{n}}(t)-\mu\mathbf{P}_{n}(t)$$

$$s^{n}\frac{d\mathbf{P}_{\mathrm{n}}(t)}{dt}=(n+1)\omega s^{n}\mathbf{P}_{n+1}(t)+\mu s^{n}\mathbf{P}_{n+1}(t)-n\omega s^{n}\mathbf{P}_{\mathrm{n}}(t)-\mu s^{n}\mathbf{P}_{n}(t).$$

The next step is to recognize that the last line of equation represents a system of equations for $0\le n\le a_0$, and write

$$\sum_{\mathrm{n=0}}^{\mathrm{a}}s^{n}\frac{d\mathbf{P}_{\mathrm{n}}(t)}{dt}=\frac{d\sum_{\mathrm{n=0}}^{\mathrm{a}}s^{n}\mathbf{P}_{\mathrm{n}}(t)}{dt}$$

$$=\omega\sum_{\mathrm{n=0}}^{\mathrm{a}}(n+1)s^{n}\mathbf{P}_{n+1}(t)+\mu\sum_{\mathrm{n=0}}^{\mathrm{a}}s^{n}\mathbf{P}_{n+1}(t)$$

$$-\omega\sum_{\mathrm{n=0}}^{\mathrm{a}}ns^{n}\mathbf{P}_{\mathrm{n}}(t)-\mu\sum_{\mathrm{n=0}}^{\mathrm{a}}s^{n}\mathbf{P}_{n}(t)$$

Following our work in the emigration and death processes, we recognize these summands as functions of $\mathbf{G}_t(s)$ and write

$$\frac{\partial\mathbf{G}_t(s)}{\partial t}=\omega\frac{\partial\mathbf{G}_t(s)}{\partial s}+\mu s^{-1}\mathbf{G}_t(s)-\omega s\frac{\partial\mathbf{G}_t(s)}{\partial s}-\mu\mathbf{G}_t(s)$$

$$=\omega(1-s)\frac{\partial\mathbf{G}_t(s)}{\partial s}+\mu\left(s^{-1}-1\right)\mathbf{G}_t(s)$$

which becomes

$$\frac{\partial\mathbf{G}_t(s)}{\partial \mathrm{t}}-\omega(1-s)\frac{\partial\mathbf{G}_t(s)}{\partial \mathrm{s}}-\mu\left(s^{-1}-1\right)\mathbf{G}_t(s)=0$$

## Solving the partial differential equation

We can recognize the last line of this equation as a partial differential equation in *t* and in *s*, but of the form that will allow us to find a general solution using a subsidiary set of equations. Write these equations as

$$\frac{dt}{1}=\frac{ds}{-\omega(1-s)}=\frac{d\mathbf{G}_t(s)}{-\mu\left(s^{-1}-1\right)\mathbf{G}_t(s)}$$

Taking these terms two at a time, we can identify information on the form of the generating function $\mathbf{G}_t(s)$

Using the first and third terms from the previous set of equations

$$\frac{dt}{1}=\frac{ds}{-\omega(1-s)}=\frac{d\mathbf{G}_t(s)}{-\mu\left(s^{-1}-1\right)\mathbf{G}_t(s)}$$

write

$$\frac{\mathrm{d}t}{1}=\frac{\mathrm{d}\mathbf{G}_t(s)}{-\mu\left(s^{-1}-1\right)\mathbf{G}_t(s)}$$

From our emigration model, we see that this is

$$\mathbf{G}_t(s)=\mathrm{e}^{\mu t\left(s^{-1}-1\right)}\Phi(\mathrm{C})$$

Using the first and second equations, we find that

$$\frac{\mathrm{d}t}{1}=\frac{\mathrm{d}s}{-\omega\,(1-s)}$$
$$\frac{\omega\,\mathrm{d}t}{1}=\frac{-\mathrm{d}s}{(1-s)}$$
$$\int\omega\,\mathrm{d}t=\int\frac{-\mathrm{d}s}{(1-s)}$$
$$\omega t+\mathrm{C}=\ln(1-s)$$
$$\mathrm{C}=-\omega t+\ln(1-s)$$

From which find $\mathrm{C}_2=\mathrm{e}^{-\omega t}(1-s)$

Combining $\mathbf{G}_t(s)=\mathrm{e}^{\mu t\left(s^{-1}-1\right)}\Phi(\mathrm{C}_2)$ with $\mathrm{C}_2=\mathrm{e}^{-\omega t}(1-s)$ gives

$$\mathbf{G}_t(s)=\mathrm{e}^{\mu t\left(s^{-1}-1\right)}\Phi\left(\mathrm{e}^{-\omega t}(1-s)\right).$$

and we are now ready to identify the specific solution to the death-emigration model.

## Specific solution

Having identified $\mathbf{G}_t(s)=\mathrm{e}^{\mu t\left(s^{-1}-1\right)}\Phi\left(\mathrm{e}^{-\omega t}(1-s)\right).$ now pursue a specific solution, beginning with the boundary conditions. At $t=0$, if there are *a* patients in the population

$$\mathbf{G}_0(s)=s^{\mathrm{a}}=\Phi(1-s)$$

Now, if we let $z=1-s$, then $s=1-z.$ and substituting this result into the above equation,

$$(1-z)^a=\Phi(z)$$

We can now write

$$\mathbf{G}_t(s) = \mathrm{e}^{\mu t\left(s^{-1}-1\right)}\Phi\left(\mathrm{e}^{-\omega t}(1-s)\right) = \mathrm{e}^{\mu t\left(s^{-1}-1\right)}\left[1-\left(\mathrm{e}^{-\omega t}(1-s)\right)\right]^{\mathrm{a}}$$

This is the generating function of the emigration-death process. It remains to invert it.

## Inversion

The inversion of $\mathbf{G}_t(s)$ can be carried out in a straightforward manner. The first term on the right hand side of equation reflects the generating function for the emigration process. The remaining term is the probability generating function of the binomial distribution with $p = e^{-\omega t}$. Thus

$$\left[\left(1-\mathrm{e}^{-\omega t}\right)+s\mathrm{e}^{-\omega t}\right]^{a} \triangleright \left\{\binom{a}{k}\mathrm{e}^{-\omega tk}\left(1-\mathrm{e}^{-\omega t}\right)^{a-k}\right\}.$$

We only have to address the parameters. In the emigration-death process, the number of subjects in the system can only decrease. In order to have n in the system at time t, we can combination of $a-n$ emigrations and deaths. If there are $m$ emigrations, than there must be $a-n-m$ deaths. Thus

$$\mathbf{P}_n(t) = \sum_{m=0}^{a-n}\frac{\mu t}{m!}e^{-\mu t}\binom{a-n-m}{k}e^{-k\omega t}\left(1-e^{-\omega t}\right)^{a\text{-}n\text{-}m\text{-}k}$$

Contagion
Death Process
The Immigration-Death Process
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Immigration-Death Model

The model in this particular section considers only the possibility of a decrease of the occurrence of disease in the population, considering two forces which reduce the size of the population. One is independent of the population size. The other is proportional to that size. As before, we expect in the derivation of the generating function $\mathbf{G}_s\left(t\right)$ the appearance of a partial differential equation in both *t* and *s*. Its solution will be straightforward.

## Prerequisites

Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations
Exponential Limit
The Exponential and Gamma Functions
Properties of Probability
General Poisson Process
Moment and Probability Generating Functions
Negative binomial measure
Binomial distribution

## Elaboration

In this circumstance, there are two processes working simultaneously. The first is the process that grows the population of single individuals by their simple arrival into the community. These are arrivals that are independent of each other and also independent of the number of individuals with the disease. This is the immigration process with which we are familiar. However, the second is the force which extinguishes affected population at a rate that is proportional to the population size. This is either an overwhelming force, or a meager one based on the number of individuals in the population.

It is important to understand that this need not be death. For example a community that has new arrivals with colon cancer. These individuals arrive at a constant rate. However, the community, not prepared for the new arrival of these patients, develops a collection of health care providers and treatment centers for colon cancer. These new treatment complexes are proportional to the total number of patients arriving in the population. Thus, this “treatment process” meets the criteria of our “death process”, i.e., the treatments are independent of each other and proportional to the number of colon cancer patients in to the community.

In addition, each of the immigration and death processes act independently of each other.

## Developing the equations

In establishing the immigration-death differential equations, we return to our mechanism of slowing time down sufficiently so that one and only one event can occur in the time period $\left(t, t+\Delta t\right)$. Specifically, how can we have a total *n* patients in the system at time $t+\Delta t$?

Assume that the emigration parameter is $\mu$ and the death parameter is $\omega$.

Then, this may occur in one of three ways. The first is that there are $n+1$ patients in the population at time *t*, and a death occurs, an event which occurs with probability $\left(n+1\right)\omega t$. In addition, with with $n-1$ patients at time *t*, a patient could enter the system with probability $\lambda\,\Delta t$. Finally, there could be *n* patients in the system with neither a death, nor an immigration occurring in time Δt. Assume that $\mathbf{P}_0\left(t\right)=0$, and $\mathbf{P}_a\left(0\right)=1$. The Chapman-Kolmogorov equations for this system are.

$$\begin{aligned}\mathbf{P}_{\mathrm{n}}(t+\Delta t) = \mathbf{P}_{n+1}(t)(n+1)\,\omega\,\Delta t \;+\; \mathbf{P}_{n\text{-}1}(t)\,\lambda\,\Delta t \\ +\mathbf{P}_n(t)\left(1-n\,\omega\,\Delta t-\lambda\Delta t\right)\end{aligned}$$

Collecting all of the terms involving Δt on the left hand side of the equation and take a limit as Δt decreases in size to zero.

$$\begin{aligned}\frac{\mathbf{P}_{\mathrm{n}}(t+\Delta t)\; -\mathbf{P}_{\mathrm{n}}(t)}{\Delta t} &= (n+1)\,\omega\,\mathbf{P}_{n+1}(t)+\lambda\,\mathbf{P}_{n\text{-}1}(t) \\ &-n\,\omega\,\mathbf{P}_{\mathrm{n}}(t)-\lambda\mathbf{P}_n(t)\end{aligned}$$

And taking a limit we get

$$\begin{aligned}\lim_{\Delta t\to 0}\frac{\mathbf{P}_{\mathrm{n}}(t+\Delta t)\; -\mathbf{P}_{\mathrm{n}}(t)}{\Delta t} &= (n+1)\,\omega\,\mathbf{P}_{n+1}(t)+\lambda\,\mathbf{P}_{n+1}(t) \\ &-n\,\omega\,\mathbf{P}_{\mathrm{n}}(t)-\lambda\,\mathbf{P}_n(t)\end{aligned}$$

leading to

$$\frac{\mathrm{d}\mathbf{P}_{\mathrm{n}}(t)}{\mathrm{d}t} = (n+1)\,\omega\,\mathbf{P}_{n+1}(t)+\lambda\,\mathbf{P}_{n\text{-}1}(t)-n\,\omega\,\mathbf{P}_{\mathrm{n}}(t)-\lambda\,\mathbf{P}_n(t)$$

## Introducing the generating function

Begin as always by first defining the generating function

$$\mathbf{G}_s(t)=\sum_{\mathrm{n=0}}^{\mathrm{a}} s^{\mathrm{n}}\mathbf{P}_{\mathrm{n}}(t)$$

and move forward with the conversion and consolidation of the equation:

$$\frac{d\mathbf{P}_{\mathrm{n}}(t)}{dt} = (n+1)\omega\mathbf{P}_{n+1}(t) + \lambda\mathbf{P}_{n\text{-}1}(t) - n\omega\mathbf{P}_{\mathrm{n}}(t) - \lambda\mathbf{P}_{n}(t)$$

$$s^n\frac{d\mathbf{P}_{\mathrm{n}}(t)}{dt} = (n+1)\omega s^n\,\mathbf{P}_{n+1}(t) + \lambda s^n\,\mathbf{P}_{n\text{-}1}(t) - n\omega s^n\mathbf{P}_{\mathrm{n}}(t) - \lambda s^n\,\mathbf{P}_{n}(t).$$

The next step is to recognize that the last line of equation represents a system of equations for $0 \le n \le \infty$, and write

$$\sum_{\mathrm{n}=0}^{\infty} s^n\frac{d\mathbf{P}_{\mathrm{n}}(t)}{dt} = \frac{d\sum_{\mathrm{n}=0}^{\infty} s^n\mathbf{P}_{\mathrm{n}}(t)}{dt}$$

$$= \omega\sum_{\mathrm{n}=0}^{\infty}(n+1)s^n\,\mathbf{P}_{n+1}(t) + \lambda\sum_{\mathrm{n}=0}^{\infty}s^n\,\mathbf{P}_{n\text{-}1}(t)$$

$$-\omega\sum_{\mathrm{n}=0}^{\infty}ns^n\mathbf{P}_{\mathrm{n}}(t) - \lambda\sum_{\mathrm{n}=0}^{\infty}s^n\,\mathbf{P}_{n}(t).$$

Following our work in the immigration and death processes, we recognize these summands as functions of $\mathbf{G}_t(s)$ and write

$$\begin{aligned}\frac{\partial\mathbf{G}_t(s)}{\partial t} &= \omega\frac{\partial\mathbf{G}_t(s)}{\partial s} + \lambda s\mathbf{G}_t(s) - \omega s\frac{\partial\mathbf{G}_t(s)}{\partial s} - \lambda\mathbf{G}_t(s)\\ &= \omega\,(1-s)\frac{\partial\mathbf{G}_t(s)}{\partial s} + \lambda\left(s-1\right)\mathbf{G}_t(s)\end{aligned}$$

which becomes

$$\frac{\partial\mathbf{G}_t(s)}{\partial \mathrm{t}} - \omega(1-s)\frac{\partial\mathbf{G}_t(s)}{\partial \mathrm{s}} - \lambda\left(s-1\right)\mathbf{G}_t(s) = 0$$

## Solving the Partial Differential Equation

We can recognize the last line of this equation as a partial differential equation in *t* and in *s*, but of the form that will allow us to find a general solution using a subsidiary set of equations. Write these equations as

$$\frac{dt}{1} = \frac{ds}{-\omega\left(1-s\right)} = \frac{d\mathbf{G}_t\left(s\right)}{\lambda\left(s-1\right)\mathbf{G}_t\left(s\right)}$$

Taking these terms two at a time, we can identify information on the form of the generating function $\mathbf{G}_t\left(s\right)$

For example, from the first and third terms note

$$\frac{dt}{1} = \frac{d\mathbf{G}_t\left(s\right)}{\lambda\left(s-1\right)\mathbf{G}_t\left(s\right)}$$

Recall from our development of the integration process, we have

$$\mathbf{G}_t(s) = e^{\lambda t(s-1)}\Phi(C)$$

The remaining requirement is to identify the form $\Phi(C)$.

Our experience with the death process also reveals that $\Phi[C] = \Phi\left[e^{-\omega t}(1-s)\right]$. Thus, we can write

$$G_t(s) = e^{\lambda t(s-1)}\Phi\left[e^{-\omega t}(1-s)\right].$$

and we are now ready to identify the specific solution to the immigration-death model.

## Specific solution

Having identified $\mathbf{G}_t(s) = e^{\lambda t(s-1)}\Phi\left(e^{-\omega t}(1-s)\right)$. now pursue a specific solution, beginning with the boundary conditions. At $t = 0$, if there are $a$ patients in the population

$$\mathbf{G}_0(s) = s^a = \Phi(1-s)$$

Now, if we let $z = 1-s$, then $s = 1-z$. and substituting this result into the above equation,

$$\Phi(z) = (1-z)^a$$

or $\Phi(z) = (1-z)^a$. We can now write

$$\mathbf{G}_t(s) = e^{\lambda t(s-1)}\left[1\text{-}\left(e^{-\omega t}(1-s)\right)\right]^a$$

This is the generating function of the immigration-death process. It remains to invert it.

## Inversion

The inversion of $\mathbf{G}_t(s)$ can be carried out in a straightforward manner. The first term on the right hand we recognize as the generating function of a Poisson random variable; the second is the probability generating function of the binomial distribution with $p = e^{-\omega t}$. Thus we can write that, in order to have $n$ subjects in the system, we have m arrivals, and $a+m-n$ departures.

$$\mathbf{P}_n(t) = \sum_{m=0}^{\infty}\frac{(\lambda t)^m}{m!}e^{-\lambda t}\binom{a+m}{a+m\text{ - }n}e^{-\omega t(a+m\text{-}n)}\left(1-e^{-\omega t}\right)^n$$

Note the similarity of this result with that of the emigration-death model. In that case, as here the solution was essential one of binomial measure. However, the adjustments are different. In the emigration-death model, emigrations decrease the number of subjects who undergo the death process. However, in this model, immigration increases the number of subjects available to undergo the "death" or "treatment" process.

## Mean and variance

We can find the expected number of patients in the system at time *t*, $\mathbf{E}\left[n\right]$, in a straightforward fashion using the concept of double expectation.

$$\mathbf{E}\left[n\right]=\mathbf{E}_m\left[\mathbf{E}\left[n\,|\,m\right]\right].$$

$\mathbf{E}\left[n\,|\,m\right]$ is the expected value of a binomial random variable with parameters $\left(a+m,\,1-e^{-\omega t}\right)$. Thus

$$\mathbf{E}\left[n\,|\,m\right]=\left(a+m\right)\left(1-e^{-\omega t}\right).$$

The final expectation is taken with respect to a Poisson random variable with parameter $\lambda$. We can therefore write

$$\mathbf{E}\left[n\,|\,m\right]=\left(a+\lambda\right)\left(1-e^{-\omega t}\right).$$

The variance is also straightforward. We note that

$$\mathbf{Var}\left[n\right]=\mathbf{Var}_m\left[\mathbf{E}\left[n\,|\,m\right]\right]+\mathbf{E}_m\left[\mathbf{Var}\left[n\,|\,m\right]\right].$$

Following the development of the mean, we can compute

$$\mathbf{E}\left[n\,|\,m\right]=\left(a+m\right)\left(1-e^{-\omega t}\right):\mathbf{Var}\left[n\,|\,m\right]=\left(a+m\right)\left(1-e^{-\omega t}\right)e^{-\omega t}.$$

The next level of moments is taken with respect to the Poisson distribution.

$$\mathbf{E}_m\left[\mathbf{Var}\left[n\,|\,m\right]\right]=\mathbf{E}_m\left[\left(a+m\right)\left(1-e^{-\omega t}\right)e^{-\omega t}\right]$$
$$=\left(a+\lambda\right)\left(1-e^{-\omega t}\right)e^{-\omega t}.$$

$$\mathbf{Var}_m\left[\mathbf{E}\left[n\,|\,m\right]\right]=\left(a+m\right)\left(1-e^{-\omega t}\right)=\lambda\left(1-e^{-\omega t}\right)^2.$$

Thus

$$\mathbf{Var}\left[n\right]=\left(a+\lambda\right)\left(1-e^{-\omega t}\right)e^{-\omega t}+\lambda\left(1-e^{-\omega t}\right)^2$$
$$=ae^{-\omega t}+\lambda\left(1-e^{-\omega t}\right).$$

Contagion
Death Process
The Emigration-Death Process
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Continuous Probability Measure

Prerequisites
Properties of Real Numbers
The Concept of the Limit
Pointwise vs. Uniform Convergence
Uniform Convergence and Continuity
An Introduction to the Concept of Measure
Elementary Set Theory
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Bernoulli Distribution – In Depth Discussion
Geometric and Negative binomial measures
General Poisson Process

We have seen a progression in the complexity of probability functions that we have discussed so far. Our first distribution, the Bernoulli, set equal values for point mass on each of two values, 0, and 1. For us then, manipulating functions of random variables that follow Bernoulli measure was simply a matter of managing two point masses.

Moving to the binomial and then the multinomial distribution complicated the counting we had to do, but not the process. Binomial $(n, p)$ measure was simply accumulating the point mass using these measures required us to keep track of not just two point masses, but $n$ of them.

The situation transformed when we came to geometric, negative binomial and Poisson measure. For each of these distributions, we continue to accrue point mass using either of these functions. However, we were no longer confined to accruing this measure over solely finite ranges of the nonnegative integers. Computing a probability for a random variable $X$ that followed a negative binomial $(r, p)$ such as $\mathbf{P}(X > 100)$ requires consideration and summation of an infinite number of events.

However, since each of these probability distributions converge (i.e., the sum over all nonnegative $k$ for each of these point mass functions is finite), we can accumulate measure or probability by summing over an infinite number of integers if need be. Thus while the range of the summands changed by moving to the negative binomial and Poisson distributions, the process by which we accumulated their measure did not; we simply summed the function over the relevant integer range, whether it was finite or infinite.

However, there is an entire collection of random variables for which simple summing is not the case. In managing this new circumstance, we will stay true to our concept of measure or accumulation; we must simply measure differently.

## Probability as a limiting process

Consider a very simple collection of random variable $X_n = \left\{ \frac{1}{1}, \frac{1}{2}, \frac{1}{3} \ldots \frac{1}{n} \right\}$ each value having probability or mass $\frac{1}{n}$. For each value of *n*, $X_n$ represents a proper family of random variables. For example, for *n* = 1, $X_1 = 1$ with probability 1. This random variable poses no problem for our concept of or measurement of probability. For *n* = 2, $X_2$ is equal to either $\frac{1}{2}$ or 1, each with probability $\frac{1}{2}$. This also represents no particular challenge to us. Similarly for *n* = 3, 4, 5, ….

For any set value of *n*, we know how to identify each distinct possible value of $X_n$ and compute the value of its probability as simply $\frac{1}{n}$, and therefore understand and manage probabilities for the entire sample space (for example, the $\mathbf{P}\left[ X_n \leq \frac{n}{n+1} \right]$.

However, extending this practice farther and farther out into this process, we note that the number of possible values of the random variable $X_n$ increases, and also, the probability that any of these values of $X_n$ decreases. We never get to the point where there are an infinite number of possible value of $X_n$ or where the $\mathbf{P}\left[ X_n \right]$ is zero, but we get close enough to wonder how we would ever manage that.

Also, consider the $\mathbf{P}\left[ X_n \leq 1 \right]$. This is easily computed as $\mathbf{P}\left[ X_n \leq 1 \right] = \sum_{k=1}^{n} \frac{1}{k}$. Again, initially, this poses no problem for us as $\mathbf{P}\left[ X_1 \leq 1 \right] = 1, \mathbf{P}\left[ X_2 \leq 1 \right] = 2\left( \frac{1}{2} \right), \mathbf{P}\left[ X_3 \leq 1 \right] = 3\left( \frac{1}{3} \right)$, etc. For any finite *n*, we compute $\mathbf{P}\left[ X_n \leq 1 \right] = n\left( \frac{1}{n} \right) = 1$. But what happens when *n* is infinity? What does $\mathbf{P}\left[ X_\infty \leq 1 \right] = \infty\left( \frac{1}{\infty} \right)$ mean? Is this 0? Infinity? 1? Something else?[*]

We have identified an infinite sequence of random variables, that, while the behavior of any one of them makes perfect sense, we can make no sense out of the behavior of the limiting process. Where is the problem?

The problem is the concept of point mass as probability.

Assigning positive probability to a point has been helpful for us so far because we could do that in a way that the total accumulated probability was one. Even when the number of points was infinite (e.g., for the Poisson distribution) we could count on $\sum_{n=0}^{\infty} \mathbf{P}\left[ n \right] = 1$. However in the current example this is not the case.[†]

An inspection of the concept of probability as point mass (Figure 1) suggests another candidate for probability.

[*] Infinity is far too complicated a concept to blithely assume that $\frac{\infty}{\infty} = 1$.

[†] We know this since the harmonic sequence $A = \sum_{n=1}^{\infty} \frac{1}{n}$ does not converge.

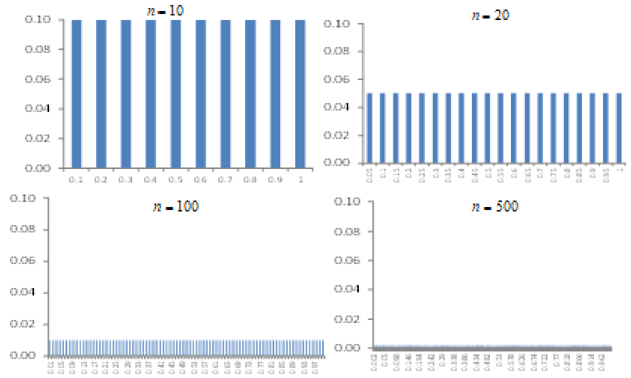


Figure 1. Shrinking probability as $n$ increases. Note the probability becomes more densely packed.

From Figure 1, we notice two things 1) the probability for each point gets smaller, and 2) there are more points over which to distribute the probability. The thickness of the probability increases as it is spread across the real line. This suggests that we assign probability to not just a single point, but to an interval of points on the real line.

This is commonly stated as assigning probability as area.

Note this is not a change in our concept of measure. We are still very much interested in the idea of measure as accumulation. However,  we have always had flexibility in how we accumulate measure. In the past, we accumulated measure by simply counting. Now we simply need another tool.

## Measure as an interval

Suppose that we want to assign probability to an interval of real numbers. We know how to determine the length of an interval on the positive real number line; the length of the interval $[0,1]$ is simply $b-a$.*

What probability that we assign this interval is up to us, just as it was up to us to decide what probability we wished to assign to a point (counting measure? Poisson measure?) There are many ways to assign measure or probability to an interval. In some circumstances, probability is its length (this is commonly known as Lebesgue measure), but we could also assign probability as $e^{-3a}-e^{-3b}$.

It is wholly up to us, as long as we are consistent with the properties of probability. However, assigned in accordance with these rules, we will be forced to acknowledge that there are some circumstances in which positive probability cannot be provided to individual points. While there are important advantages to computing probability over an interval of real numbers, a relative disadvantage is that we cannot assign a nonzero value to any particular point unless we stipulate that there are particular values that will have positive probability.

In fact many commonly used probability measures are based on continuous random variables. The uniform, normal, and exponential are just a few of the distributions that we will discuss that assign different measure or probability to the real line. In each case, particular values of random variables that follow these distributions do not have positive probability assigned to

* Recall that this is what is known as Lebesgue measure.

them. In fact, any discrete or countable collection of values for these random variables ( e.g., $X=\frac{1}{8}$, or $X$ being any integer greater than 5) have no positive value assigned to any of them.[*] While probability can accumulate using interval length, it cannot accumulate by actually collecting individual points and trying to "measure" them. As we saw with irrational numbers, there are far too many of them to bound the measure of the real line by a finite constant.

Probability has prospered with the development of this tool using the Riemann integral to provide these computations.

Consider for example a function $f(x)=2\mathbf{1}_{0\le x\le \frac{1}{2}}$. This is a straight horizontal line of height two over the real numbers from zero to $\frac{1}{2}$ (Figure 2).

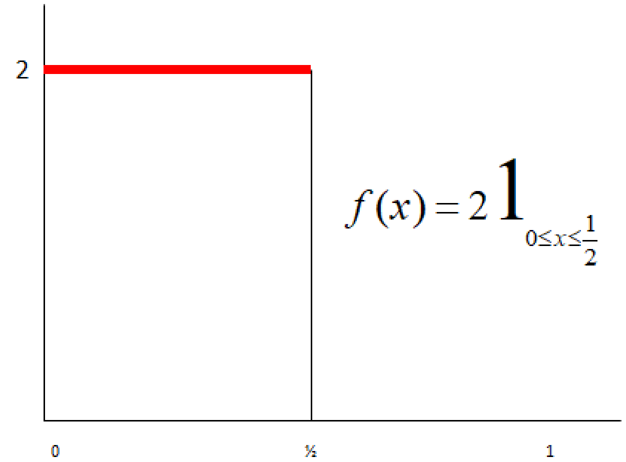


**Figure 2**. Example of a function where probability is area.

If we replace for a moment the concept of probability as point mass with the notion of probability as area, we can assure ourselves of some findings that make us more comfortable with the idea of probability as area.

First, in this example, what is $\int_{\Omega} d\mathbf{P}$ ? We fully expect this quantity should be one. Carrying out the measure of this interval, we find $\int_{\Omega} d\mathbf{P}=\int_{\Omega} 2\mathbf{1}_{0\le x\le \frac{1}{2}}$, which simplifies to $=2\int_{0}^{\frac{1}{2}} dx=(2)\left(\frac{1}{2}\right)=1.$[†] We also note that the closed interval $\left[0,\ \frac{1}{2}\right]$ has the same probability as the open $\left(0,\ \frac{1}{2}\right)$ since the probability of the end points of 0 and $\frac{1}{2}$ are each zero. We also see that intervals outside of the interval $\left[0,\ \frac{1}{2}\right]$ have probability zero.

For an interval $(a,\, b)$ such that $0\le a<b\le\frac{1}{2}$ we can find the probability $\mathbf{P}\left[a<X<b\right]$.

[*] We will see later that we can combine random variables mixing discrete and continuous measuring tools simultaneously. Here we are only talking about a continuous measuring tool.

[†] The other properties of probability discussed in can be shown to be satisfied.

$$\mathbf{P}\left[a < X < b\right] = \int_a^b d\mathbf{P} = \int_a^b 2\mathbf{1}_{0 \le x \le \frac{1}{2}} dx = 2\int_a^b dx = 2(b-a).$$

Another way to handle the probability of interval of the random variable *X* is to write the interval as an element function and takes its expectation. In this case we would write the event that $a < X < b$ as $\mathbf{1}_{a<X<b}$, and write

$$\mathbf{P}\left[a < X < b\right] = \mathbf{E}_X\left[\mathbf{1}_{a<X<b}\right] = \int_\Omega \mathbf{1}_{a \le x \le b} d\mathbf{P}$$

$$= \int_\Omega \mathbf{1}_{a \le x \le b} 2\mathbf{1}_{0 \le x \le \frac{1}{2}} dx = 2\int_a^b dx = 2(b-a).$$

Note that we used $\mathbf{1}_{a<X<b}\mathbf{1}_{0<X<\frac{1}{2}} = \mathbf{1}_{a<X<b}$ for $0 \le a < b \le \frac{1}{2}$. Also note that we expressed the expectation as $\mathbf{E}_X\left[\ \right]$ to be clear about the probability distribution whose expectation we wanted.

The above formulation with element functions can be very helpful. Suppose we want the expected value of *X* for $0 \le a < b \le \frac{1}{2}$.We can find this by noting that $\mathbf{E}_X\left[g(x)\right] = \int_\Omega g(x) d\mathbf{P}$, choose $g(x) = x\mathbf{1}_{a<X<b}$, and write

$$\mathbf{E}_X\left[x\mathbf{1}_{a<X<b}\right] = \int_\Omega x\mathbf{1}_{a<X<b} 2\mathbf{1}_{0 \le x \le \frac{1}{2}} dx = 2\int_a^b x dx = b^2 - a^2.$$

## Cumulative and density functions

We are already comfortable with the concept of a probability function. We introduced it as the probability of a particular point, a definition that served us well in the probability-as-point-mass environment. However, this concept does not hold for us all the time. In order to manage this new case of continuous random variables, we now rename the probability function as a *probability density function*

The cumulative distribution function retains its original concept. However, since the random variable *X* is not continuous, the cumulative distribution function is also continuous. This continuity is a foundation concept, helping us to now identify a new relationship between the cumulative distribution function $F_X(x)$ and the density function $f_X(x)$. We can now write

$$F_X(x) = \int_{-\infty}^{x} f_X(y) dy.$$

Alternatively, we can write

$$f_X(x) = \frac{dF_X(x)}{dx} {}^{*}$$

[*] We assume here that the derivative does exist.

We will find the ability to shuttle back and forth between the cumulative and density functions very helpful for these new continuous random variables.

This is one of the advantages of the measure concept. It permits us to discuss the measure of an event without differentiating mass function from density function. Measure simply uses a tool – that tool can be discrete, continuous, or a combination of both.

## Probability as different measures

As an example of a random variables that has both discrete and continuous properties, consider the random variable $X$ which is continuous on $\left(0,\frac{3}{2}\right)$. It takes on the value one for the interval $\left(\frac{1}{2},1\right)$ and the value $\frac{1}{2}$ for the other two intervals of equal length. What is the measure in this case? We can write the measure $\mathbf{P}$ as

$$\mathbf{P}=\frac{1}{2}1_{0<x<\frac{1}{2}}+1_{\frac{1}{2}<x<1}+\frac{1}{2}1_{1<x<\frac{3}{2}}$$

Figure 3 provides a clue as to how to measure this function.

The figure tells us how to compute it. We see that the measure that is to be applied to each interval depends on the interval itself. Thus we compute

$$\begin{aligned}\int_\Omega d\mathbf{P}&=\int_\Omega\left(\frac{1}{2}1_{0<x<\frac{1}{2}}+1_{\frac{1}{2}<x<1}+\frac{1}{2}1_{1<x<\frac{3}{2}}\right)dx\\&=\int_\Omega\frac{1}{2}1_{0<x<\frac{1}{2}}\,dx+\int_\Omega 1_{\frac{1}{2}<x<1}\,dx+\int_\Omega\frac{1}{2}1_{1<x<\frac{3}{2}}\,dx\\&=\frac{1}{4}+\frac{1}{2}+\frac{1}{4}=1.\end{aligned}$$

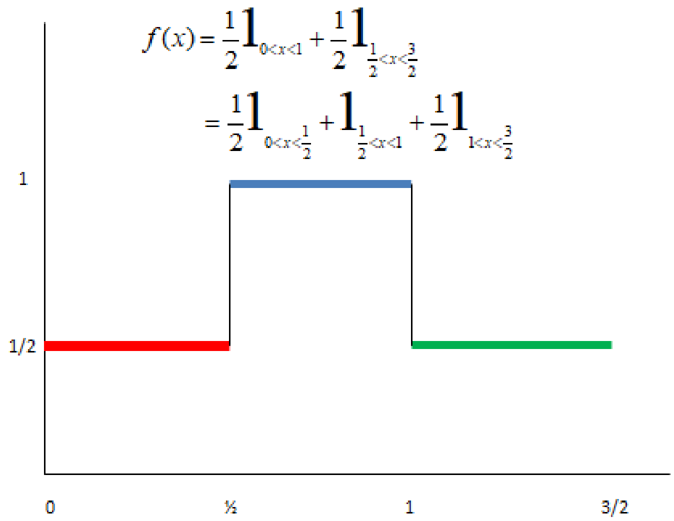


**Figure 3**. Example of a function where the measure is interval dependent.

What worked for us here is dividing the function into regions of the real line such that the value of the function was constant. Then we just multiplied the function value by the interval length. This elementary example is the heart of Lebesgue integration, and is analogous to the process by which we built up measurable functions from linear combinations of simple functions. With this approach we can combine complicated probability function yet still verify the basic properties that they must satisfy, as well as compute probabilities and moments.

To find $\mathbf{P}\left[\frac{5}{8} < X < \frac{9}{8}\right]$, we compute

$$\mathbf{P}\left[\frac{5}{8} < X < \frac{9}{8}\right] = \int_{\frac{5}{8}}^{\frac{9}{8}} \left( \frac{1}{2}\mathbf{1}_{0<x<\frac{1}{2}} + \mathbf{1}_{\frac{1}{2}<x<1} + \frac{1}{2}\mathbf{1}_{1<x<\frac{3}{2}} \right) dx$$

$$= \int_{\frac{5}{8}}^{1} dx + \frac{1}{2}\int_{1}^{\frac{9}{8}} dx = \frac{7}{16}.$$

We only have to simplify the function so that we identify sets of the real number line where there are common values of the function then multiply the function value by the interval width. So, combining these probability functions is a straightforward process using the concept of Lebesgue measure.

## Probability as area *and* point mass

We can take this one step further, and not just combined functions which assign different measures to intervals, but we can combine point mass functions (traditionally known as discrete functions) with "measure as area" probability functions as well.

Consider the following very elementary example

$$f(x) = \frac{1}{2}\mathbf{1}_{0<x<1} + \frac{1}{2}\mathbf{1}_{x=\frac{1}{2}}.$$

Here, probability is measured as both area on (0,1), and as a single point mass with a height of one at $x = \frac{1}{2}$ (Figure 4).

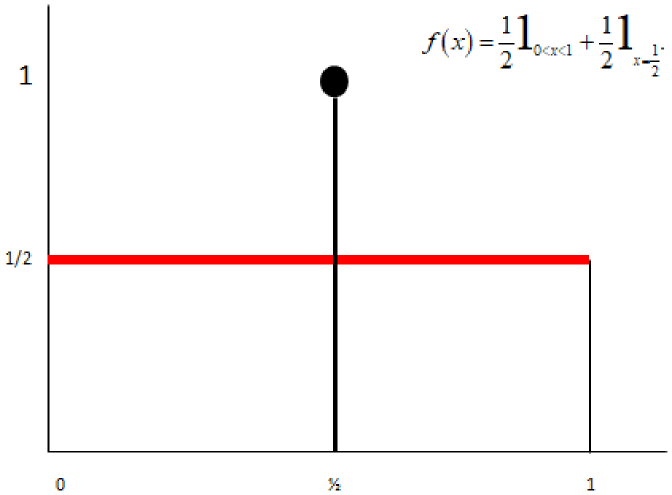


**Figure 4.** Probability as both area and point mass.

If we think of probability as only interval measure (sometimes referred to as Lebesgue measure) or as only point mass, $f(x)$ cannot be a probability function. However, can we allow it to be both? How would we accumulate probability in this matter?

Remembering our first discussion about the concept of measure theory we can see how we might try this. For example to compute $\int_\Omega d\mathbf{P}$ we can start at $x=0$ and accumulate probability allowing x to increase.* In this open interval we accumulate probability using the measure as area concept to $x=\frac{1}{2}$. This gives us $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)=\frac{1}{4}$. At $x=\frac{1}{2}$, we now switch to point mass, accumulating another $\frac{1}{2}$. Finally for x = $\frac{1}{2}$ to 1, we return to accumulating probability as area, accruing $\frac{1}{4}$. The sum of these probabilities is one after all.

## Continuous joint distributions

We have developed the concept of joint distributions when the probability measure is discrete. However, the role of joint distributions when the random variables have continuous measure is also of great value and is the topic of the following set of discussions. In fact many of the definitions and concepts proceed similarly to their discrete analog.

The joint distribution of two random variables $X$ and $Y$ is simply $f_{XY}(x,y)$. It defines measure not just in the $x$-domain, but that of $y$ as well. For example, if we define

$$f_{XY}(x,y)=1_{0\le x\le 1,0\le y\le 1},$$

then measure is assigned on the unit square, with all other regions of the $(x,y)$ plane having measure 0 assigned to them. Thus we can show that

* We could start at $x=-\infty$. and accumulate probability zero all the way to the beginning of the (0, 1) interval as well.

$$\int_{\Omega_{x,y}} d\mathbf{P} = \int_{\Omega_{x,y}} f_{XY}(x,y) = \int_{\Omega_{x,y}} 1_{0\le x\le 1, 0\le y\le 1} = \int_0^1\int_0^1 dxdy = 1.^*$$

We can find probabilities on a set $A$ where $\left\{x,y: 0 \le x \le \frac{1}{4}, 0 \le y \le \frac{3}{4}\right\}$ by simply defining the indicator variable $1_A = 1_{0\le x\le \frac{1}{4},\, 0\le y\le \frac{3}{4}}$ and computing

$$\mathbf{P}[A] = \mathbf{E}\left[1_A\right] = \int_{\Omega_{x,y}} 1_A d\mathbf{P} = \int_{\Omega_{x,y}} 1_{0\le x\le \frac{1}{4},\, 0\le y\le \frac{3}{4}} 1_{0\le x\le 1, 0\le y\le 1}$$

$$= \int_0^{\frac{3}{4}}\int_0^{\frac{1}{4}} dxdy = \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{3}{16}.$$

It is not two difficult to see that if each of $X$ and $Y$ follow the uniform distribution on $(0,1)$, then $f_{XY}(x,y) = f_X(x)f_Y(y)$. This is the definition of independence, a statement that is the exact counterpart to that for discrete random variables. It also follows that probabilities involving only $X$ are based on the measuring tool $f_X(x)$, since the marginal distribution of $X$ (or the distribution of $X$ by itself) is

$$\int_{\Omega_Y} f_{X,Y}(x,y) = \int_{\Omega_Y} f_X(x)f_Y(y) = f_X(x)\int_{\Omega_Y} f_Y(y) = f_X(x).$$

Thus $\mathbf{P}[X \le 1]$ we see to be $\frac{1}{2}$, a probability that we can compute without giving any consideration to the random variable $Y$.

However, suppose the joint measuring tool for two continuous random variables $X$ and $Y$ was $f_{XY}(x,y) = \frac{1}{2}1_{0\le x\le y<2}$. This is a probability density for which $X$ is clearly dependent on $Y$ and is confined to a smaller region that the original independent random variables (Figure 1).

We can convince ourselves easily that $\int_{\Omega_{x,y}} d\mathbf{P} = 1$ from

$$\int_{\Omega_{x,y}} d\mathbf{P} = \frac{1}{2}\int_0^2\int_0^y dxdy = \frac{1}{2}\int_0^2 y\,dy = \left[\frac{y^2}{4}\right]_0^2 = \frac{4}{4} = 1.$$

These two densities provide different solutions for the same probabilities. Recall that when $X$ and $Y$ were independent, $\mathbf{P}[X \le 1] = \frac{1}{2}$. For this new joint density, the situation is more complicated and we must consider the two cases that $X \le 1$ when $Y \le 1$ or when $Y > 1$ (Figure 5).

Thus,

[*] Note that we are converting a double integral into two iterated integrals due to Fubini's theorem.

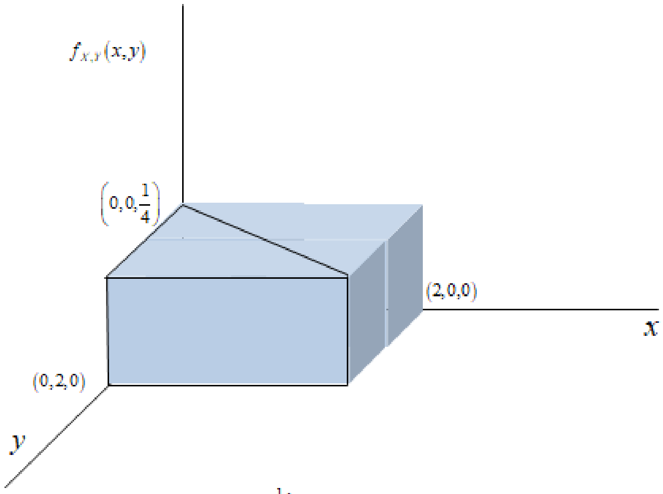


**Figure 5.** The joint density $f_{X,Y}(x,y)=\frac{1}{2}1_{0\le x\le y<2}$ has positive measure on the wedge

$$\mathbf{P}\left[X\le 1\right]=\mathbf{P}\left[X\le 1\cap Y\le 1\right]+\mathbf{P}\left[X\le 1\cap Y>1\right]$$

$$=\frac{1}{2}\int_0^1\int_0^y \mathbf{1}_{0\le x\le y<2}\,dxdy+\frac{1}{2}\int_1^2\int_0^1 \mathbf{1}_{0\le x\le y<2}\,dxdy$$

$$=\frac{1}{2}\int_0^1 y\,dy+\frac{1}{2}\int_1^2 dy=\frac{1}{2}\left(\left[\frac{y^2}{2}\right]_0^1+\left[y\right]_1^2\right)$$

$$=\frac{1}{2}\left(\frac{1}{2}+1\right)=\frac{3}{4}.$$

We can also identify the marginal probability of *X*. We compute

$$f_X(x)=\int_{\Omega_y} f_{X,Y}(x,y)=\int_{\Omega_y}\frac{1}{2}\mathbf{1}_{0\le x\le y<2}=\frac{1}{2}\mathbf{1}_{0\le x<2}\int_x^2 dy$$

$$=\frac{1}{2}\mathbf{1}_{0\le x<2}\left(2-x\right)=\left(1-\frac{x}{2}\right)\mathbf{1}_{0\le x<2}.$$

Note that this density is quite different than that of the random variable *X* when *X* and *Y* were independent .

We can find $\mathbf{P}\left[X\le 1\right]$ directly with no consideration of the random variable *Y*.

$$\mathbf{P}\left[X\le 1\right]=\int_0^1 d\mathbf{P}=\int_0^1\left(1-\frac{x}{2}\right)\mathbf{1}_{0\le x<2}=\int_0^1\left(1-\frac{x}{2}\right)dx$$

$$=\left[x-\frac{x^2}{4}\right]_0^1=1-\frac{1}{4}=\frac{3}{4},$$

The solution that we obtained previously. Similarly, we can find the marginal distribution of the random variable *Y*.

$$f_Y(y)=\int_{\Omega_x} f_{X,Y}(x,y)=\int_{\Omega_x}\frac{1}{2}\mathbf{1}_{0\le x\le y<2}=\frac{1}{2}\mathbf{1}_{0\le y<2}\int_0^y dx$$

$$=\frac{1}{2}\mathbf{1}_{0\le y<2}\,y=\frac{y}{2}\mathbf{1}_{0\le y<2}.$$

Note that while the joint distribution of our two random variables seems "similar" to the original joint distribution (i.e., uniform on unit square), the marginal distributions in this case are not uniform.

Two important formulas for us are

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}: \quad f_{Y|X}(y \mid x) = \frac{f_{Y,X}(x,y)}{f_X(x)}$$

This now permits us to compute the conditional distribution of $X$ given the value of the random variable $Y$, $f_{X|Y}(x \mid y)$, as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{2}1_{0 \le x \le y < 2}}{\frac{y}{2}1_{0 \le y < 2}} = \frac{1}{y}1_{0 \le x \le y < 2}.$$

This is the uniform distribution. We note that $f_{X|Y}(x \mid y)$ is a function of not just the random variable $X$ but $Y$ as well. In this case the upper bound of the uniform distribution for $X$ is $Y$, but that $Y$ must be less than two.

In a similar fashion,

$$f_{Y|X}(y \mid x) = \frac{f_{Y,X}(x,y)}{f_X(x)} = \frac{\frac{1}{2}1_{0 \le x \le y < 2}}{\left(1 - \frac{x}{2}\right)1_{0 \le x < 2}} = \frac{1}{(2-x)}1_{0 \le x \le y < 2}.$$

The conditional distribution of the random variable y is also uniformly distributed. The lower bound of this distribution is $2 - x$ and its upper bound is 2.

## Joint and conditional expectations

We can formulate the expectation of a function of the random variables $X$ and $Y$, $g(x,y)$ simply as

$$\mathbf{E}\left[g(x,y)\right] = \int_{\Omega_{X,Y}} g(x,y)\,d\mathbf{P} = \int_{\Omega_{X,Y}} g(x,y) f_{X,Y}(x,y).$$

For example, if $f_{X,Y}(x,y) = \frac{1}{2}1_{0 \le x \le y < 2}$, and $g(x,y) = (y-x)^2$, we write

$$
\begin{aligned}
\mathbf{E}\left[g(x,y)\right] = \mathbf{E}\left[(y-x)^2\right] \int\limits_{\Omega_{X,Y}} (y-x)^2\, d\mathbf{P} &= \int\limits_{\Omega_{X,Y}} (y-x)^2 f_{X,Y}(x,y) \\
&= \frac{1}{2}\int_0^2\int_0^y \left(y^2 - 2xy + x^2\right)\mathbf{1}_{0\le x\le y<2}\, dxdy \\
&= \frac{1}{2}\left[\int_0^2 y^2 \int_0^y dxdy - 2\int_0^2 y \int_0^y x\, dxdy + \int_0^2\int_0^y x^2 dxdy\right] \\
&= \frac{1}{2}\left[\int_0^2 y^3 dy - 2\int_0^2 \frac{y^3}{2} dy + \int_0^2 \frac{y^3}{3} dy\right] \\
&= \frac{1}{2}\left[4 - 4 + \frac{4}{3}\right] = \frac{2}{3}.
\end{aligned}
$$

However, we can also compute conditional expectations. To follow our development, we write

$$
\begin{aligned}
\mathbf{E}_{X|Y}\left[g(x)\right] &= \int\limits_{\Omega_X} g(x)\, d\mathbf{P} = \int\limits_{\Omega_X} g(x) f_{X|Y}(x\,|\,y) \\
\mathbf{E}_{Y|X}\left[g(y)\right] &= \int\limits_{\Omega_Y} g(y)\, d\mathbf{P} = \int\limits_{\Omega_Y} g(y) f_{Y|X}(y\,|\,x)
\end{aligned}
$$

From our previous work in this section and our knowledge of the moments of the uniform distribution, we compute $\mathbf{E}_{X|Y}\left[x\right] = \frac{y}{2}$, $\mathbf{E}_{Y|X}\left[x\right] = \frac{x}{2} + 1$.

### *Double expectations*

A concept of interest that commonly simplifies our work in computing expectations for joint probabilities is that of taking sequential expectations. We know that that

$$
\mathbf{E}\left[g(x,y)\right] = \int\limits_{\Omega_{X,Y}} g(x,y)\, d\mathbf{P} = \int\limits_{\Omega_{X,Y}} g(x,y) f_{X,Y}(x,y).
$$

Since $f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, or $f_{X,Y}(x,y) \;=\; f_{X|Y}(x\,|\,y)\, f_Y(y)$, so we may write $\mathbf{E}\left[g(x,y)\right] = \int\limits_{\Omega_{X,Y}} g(x,y) f_{X|Y}(x\,|\,y)\, f_Y(y)$. We now simply expand the integration to see

$$
\begin{aligned}
\mathbf{E}\left[g(x,y)\right] &= \int\limits_{\Omega_{X,Y}} g(x,y) f_{X|Y}(x\,|\,y)\, f_Y(y) \\
&= \int\limits_{\Omega_Y}\left[\int\limits_{\Omega_{X|Y}} g(x,y)\, f_{X|Y}(x\,|\,y)\right] f_Y(y) \\
&= \int\limits_{\Omega_Y} \mathbf{E}_{X|Y}\left[g(x,y)\right] f_Y(y) \\
&= \mathbf{E}_Y\left[\mathbf{E}_{X|Y}\left[g(x,y)\right]\right]
\end{aligned}
$$

This result that $\mathbf{E}\left[g(x,y)\right]= \mathbf{E}_Y\left[\mathbf{E}_{X|Y}\left[g(x,y)\right]\right]$ or equivalently $\mathbf{E}_X\left[\mathbf{E}_{Y|X}\left[g(x,y)\right]\right]$ can provide important simplifications in our calculations for example, when we identify the distribution of a matrix determinant.

The concept of double expectation argument also applies to computing the variance. We may write

$\mathbf{Var}\left[X\right]=\mathbf{Var}_X\left[\mathbf{E}\left[Y\mid X\right]\right]+\mathbf{E}_X\left[\mathbf{Var}\left[Y\mid X\right]\right]$. We will apply this to stochastic processes, e.g., the immigration-death model

### *Example: Myocardial infarction and death*

One of the sequela of a heart attack or myocardial infarction is death. When deaths occur following a heart attack they are more likely to be earlier rather than later, that is, the six month post myocardial infarction rate is greater than the six month to one year rate, which is greater than the subsequent six month rate, etc.

A study is being conducted that follows patients at elevated risk of a heart attack  Let *R* be the random variable reflecting the time from the beginning of the study until the time suffers a heart attack. Let *S* be the survival time of the patient up to time *T* which is the time the follow-up period ends.  Then let's define the joint distribution of *R* and *S* as

$$f_{R,S}\left(r,s\right)=\left(\lambda+\mu\right)\lambda e^{-(\lambda s+\mu r)}1_{0\le r\le s\le\infty}$$

Since the time to infarct must precede the survival time which is terminated at the end of the study, we are required to restrict the region of positive measure to $1_{0\le r\le s\le\infty}$. We can demonstrate

$$\begin{aligned}\int_{\Omega_{R,S}} d\mathbf{P} &= \int_{\Omega_{R,S}} f_{R,S}\left(r,s\right)=\int_{\Omega_{R,S}}\left(\lambda+\mu\right)\lambda e^{-(\lambda s+\mu r)}1_{0\le r\le s\le\infty}\\ &=\frac{\lambda+\mu}{\mu}\int_0^\infty \lambda e^{-\lambda s}ds\int_0^s \mu e^{-\mu r}dr=\frac{\lambda+\mu}{\mu}\int_0^\infty \lambda e^{-\lambda s}\left(1-e^{-\mu s}\right)ds\\ &=\frac{\lambda+\mu}{\mu}\left[1-\frac{\lambda}{\lambda+\mu}\right]=\frac{\lambda+\mu}{\mu}\left[\frac{\mu}{\lambda+\mu}\right]\quad=1\end{aligned}$$

We can also find the marginal density of *R* as

$$\begin{aligned}f_R\left(r\right) &= \int_{\Omega_S} f_{R,S}\left(r,s\right) = \int_{\Omega_S}\left(\lambda+\mu\right)\lambda e^{-(\lambda s+\mu r)}1_{0\le r\le s\le\infty}\\ &=\left(\lambda+\mu\right)1_{0\le r\le\infty}e^{-\mu r}\int_r^\infty \lambda e^{-\lambda s}ds\quad=\left(\lambda+\mu\right)e^{-\mu r}e^{-\lambda r}1_{0\le r\le\infty}\\ &=\left(\lambda+\mu\right)e^{-(\lambda+\mu)r}1_{0\le r\le\infty}.\end{aligned}$$

Thus, *R* follows a negative exponential measure with parameter $\lambda+\mu$. We can easily compute that a heart attack occurs in the $\left[0,T\right]$ interval as

$1-e^{-(\lambda+\mu)T}$. The marginal survival distribution, $f_S\left(s\right)$ can also be computed,

$$f_S(s) = \int_{\Omega_S} f_{R,S}(r,s) = \int_{\Omega_R} (\lambda+\mu)\lambda e^{-(\lambda s+\mu r)} dr\, ds\mathbf{1}_{0\le r\le s\le\infty}$$

Continuing,

$$\begin{aligned}&=\frac{\lambda+\mu}{\mu}\lambda e^{-\lambda s}\mathbf{1}_{0\le s\le\infty}\int_0^s \mu e^{-\mu r}dr = \frac{\lambda+\mu}{\mu}\lambda e^{-\lambda s}\left(1-e^{-\mu s}\right)\mathbf{1}_{0\le s\le\infty}\\ &=\frac{\lambda}{\mu}(\lambda+\mu)e^{-\lambda s}\left(1-e^{-\mu s}\right)\mathbf{1}_{0\le s\le\infty}.\end{aligned}$$

We can compute $\mathbf{E}[S]$ as

$$\begin{aligned}\mathbf{E}[S] &= \int_{\Omega_S} s\,d\mathbf{P} = \int_{\Omega_S} s\frac{\lambda}{\mu}(\lambda+\mu)e^{-\lambda s}\left(1-e^{-\mu s}\right)\mathrm{ds}\,\mathbf{1}_{0\le s\le\infty}\\ &=\frac{\lambda}{\mu}(\lambda+\mu)\left[\int_0^\infty se^{-\lambda s}ds-\int_0^\infty se^{-(\lambda+\mu)s}ds\right]\\ &=\frac{\lambda}{\mu}(\lambda+\mu)\left[\int_0^\infty se^{-\lambda s}ds-\int_0^\infty se^{-(\lambda+\mu)s}ds\right]\\ &=\frac{\lambda}{\mu}(\lambda+\mu)\left[\frac{1}{\lambda^2}-\frac{1}{(\lambda+\mu)^2}\right]\end{aligned}$$

relying on what we know about the gamma function. We can compute the probability a death occurs in time $[0,T]$ as

$$\begin{aligned}\mathbf{P}[S\le T] &= \frac{\lambda}{\mu}(\lambda+\mu)\int_0^T e^{-\lambda s}\left(1-e^{-\mu s}\right)ds\\ &= \frac{(\lambda+\mu)}{\mu}\left[\lambda\left(1-e^{-\lambda T}\right)-\frac{\lambda}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)T}\right)\right]\\ &= \frac{(\lambda+\mu)}{\mu}\left(1-e^{-\lambda T}\right)-\frac{\lambda}{\mu}\left(1-e^{-(\lambda+\mu)T}\right).\end{aligned}$$

With these marginal results available to us, we can compute the conditional probability density functions. For example the rule that governs the probability of a death given the infarct time or $f_{S|R}(s\,|\,r)$ is

$$f_{S|R}(s\,|\,r)=\frac{f_{R,S}(r,s)}{f_R(r)}=\frac{(\lambda+\mu)\lambda e^{-(\lambda s+\mu r)}\mathbf{1}_{0\le r\le s\le\infty}}{(\lambda+\mu)e^{-(\lambda+\mu)r}\mathbf{1}_{0\le r\le\infty}}=\lambda e^{-\lambda(s-r)}\mathbf{1}_{r\le s<\infty.}$$

This is a negative exponential random variable, with its lower bound truncated at $s=r.$ The conditional distribution of $f_{R|S}(r\,|\,s)$ is

$$\begin{aligned}f_{R|S}(r\,|\,s)&=\frac{f_{R,S}(r,s)}{f_S(s)}=\frac{(\lambda+\mu)\lambda e^{-(\lambda s+\mu r)}\mathbf{1}_{0\le r\le s\le\infty}}{\frac{\lambda}{\mu}(\lambda+\mu)e^{-\lambda s}\left(1-e^{-\mu s}\right)\mathbf{1}_{0\le s\le\infty}}\\ &=\frac{\mu e^{-\mu r}}{\left(1-e^{-\mu s}\right)}\mathbf{1}_{0\le r\le s}.\end{aligned}$$

This is another truncated negative exponential distribution, “trapped” between 0 and $s$. It is these types of distributions that are commonly used in designing clinical trials where the probabilities of events must be estimated to compute the minimum number of subjects required for the study.

Next sections

Uniform and Beta Measure
Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# Uniform and Beta Measure

The uniform probability distribution is perhaps one of the easiest distributions using the concept of measure as interval length, and is a natural one to start with. We will begin with some simple concepts first, providing some examples of how this function is used. In doing so, we will provide an introduction to what is commonly known as the transformation of variables, and also begin to work with some simple linear combination of random variables. We will then introduce the family of beta distributions, a family of probability functions of which the uniform distribution is a member.

## Prerequisites

Properties of Probability
Gamma Function
Measurable Functions
Measure and Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Introduction to Continuous Probability Functions

### *Introduction*

The hallmark of the uniform probability distribution is the observation that intervals of equal length have equal probability (Figure 1).[*]

[*] This is Lebesgue measure.

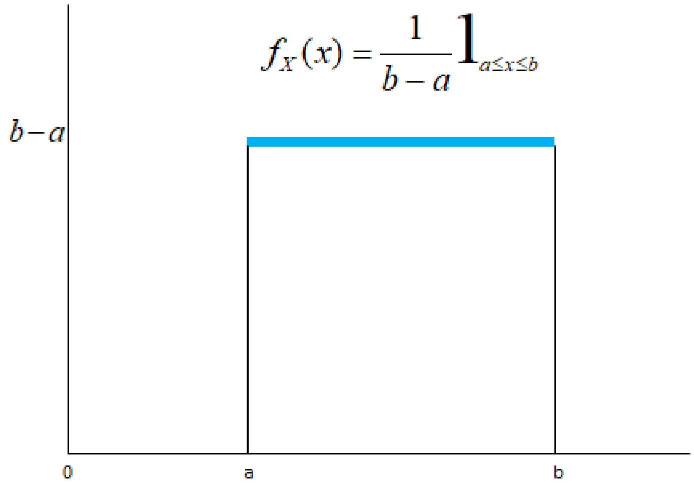


**Figure 1**. Uniform distribution on the [a, b] interval. Intervals of equal length have equal probability.

We define uniform measure (the probability function) as

$$f_X(x) = \frac{1}{b-a}\mathbf{1}_{a \le x \le b}$$

and say that *X* follows a *U*(*a*, *b*) distribution. We can see at once that

$$\int_\Omega d\mathbf{P} = \int_\Omega \frac{1}{b-a}\mathbf{1}_{a \le x \le b} = \frac{1}{b-a}\int_\Omega \mathbf{1}_{a \le x \le b} = \frac{1}{b-a}\int_a^b dx = \frac{b-a}{b-a} = 1.$$

### *Probability and cumulative distribution function*

In order to find $\mathbf{P}\left[c < X < d\right]$ when $a \le c < d \le b$ we write a $\mathbf{E}_X\left[\mathbf{1}_{c<X<d}\right]$ and compute

$$\begin{aligned}\mathbf{E}_X\left[\mathbf{1}_{c<X<d}\right] &= \int_\Omega \mathbf{1}_{c<X<d}\frac{1}{b-a}\mathbf{1}_{a \le x \le b}\,dx = \frac{1}{b-a}\int_\Omega \mathbf{1}_{c<X<d}\mathbf{1}_{a \le x \le b}\,dx \\ &= \frac{1}{b-a}\int_\Omega \mathbf{1}_{c<X<d}\,dx = \frac{1}{b-a}\int_\Omega \mathbf{1}_{c<X<d}\,dx = \frac{1}{b-a}\int_c^d dx \\ &= \frac{d-c}{b-a}.\end{aligned}$$

Use of the element functions may appear a bit tedious, but they help us by keeping the regions of measure clear and unambiguous. Expanding on the previous example, we can see that the cumulative distribution function $F_X(x)$ is simply $F_X(x) = \frac{x-a}{b-a}\mathbf{1}_{a<x<b}$ and is a linear function of *x*.

## Example – Laceration lengths

Assume in a rural emergency department that the lacerations of patients arriving for wound repair are uniformly distribution between 1 and 6 cms. What is the probability that lacerations are greater than 4 cms in length?

This is simply

$$\mathbf{P}\left[4 < x < 6\right] = \mathbf{E}\left[\mathbf{1}_{4<x<6}\right] = \int_{\Omega} \mathbf{1}_{4<x<6}\, d\mathbf{P} = \frac{1}{5}\int_{\Omega} \mathbf{1}_{4<x<6}\mathbf{1}_{1<x<6}\, dx$$

$$= \frac{1}{5}\int_{4}^{6} dx = \frac{2}{5}.$$

If we assume that laceration lengths are independent of each other, what is the probability that of the next twelve patients, no more than five have lacerations as long or longer than 4 cms?

For this probability we return to the binomial distribution, where a "success" is defined as a laceration of at least four cms in length. The uniform distribution supplies this probability and we write

$$\mathbf{P}\left[Y \le 5\right] = \sum_{k=0}^{5} \binom{12}{k} (0.40)^{5} (0.60)^{12-k} = 0.665.$$

While we saw that the ultimate probability problem to be solved was one involving a binomial random variable, its probability needed to be supplied by the uniform distribution.

■

## Moments of the uniform distribution

Means and variances from the uniform distribution are readily available. To compute $\mathbf{E}\left[X\right]$, we simply write

$$\mathbf{E}\left[X\right] = \int_{\Omega} x\, d\mathbf{P} = \int_{\Omega} x \frac{1}{b-a}\mathbf{1}_{a \le x \le b} = \frac{1}{b-a}\int_{a}^{b} x\, dx$$

$$= \frac{b^2 - a^2}{2\left(b-a\right)} = \frac{b+a}{2}$$

The variance requires some additional algebra but can be shown to be $\dfrac{\left(b-a\right)^2}{12}$.

## Layered cake: Continuous variables

Recall that we demonstrated for nonnegative discrete random variables, $\mathbf{E}\left[X\right] = \sum_{k=0}^{\infty} \mathbf{P}\left[X > k\right]$. The analogue for continuous nonnegative random variables is $\mathbf{E}\left[X\right] = \int_{0}^{\infty} 1 - F_X(x)\, dx.$

The proof is based on the observation that there are multiple (in this circumstance, there are two relevant) ways to take the measure of a region. Begin by writing $\int_{0}^{\infty} 1 - F_X(x)\, dx = \int_{0}^{\infty} dt \int_{t}^{\infty} f_X(x) dx.$ This divides the region of nonnegative reals into two region, $t \le x \le \infty$, and $0 \le t \le \infty$. Also note the difference in the measures. The random variable $X$ is using its probability density as the measuring tool, but $t$ is simply using Lebesgue measure, or measure as interval length.

Yet, the same region may be mapped as $0 \le x \le t$ and $0 \le x \le \infty$. Thus, we may continue by writing

$$\int_0^\infty 1 - F_X(x)\,dx = \int_0^\infty dt \int_t^\infty f_X(x)dx = \int_0^\infty f_X(x)dx \int_0^x dt.$$

The measuring tools have stayed the same, but the regions have changed. We may finish as

$$\int_0^\infty 1 - F_X(x)\,dx = \int_0^\infty dt \int_t^\infty f_X(x)dx = \int_0^\infty f_X(x)dx \int_0^x dt$$

$$= \int_0^\infty x f_X(x)dx = \mathbf{E}\left[X\right].$$

And we have the layered cake demonstration for continuous random variables.*

## Transformation of random variables

Commonly we will find that we are not ultimately interested in a random variables whose probability function we know, but instead in a random variable related to one that we know. In order to obtain the function in which we have interest, we must become facile with converting or transforming one random variable to another.

Let's start with the random variable $X$ that follows a $U$(0,2) distribution, allowing us to write $f_X(x) = \frac{1}{2}\mathbf{1}_{0<x<2}$. This measuring tool has all of the properties that we have demonstrated thus far in this section. Now suppose we have a new random variable, $Y = 3X$. What is the probability distribution of $Y, f_Y(y)$?

Before we begin a formal evaluation, we can think through what we would expect. The minimum value of $Y$ will be zero, since this is the minimum value of $x$. However, the maximum value will be (3)(2) = 6, three times the maximum value of $X$. Nothing about this transformation suggest that we will get away from intervals of equal length having equal probability, but the probability will need to be spread out covering not just (0, 2) interval but the (0, 6) interval.

Now, how can we demonstrate that? There two new ways that we will explore in this section.† The first is to remember that if we find the cumulative distribution function $F_Y(y)$ and it is continuous, we only need differentiate it to obtain the probability density function. Remembering that for $Y = 3X$, we begin

$$F_Y\left(y\right) = \mathbf{P}\left[Y \le y\right] = \mathbf{P}\left[3X \le y\right].$$

Note here that we have changed the random variable that we are working with from $Y$ to $X$. Continuing, we have

$$\mathbf{P}\left[3X \le y\right] = \mathbf{P}\left[X \le \frac{y}{3}\right] = F_X\left(\frac{y}{3}\right).$$

* If it is true that the measure theoretic approach covers both discrete and continuous random variables, then it stands to reason that one measure theoretic approach to the layered cake observation would suffice. This is the case.

† A third way is to find the moment generating function of the new randon variables Y that we have already discussed

So we now have the relationship $F_Y(y) = F_X\left(\frac{y}{3}\right)$.

We know though that if $W$ is U($a$, $b$) then the cumulative distribution function that $F_W(w) = \frac{w-a}{b-a}\mathbf{1}_{a<W<b}$. In this case $F_X(x) = \frac{x}{2}\mathbf{1}_{0<X<2}$, and therefore $F_X\left(\frac{y}{3}\right) = \frac{y}{6}\mathbf{1}_{0<Y<6}$. Note how the region of the element function changed now that we are writing the cumulative distribution function as a function of $y$. Since $F_Y(y) = \frac{y}{6}\mathbf{1}_{0<Y<6}$. is a smooth function with existing derivatives on $0 \le y \le 6$, we write $f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{6}\mathbf{1}_{0\le Y\le 6}$

Which is the density for a $U$(0,6) random variable.

Another approach is to use the following helpful formula from transformation of variable

$$f_Y(y) = f_X(y)\left[dx \to dy\right]\left[\Omega_x \to \Omega_y\right]$$

This is a transformation from $X$ to $Y$. It contains three components. The first is to write the probability density in $X$ in terms of $Y$. The second is to include the derivative, which is the scale factor. The third component changes the set on which $X$ has positive probability to that which has $Y$. In this example we have

$$f_Y(y) = f_X(x) = \frac{1}{2}\frac{1}{3}\mathbf{1}_{0<y<6} = \frac{1}{6}\mathbf{1}_{0<y<6}$$

As another example, assume $X$ follows a $U$(0,1) distribution. Let $Y = a + (b-a)X$. If we use the density approach we begin with $f_X(x) = \mathbf{1}_{0<x<1}$. Consider first the new range of $y$. For $x = 0$, then $y = a$, and for $x = 1$, $y = b$. Also, $X = \frac{Y-a}{b-a}$ then $dX = \frac{dY}{b-a}$. We now have $f_Y(y) = \frac{1}{b-a}\mathbf{1}_{a<Y<b}$, and we have seen how we can convert a $U$(0,1) to a $U$($a$, $b$).

As another example, let $X$ follow a U(0,1) distribution. Let $Y = +\sqrt{X}$ so the transformation is one to one. What then is $f_y(y)$?

The $+\sqrt{X}$ is between 0 and 1 for $0 < X < 1$, so the region does not change. Since $X = Y^2$ then $dX = 2Ydy$ and we write $f_Y(y) = 2y\mathbf{1}_{0<y<1}$. We can see that this integrates to one on (0,1).

However, suppose we let $W = \sqrt{X}$? Although $X$ is nonnegative, $W$ is not. A methodologic approach is to find the density for $y \ge 0$ and $y < 0$, and then make the adjustment for the two different mappings.

We already have the solution for positive $Y > 0$. For $Y < 0$, we note that , $-1 \le Y \le 0$, $X = (-Y)^2$, $|dX| = 2|Y|dy$, and we would have $f_Y(y) = 2|y|\mathbf{1}_{-1<y<0}$. So our first attempt at the solution would be

$$f_Y(y) = 2y\mathbf{1}_{0<y<1} + 2|y|\mathbf{1}_{-1<y<0} \text{ ??}$$

However, the one to two mapping needs to be taken into account. We do this by halving the probability for each region. The ½ is needed to ensure that the measure over the sample space is one. We therefore conclude

$$f_Y\left(y\right) = y\mathbf{1}_{0<y<1} + \left|y\right|\mathbf{1}_{-1<y<0}.$$

Which integrates to one on the set $-1 \le Y \le 1$.

It is very easy to fall into the trap of simply computing $f_Y(y) = f_X(y)\left|\frac{dx}{dy}\right|$ when carrying out transformations, implicitly assuming that $f_Y(y) = f_X(y)\left[dx \to dy\right]$. However, without explicit consideration of the region, it is all too easy to obtain the wrong result, never being quite sure why it was wrong. In order to be sure that the entire transformation is covered, it is critical to consider the region as well, or

$$f_Y(y) = f_X(y)\left[dx \to dy\right]\left[\Omega_x \to \Omega_y\right].$$

## Probability integral transformation

Perhaps one of the most useful transformation involving the uniform distribution is not the transformation away from the uniform distribution but a transform into it. Let *X* have a probability distribution with cumulative distribution function $F_X(x)$. Let's define the new random variable $Y = F_X(x)$. What is the actual probability distribution of *Y*?

Here, the experiment is to choose a value of *x* randomly, then compute $\mathbf{F}_X(x)$. We know at once that $0 \le Y \le 1$. Is there anything else that we can deduce? Using the cumulative distribution approach, we can calculate

$$\begin{aligned}\mathbf{F}_Y(y) &= \mathbf{P}\left[Y \le y\right] = \mathbf{P}\left[\mathbf{F}(x) \le y\right] = \mathbf{P}\left[x \le \mathbf{F}^{-1}\left(y\right)\right] \\ &= \mathbf{F}\left(\mathbf{F}^{-1}\left(y\right)\right) = y.\end{aligned}$$

This is the cumulative distribution function of the *U*(0,1) distribution. This transform, known as the probability integral transformation, is the basis for converting uniformly distributed random variables which are commonly easy to generate into random variables that follow more complex distributions.

## Sums of uniform random variable

The distribution function technique just discussed is very helpful when considering the sums of uniform random variables. These types of examples afford fine practice in managing the event space, which in this case is the region of integration. Letting *X* and *Y* are i.i.d. *U*(0,1), the complete computations for the probability density function for *X* + *Y*, *X* – *Y*, *XY* and $\frac{X}{Y}$ are available, Taking the result from this work, we that if $Z = X + Y$, then

$$f_Z(z) = z\mathbf{1}_{0\le z\le 1.} + \left(2-z\right)\mathbf{1}_{1<z\le 2}.$$

We first note that this is not the density for a uniformly distributed random variable. In fact, the density increases to a value of 1, then decreases to zero for the largest possible value $z = 2$. We also see that there is a radical departure in the shape of the distribution (Figure 2).

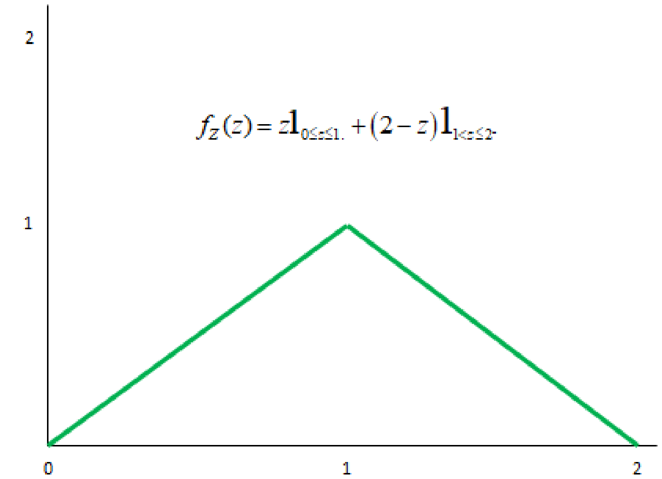


**Figure 2.** Probability density function of the sum of two U( 0,1) random variables.

The uniform distributions from which the random variable $Z$ was formed provide equal probabilities for equal interval lengths, regardless of the location of the interval on [0,1]. However, the density of $Z$ is quite different. It places higher probability on intervals closer to one, with lower probabilities in the intervals closer to the extremes. For example. If we want to

$$\mathbf{P}\left[Z \le \frac{1}{2}\right] = \mathbf{E}\left[1_{0\le z\le \frac{1}{2}}\right] = \int_{\Omega_x} 1_{0\le z\le \frac{1}{2}}\, d\mathbf{P} = \int_{\Omega_x} 1_{0\le z\le \frac{1}{2}}\left[z1_{0\le z\le 1.} + (2-z)1_{1<z\le 2.}\right]$$

find $\mathbf{P}\left[Z \le \frac{1}{2}\right]$, we compute

$$= \int_0^{\frac{1}{2}} z\,dz = \left.\frac{z^2}{2}\right]_0^{\frac{1}{2}} = \frac{1}{8}.$$

Another interval of equal length is $\frac{3}{4} \le x \le \frac{5}{4}$. To compute $\mathbf{P}\left[\frac{3}{4} \le x \le \frac{5}{4}\right]$, we find

$$\mathbf{P}\left[\frac{3}{4} \le x \le \frac{5}{4}\right] = \mathbf{E}\left[1_{\frac{3}{4}\le x\le \frac{5}{4}}\right] = \int_{\Omega_x} 1_{\frac{3}{4}\le x\le \frac{5}{4}}\, d\mathbf{P} = \int_{\Omega_x} 1_{\frac{3}{4}\le x\le \frac{5}{4}}\left[z1_{0\le z\le 1.} + (2-z)1_{1<z\le 2}\right]$$

Continuing

$$= \int_{\frac{3}{4}}^{1} z\,dz + \int_{1}^{\frac{5}{4}} (2-z)\,dz \;=\; \left.\frac{z^2}{2}\right]_{\frac{3}{4}}^{1} + \left.\left(2z - \frac{z^2}{2}\right)\right]_1^{\frac{5}{4}}$$

$$= \left(\frac{1}{2} - \frac{9}{32}\right) + \left(\frac{10}{4} - \frac{25}{32}\right) - \left(2 - \frac{1}{2}\right) = \frac{7}{16}.$$

It's useful to compare $\frac{\mathbf{P}\left[\frac{3}{4} \le x \le \frac{5}{4}\right]}{\mathbf{P}\left[0 \le x \le \frac{1}{2}\right]} = \frac{7/16}{1/8} = 3.5$. For intervals of length, the probability of the central interval is 3.5 times as large as the probability as the extreme interval. This tendency to construct the central movement of probability from probability distributions that have no particular property of central tendency is at the heart of the central limit theorem, perhaps the most useful probability theorem of them all.

## The Beta distribution

This distribution represents a family of distributions. It is the first of several distributions that we will discuss that has two parameters, *α* and *β*. It's probability density function $f_X(x)$ is

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} 1_{0 \le x \le 1}$$

Note that $\alpha = \beta = 1$ reduces this to the *U*(0,1) distribution.

The demonstration that this function integrates to one requires the use of double integrals a transformation and is straightforward. It is worth reviewing as it will take some of the mystery out of this distributional form.

The distribution has many shapes, governed by its two parameters (Figure 3).This flexibility has broadened its use, particularly in Bayesian modeling.

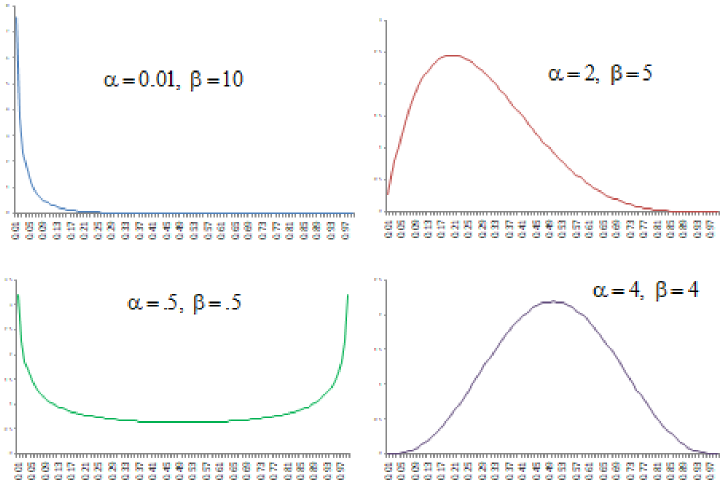


**Figure 3.** Different forms of the beta distribution

### *Moments of the beta distribution*

Computation of the moments of the beta distribution are not only useful for their results, but for the use of a tool of a great interest in computing probabilities. Let's begin with the first moment of a random variable *X* that follows a beta (α, β) distribution.

$$\begin{aligned}\mathbf{E}\left[X\right] &= \int_{\Omega_X} x d\mathbf{P} = \int_0^1 x\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\,dx \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta-1}\,dx\end{aligned}$$

The “trick” is so see that the term in the integrand $x^{\alpha}(1-x)^{\beta-1}$ is “almost the density of a beta distribution. We need the constant $\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}$ to make it so. Since that is not the constant we already have, we simply remove the present constant from the integral, multiplying the resultant integrand by $\frac{\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}}{\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}}$ to keep the equality. Thus,

$$\begin{aligned}\mathbf{E}\left[X\right] &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta-1}\,dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{\alpha}(1-x)^{\beta-1}\,dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta-1}\,dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\end{aligned}$$

And noting that $\Gamma(\alpha+\beta+1)=(\alpha+\beta)\Gamma(\alpha+\beta)$ and $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$, we can further simply to see $\mathbf{E}\left[X\right]=\frac{\alpha}{\alpha+\beta}$.

We can carry out this same style of computation to find $\mathbf{E}\left[X^k\right]=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$. Thus

$$\begin{aligned}\mathbf{E}\left[X^2\right] &= \frac{\Gamma(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{(\alpha+1)\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.\end{aligned}$$

We now find

$$
\begin{aligned}
\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}^2[X] \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha(\alpha+1)(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)} - \frac{\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha(\alpha+1)(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)} - \frac{\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}
$$

Next section
Working with Transformations of Random Variables
Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# Variance of the Uniform Distribution

Let $f_X(x) = \frac{1}{b-a}\mathbf{1}_{a \le x \le b}$. We already identified the mean of this distribution. To find its variance we begin with $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}^2[X]$.

$$\mathbf{E}\left[X^2\right] = \int_{\Omega} x^2 d\mathbf{P} = \int_{\Omega} x^2 \frac{1}{b-a}\mathbf{1}_{a \le x \le b} = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)}$$

Noting that $b^3 - a^3 = (b-a)(a^2 + ab + b^2)$, then

$$\mathbf{E}\left[X^2\right] = \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} = \frac{(a^2+ab+b^2)}{3}.$$

We can now write

$$\begin{aligned}\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}^2[X] &= \frac{(a^2+ab+b^2)}{3} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{4a^2+4ab+4b^2-3(a^2+b^2+2ab)}{12} \\ &= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}\end{aligned}$$

Uniform and Beta Measure

# Functions of Two Uniform Random Variables

The following examples are excellent examples of the need to pay attention to pay to the regions of interest, as well as in some simple multiple integration.

## The sum of two i.i.d. U(0,1)

Let *X* and *Y* be i.i.d. U(0.1) random variables. Our goal is to find the distribution of $Z = X + Y$. We will use the distribution approach to first identify the cumulative distribution function of *Z*, $F_Z(z)$ and then, if a derivative exists, differentiate to obtain the density.

Since *X* and *Y* are independent, we can write their joint probability density function as the product of the individual ones, or $f_{X,Y}(x,y) = 1_{0<x<1}1_{0<y<1}$.

Since we are examining a function of two random variables, we will need to conduct not just one integral but two. This is called a double integral. We will manage this by essentially converting the double integral into an iterated integral, evaluating this system by first integrating over one of the variables, keeping the other constant, then finally over the second integral, keeping the first constant. The order of the integrals is up to us, and we chose one to simplify the evaluations. This is called Fubini's Theorem. A basic example of this conversion from a double integral to an iterated integral would be to show that the integral over the uniform square (the square with vertices (0, 0), (1,0), (1,1), and (1,0)) is one. We write

$$\mathbf{P}\left[(X,Y) \subset \text{uniform square}\right] = \iint\limits_{(X,Y)\subset\text{uniform square}} f_{X,Y}(x,y)$$

$$= \iint\limits_{(X,Y)\subset\text{uniform square}} 1_{0<x<1}1_{0<y<1} = \int_0^1 dy \int_0^1 dx = 1.$$

Note that we converted the iterated integral over the joint region into a double integral, one over the region of *X*, the second over *Y*.

Before we proceed with the computation of the cumulative distribution function of the sum, do any formal calculations, or sketch out any integrals, we must first understand the region that shapes the event. We will find that this process is critical to performing the correct computations.

In this particular circumstance, we first realize that while each of $X$ and $Y$ are bounded by one, the sum is bounded by two. The distribution approach to identifying the density function of the sum requires us to find $\mathbf{P}[X+Y<Z]$, so this region's configuration is critical to the solution of the problem. However, we see the shape of this region depends on the value of Z (Figure 1).

For any $Z \leq 1$, the region is simply a triangle, and each portion of this region has positive probability (Figure 1). Managing this region should pose no difficulty.

However, the situation is more complicated for $1<Z<2$. Here, only a portion of the region where $X+Y<Z$ has positive probability and this area of integration poses a challenge. However, it does appear that there is a triangle region that reflects the event $X+Y<Z$. It may be best to compute the probability of this region, then compute its complement.

In any event, at least at first blush, it appears that the geometry of these regions are quite different. It may be prudent for us to proceed under the assumption that these two different circumstances should be treated as different cases. We will mark Case 1 as $Z \leq 1$. and Case 2 as $1<Z \leq 2$.

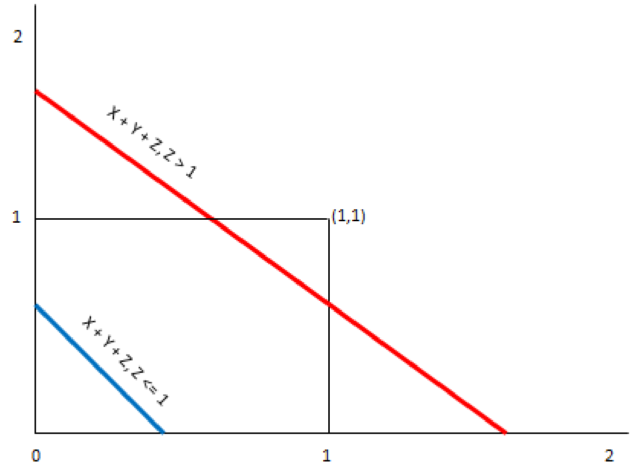


**Figure 1.** The approach to identifying the sum of two random variables whose positive measure is on the [0,1] interval.

Case 1: $Z \leq 1$.
We can examine this circumstance in more detail (Figure 2).

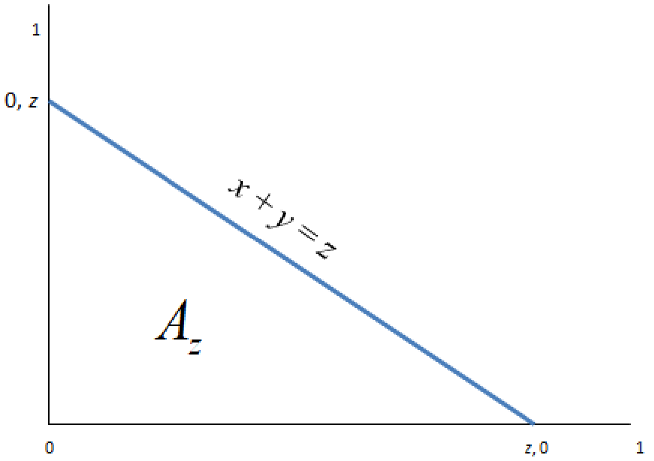


**Figure 2**. The sum of two U(0, 1) random variables. Case 1. $Z \leq 1$

The computation for this case is as straightforward as we envision. Keeping in mind that *z* is a constant, we write

$$\mathbf{P}\left[X+Y\le Z\right]=\iint\limits_{X+Y\le Z} f_{X,Y}\left(x,y\right)=\iint\limits_{X+Y\le Z} 1_{0<x<1}1_{0<y<1}=\int_0^z dy\int_0^{z-y} dx$$

$$=\int_0^z\left(z-y\right)dy=\left[zy-\frac{y^2}{2}\right]_0^z=z^2-\frac{z^2}{2}=\frac{z^2}{2}$$

So $F_Z(z)=\frac{z^2}{2}1_{0<z\le 1}$. Note that $F_Z\left(0\right)=0$ which we expect. We also observe that $F_Z\left(1\right)=\frac{1}{2}$.

Case 2: $1<Z\le 2$.
As we discovered earlier, this region bears closer inspection (Figure 3).

The direct computation of $\mathbf{P}\left[X+Y\le Z\right]$ is complicated by the shape of this region. We will instead compute $\mathbf{P}\left[X+Y>Z\right]$, and then add the additional step $\mathbf{P}\left[X+Y\le Z\right]$ = $1-\mathbf{P}\left[X+Y>Z\right]$. We begin with

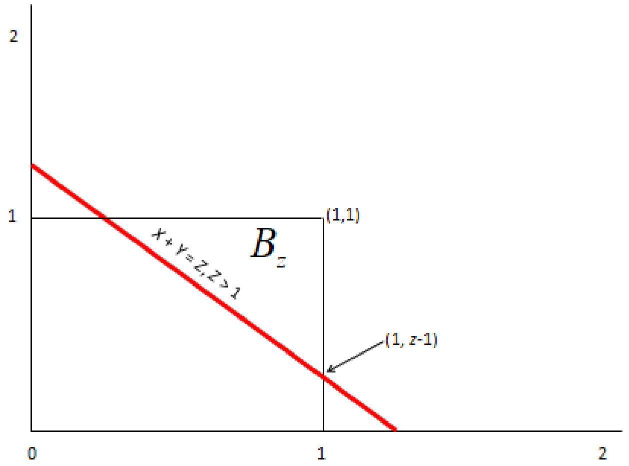


**Figure 3**. The sum of two U( 0,1). Case 2. Z > 1

$$\mathbf{P}\left[X+Y>Z\right]=\iint\limits_{X+Y>Z} f_{X,Y}\left(x,y\right)=\iint\limits_{X+Y>Z} 1_{0<x<1}1_{0<y<1}=\int_{z-1}^{1} dy\int_{z-y}^{1} dx$$

$$=\int_{z-1}^{1}\left(1-z+y\right)dy=\left[\frac{y^2}{2}-(z-1)y\right]_{z-1}^{1}$$

Continuing,

$$\mathbf{P}\left[X+Y>Z\right]=\frac{1}{2}-(z-1)-\left[\frac{(z-1)^2}{2}-(z-1)^2\right]$$
$$=\frac{1}{2}+\frac{(z-1)^2}{2}-(z-1).$$

So, our presumptive finding for the cumulative distribution function when $Z>1$ is $F_Z(z)=\left[\frac{1}{2}+(z-1)-\frac{(z-1)^2}{2}\right]1_{1<Z\le 2}$ . Note that $F_Z(z)=1$ for $z=2$, a result that we would expect. In addition, $F_Z(1)=\frac{1}{2}$ which is what we also found for Case 1. Thus, we write

$$F_Z(z)=\frac{z^2}{2}1_{0<z\le 1}+\left[\frac{1}{2}+(z-1)-\frac{(z-1)^2}{2}\right]1_{1<Z\le 2}$$

We now take a derivative to find the probability density $f_Z(z)$. For $0\le z\le 1$, we find $f_Z(z)=z1_{0\le z\le 1.}$ For the larger values of $z$, we compute $(2-z)1_{1<z\le 2}$. We can now write

$$f_Z(z)=z1_{0\le z\le 1.}+(2-z)1_{1<z\le 2}.$$

So the sum of two i.i.d. U(0,1) random variables is not uniform. Some of the implications of this finding are discussed in the uniform distribution section.

### *The difference of two U(0,1) random variables*

Having now established a motif for managing this style of U(0,1) variable manipulation, we can compute the distribution of $Z=Y-X$, where $X$ and $Y$ are each independent and i.i.d. U(0,1).

The range is different, with $Z$ now going from -1 to 1. We first consult a figure to explore the regions of interest. (Figure 4)

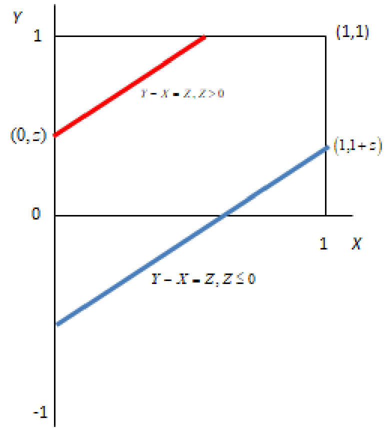


**Figure 4.** The approach to identifying the difference of two U(0,1) random variables, $Y$ - $X$

While the region of interest is the same, the integration that we must carry out is different. For $-1 < z \le 0,$ the probability of the region where $X + Y \le Z$ is straightforward to evaluate, and we write

$$\mathbf{P}\left[X + Y \le Z\right] = \iint\limits_{X+Y\le Z} f_{X,Y}\left(x, y\right) = \iint\limits_{X+Y\le Z} 1_{0<x<1}1_{0<y<1} = \int\limits_{0}^{1+z} dy \int\limits_{y-z}^{1} dx$$

$$= \int\limits_{0}^{1+z} \left(1 - y + z\right) dy = \left[(1+z)y - \frac{y^2}{2}\right]_0^{1+z} = \frac{(1+z)^2}{2}.$$

Note that when $z = -1, F_Z(z) = 0,$ a finding that tracks with our intuition. Also, when $z = 0, F_Z(z) = \frac{1}{2}.$

To compute $F_Z(z)$ for positive $z$, an evaluation of the area suggest that we might first compute $1 - F_Z(z).$

$$1 - F_Z(z) = \iint\limits_{X+Y>Z} f_{X,Y}\left(x, y\right) = \iint\limits_{X+Y>Z} 1_{0<x<1}1_{0<y<1} = \int\limits_{z}^{1} dy \int\limits_{0}^{y-z} dx$$

$$= \int\limits_{z}^{1} \left(y - z\right) dy = \left[\frac{y^2}{2} - zy\right]_z^1 = \left(\frac{1}{2} - z\right) - \left(\frac{z^2}{2} - z^2\right)$$

$$= \frac{1}{2} + \frac{z^2}{2} - z,$$

And $F_Z(z) = z - \frac{z^2}{2} + \frac{1}{2}.$Note that $F_Z\left(0\right) = \frac{1}{2},$ which matches with the finding for the cumulative distribution function for negative $z$, and also $F_Z\left(1\right) = 1,$ which again confirms our intuition. Thus we may write

$F_Z\left(z\right) = \frac{(1+z)^2}{2} 1_{-1\le z\le 0} + \left(z - \frac{z^2}{2} + \frac{1}{2}\right) 1_{0<z\le 1}.$ We now compute

$$f_Z\left(z\right) = \frac{dF_Z\left(z\right)}{dz} = \left(\mathrm{z} + 1\right) 1_{-1\le z\le 0} + \left(1 - z\right) 1_{0<z\le 1}$$

## Product of two U(0,1) variables

This circumstance may at first blush appear to be difficult, but it is among the simplest of all the cases that we will consider here. Again, we let $X$ and $Y$ each be U(0,1). We want the probability function of $Z = XY$. We know that $0 \le z \le 1,$ and a simple diagram convinces us that there is only one case to consider (Figure 5).

In this circumstance regardless of the value of $Z$, we are best served by computing the complement of the cumulative distribution function. Thus, we find

$$1-F_Z(z)=\iint\limits_{XY>Z} f_{X,Y}(x,y)=\iint\limits_{XY>Z} 1_{0<x<1}1_{0<y<1}=\int_z^1 dy\int_{\frac{z}{y}}^1 dx$$

$$=\int_z^1\left(1-\frac{z}{y}\right)dy=\left[y-z\ln(y)\right]_z^1=\left(1-z\ln(1)\right)-\left(z-z\ln(z)\right)$$

$$=1-z+z\ln(z),$$

Or $F_Z(z)=z-z\ln(z)$.

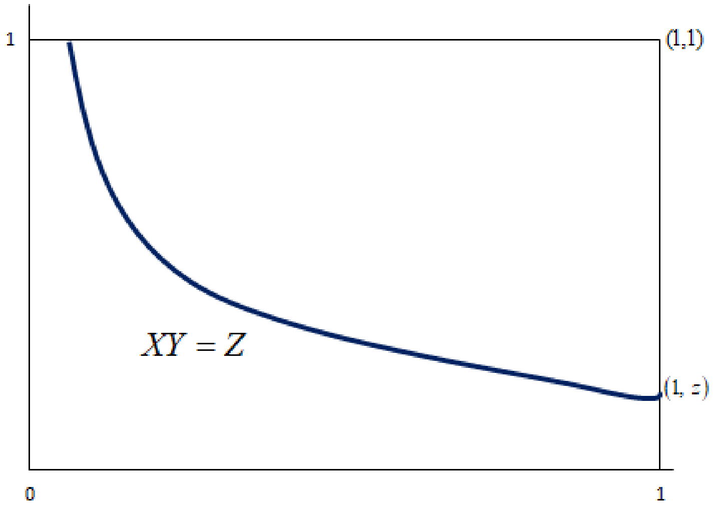


**Figure 5.** The region of integration to compute the product of two U(0,1) random variables.

Does $F_Z(z)=z-z\ln(z)$ make sense? We see that $F_Z(0)=0,$ and $F_Z(1)=1.$ We proceed to find

$$f_Z(z)=\frac{dF_Z(z)}{dz}=1-\left(1+\ln(z)\right)=-\ln(z).$$

## Quotient of two U(0,1) random variables

Let $X$ and $Y$ be i.i.d., U(0,1) random variables. We seek the probability density function of their quotient $Z=\frac{Y}{X}$. We can see the regions of integration of interest by writing this relationship as $Y=ZX$ (Figure 6).

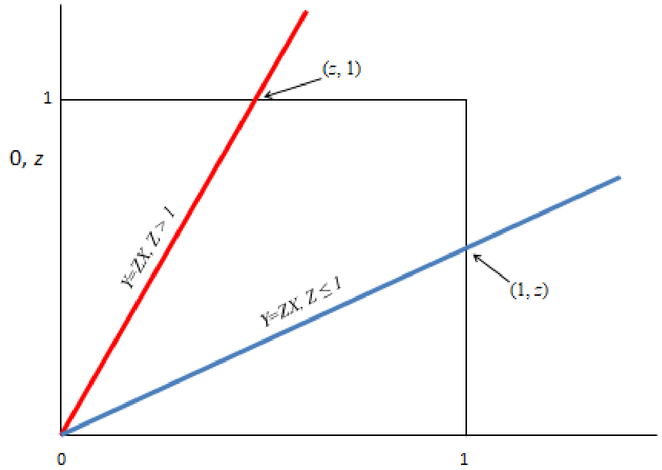


**Figure 6**. Regions of interest in computing the quotient of two U(0,1) random variables

Here, as in many of our other cases, we see that there are two cases. We will solve the easiest case first for $z \le 1$, by simply computing

$$F_Z(z) = \mathbf{P}\left[\frac{Y}{X} \le Z\right] = \iint\limits_{\frac{Y}{X} \le Z} f_{X,Y}(x,y) = \iint\limits_{\frac{Y}{X} \le Z} 1_{0<x<1} 1_{0<y<1} = \int_0^z dy \int_{\frac{y}{z}}^1 dx$$

$$= \int_0^z \left(1 - \frac{y}{z}\right) dy = \left[y - \frac{y^2}{2z}\right]_0^z = \frac{z}{2}.$$

We see that for this case $F_Z(0) = 0$ which makes sense, and $F_Z(1) = \frac{1}{2}$.

For the case of $Z > 1$ we compute

$$1 - F_Z(z) = \mathbf{P}\left[\frac{Y}{X} > Z\right] = \iint\limits_{\frac{Y}{X} > Z} f_{X,Y}(x,y) = \iint\limits_{\frac{Y}{X} > Z} 1_{0<x<1} 1_{0<y<1} = \int_0^1 dy \int_0^{\frac{y}{z}} dx$$

$$= \int_0^1 \frac{y}{z} dy = \left[\frac{y^2}{2z}\right]_0^1 = \frac{1}{2z}$$

So $F_Z(z) = 1 - \frac{1}{2z}$. In this case $F_Z(1) = \frac{1}{2}$, a finding that matches with our earlier case. And since Z has no upper bound, $\lim_{z \to \infty} F_Z(z) = 1$. So we conclude $F_Z(z) = \frac{z}{2} 1_{z \le 1} + \frac{1}{2z} 1_{z>1}$, and we are ready to find the probability density function $f_Z(z)$. Compute

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{1}{2} 1_{z \le 1} + \frac{1}{2z^2} 1_{z>1}.$$

Uniform and Beta Measure

# The Beta Function

The beta function will be most useful to us when we introduce the beta probability distribution, and also when we derive the t and F distributions. Its appearance is daunting at first.

## Definition of the beta function

The beta function is

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

And the probability distribution in which we are ultimately interested is

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} 1_{0 \le x \le 1}$$

We see the difference between these two function is the collection of gamma functions. To gain some insight into this, let's begin with two gamma functions,

$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx,$ and $\Gamma(\beta) = \int_0^\infty y^{\beta-1}e^{-y}dy.$ We write

$$\Gamma(\alpha)\Gamma(\beta) = \int_0^\infty x^{\alpha-1}e^{-x}dx \int_0^\infty y^{\beta-1}e^{-y}dy = \int_0^\infty\int_0^\infty x^{\alpha-1}y^{\beta-1}e^{-(x+y)}dxdy$$

using Fubini's theorem to write an iterated integral as a double integral.

We then carry out the transformation $x = vw; y = v(1-w)$. This is a two variable to two variable transformation. Following the rules we discussed in the uniform distribution discussion of transformation of variables. We write the transformation as

$$g_{V,W}\left(v,w\right)=g_{X,Y}\left(v,w\right)\mathbf{J}\left(x,y\rightarrow v,w\right)\left(1_{x,y}\rightarrow 1_{v,w}\right)$$

To manage the region of integration. We first write $v=x+y$ and $w=\dfrac{x}{x+y}$. Thus, on the set where $0\le x\le\infty$ and $0\le y\le\infty$, then $0\le v\le\infty$, and $0\le w\le 1$.

Proceeding, we note

$$\mathbf{J}\left(x,y\rightarrow v,w\right)=\begin{vmatrix}\dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v}\\ \dfrac{\partial x}{\partial w} & \dfrac{\partial y}{\partial w}\end{vmatrix}=\begin{vmatrix}w & 1-w\\ v & -v\end{vmatrix}=v$$

Thus

$$\begin{aligned}g_{V,W}\left(v,w\right)&=g_{X,Y}\left(v,w\right)\mathbf{J}\left(x,y\rightarrow v,w\right)\left(1_{x,y}\rightarrow 1_{v,w}\right)\\&=\left(wv\right)^{\alpha-1}\left(v\left(1-w\right)\right)^{\beta-1}e^{-v}v1_{0\le v\le\infty}1_{0\le w\le 1}\\&=w^{\alpha-1}\left(1-w\right)^{\beta-1}v^{\alpha+\beta-1}e^{-v}1_{0\le v\le\infty}1_{0\le w\le 1}\end{aligned}$$

$$\begin{aligned}g_{W}\left(w\right)&=\int_{\Omega_V}g_{V,W}\left(v,w\right)dw\\&=\int_{\Omega_V}w^{\alpha-1}\left(1-w\right)^{\beta-1}v^{\alpha+\beta-1}e^{-v}1_{0\le w\le 1}dv\\&=w^{\alpha-1}\left(1-w\right)^{\beta-1}1_{0\le w\le 1}\int_{\Omega_V}v^{\alpha+\beta-1}e^{-v}dv.\end{aligned}$$

Now, we know that $\int_{\Omega_V}v^{\alpha+\beta-1}e^{-v}dv=\Gamma\left(\alpha+\beta\right)$. Thus we have

$$\Gamma(\alpha)\Gamma(\beta)=\int_0^1 w^{\alpha-1}\left(1-w\right)^{\beta-1}\int_{\Omega_V}v^{\alpha+\beta-1}e^{-v}dv=\int_0^1 w^{\alpha-1}\left(1-w\right)^{\beta-1}\Gamma\left(\alpha+\beta\right)$$

Thus

$$\Gamma(\alpha)\Gamma(\beta)=\int_0^1 w^{\alpha-1}\left(1-w\right)^{\beta-1}dw\ =\Gamma\left(\alpha+\beta\right)$$

and

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma\left(\alpha+\beta\right)}=\int_0^1 w^{\alpha-1}\left(1-w\right)^{\beta-1}dw$$

Therefore

$$\int_0^1\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma(\alpha)\Gamma(\beta)}w^{\alpha-1}\left(1-w\right)^{\beta-1}dw=1$$

Uniform and Beta Measure

# Moments of the Beta Distibution

We will find the $k^{\text{th}}$ moment of the Let's begin with the first moment of a random variable *X* that follows a Beta (α, β) distribution.

$$
\begin{aligned}
\mathbf{E}\left[X^k\right] &= \int_{\Omega_X} x^k d\mathbf{P} = \int_0^1 x^k \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 \frac{\Gamma(\alpha+\beta+k)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+\beta+k)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}
\end{aligned}
$$

Uniform and Beta Measure

# Transformations of Variables

## Introductory remarks

We have seen that measure theory is at its root, simply measuring or weighing a collection of numbers (be they discrete, an interval, or some other combination) using a measuring tool, which in probability we call a probability mass or probability density function. The source of some of these measuring tools may be by direct computation from first principles, e.g., the Poisson distribution. Some are from a computation using moment generating functions, or from taking direct measures e.g., the sum of two uniform random variables. However, another useful tool is the development of a transformation.

Prerequisites
Lebesgue Integration Theory and the Bernoulli Distribution

Uniform and Beta Measure
Continuous Probability Measure

Transformations allow us to move smoothly from the measuring tool of one random variable to the measuring tool for another, simply by knowing the relationship between two random variables. We will start with some very simple examples to demonstrate how this tool works.

Let's begin with a random variable $X$ that follows a uniform distribution on $[0,1]$. Let's suppose that we have the random variable $Y=-X$. What is measuring tool for $Y$?

How are these random variables different? Well to start, our intuition tells us that their ranges of possible values are not the same, since $Y$ takes positive measure only on the $[-1,0]$ range. However, we can use the random variable $X$ to help us find the probability that $Y$ lies in a given interval. For example, it follows that $\mathbf{P}\left[-\frac{3}{4}\le y\le-\frac{1}{2}\right]=\mathbf{P}\left[\frac{1}{2}\le x\le\frac{3}{4}\right]=\frac{1}{4}$. Once we made the range change, finding the probability was straightforward. We can write this in terms of measuring tools $f_X(x)=1_{0\le x\le1}$, and $f_Y(y)=1_{-1\le y\le0}$. This is the solution.

For another example, let $W=2X$. Being sensitized to the role of the range in a transformation of random variables. We see that $W$ is a transformation that maps $[0,1]$ to $[0,2]$. This transformation actually does some "stretching", and to find probabilities for $W$ we have to "compress" For example, $\mathbf{P}[1 \le w \le 2] = \mathbf{P}\left[\frac{1}{2} \le x \le 1\right] = \frac{1}{2}$. We must take this compression factor of two into account.

Essentially, we find probabilities for *w* by using the measuring tool for *x* and then doing this compression, plus consideration of the range of *w*. This compression can be generally expressed as taking the derivative of *w* with respect to *x*. In this case $x = \frac{w}{2}, dx = \frac{dw}{2}$. This is the additional ingredient we needed for the transformation, and we can write $f_W(w) = \frac{1}{2}\mathbf{1}_{0 \le w \le 2}$.

Now consider the example of a random *R* that assigns measure to the $[0,1]$, in accordance with the tool $f_R(r) = 2r\mathbf{1}_{0 \le r \le 1}$. Now, let's define $S = 10R$. What is $\mathbf{P}[1 \le r \le 7]$? How do we find the measuring tool for *S*?

We know the new range is $[0,10]$. We also know that $r = \frac{s}{10}$, so $dr = \frac{ds}{10}$. This incorporates the compression as we get back to the space of positive measure for *r*. However, we must assess the measuring tool for *r* in terms of *s*. In this case this means noting $2r = 2\left(\frac{s}{10}\right) = \frac{s}{5}$.

Thus $f_S(s) = 2\left(\frac{s}{10}\right)\left(\frac{ds}{10}\right)c = \frac{s}{50}ds\mathbf{1}_{[0 \le s \le 10]}$.

## Process

Thus the entire process requires three steps

1. Change the range
2. Find the scale factor
3. Replace the original variable in the measuring tool with the new variable.

We can write this symbolically as

$$f_S(s) = f_R(s)[dr \to ds][\Omega_R \to \Omega_S].$$

Applying this formulation directly to our previous problem, we have $f_S(s) = 2\frac{s}{10}\left(\frac{1}{10}\right)\mathbf{1}_{0 \le s \le 10} = \frac{s}{50}\mathbf{1}_{0 \le s \le 10}$, and we find the measure of the interval $[1,7]$ directly as

$$\mathbf{E}\left[\mathbf{1}_{[1,7]}\right] = \int_1^7 d\mathbf{P} = \int_1^7 \frac{s}{50}\mathbf{1}_{0 \le s \le 10} ds = \int_1^7 \frac{s}{50} ds = \left[\frac{s^2}{100}\right]_1^7$$
$$= \frac{49}{100} - \frac{1}{100} = \frac{12}{25}.$$

Let's try another example. Let *X* be a random variable with measuring tool $f_X(x) = x^3\mathbf{1}_{0 \le x \le 1.}$ Define a new random variable $Y = 2X^3 + 7$. We need the measuring tool for *Y*. Note that for $0 \le X \le 1, \quad 7 \le Y \le 9$. We find $X = \left(\frac{Y-7}{2}\right)^{\frac{1}{3}}, dx = \frac{1}{6}\left(\frac{Y-7}{2}\right)^{-\frac{2}{3}}$ and we write

$$
\begin{aligned}
f_Y(y) &= f_X(s)\left[dx \to dy\right]\left[\Omega_X \to \Omega_Y\right] \\
&= \left(\left(\frac{Y-7}{2}\right)^{\frac{1}{3}}\right)^3 \frac{1}{6}\left(\frac{Y-7}{2}\right)^{-\frac{2}{3}} 1_{7\le y\le 9} \\
&= \frac{1}{6}\left(\frac{Y-7}{2}\right)^{\frac{1}{3}} 1_{7\le y\le 9}.
\end{aligned}
$$

These transformations which are bijective (i.e., one to one) are straightforward. However consider the random variable *X* with probability density function $f_X(x) = \frac{1}{2}1_{-1\le x\le 1,}$ that is, measure is spread uniformly across $[-1,1]$. Now define $Y = |X| = X1_{0\le X\le 1} - X1_{-1\le X\le 0}$. This function maps negative values of *X* to their absolute value, and positive values to their same positive value as well, making this a two to one mapping. We express this by saying if $\Omega_X = 1_{-1\le x\le 1}$ and $\Omega_Y = 1_{0\le x\le 1}$, then $\Omega_X \to \Omega_Y = 21_{0\le x\le 1}$ Our intuition tells us that intervals of equal length have the same probability, as is the case for *X*, should hold the same under *Y*. The factor from the derivative is one*

If we were to apply $f_Y(y) = f_X(y)\left[dx \to dy\right]\left[\Omega_X \to \Omega_Y\right]$ we would compute.

$$
\begin{aligned}
f_Y(y) &= f_X(y)\left[dx \to dy\right]\left[\Omega_X \to \Omega_Y\right] \\
&= \left(\frac{1}{2}\right)(1)\left(21_{0\le x\le 1}\right) = 1_{0\le x\le 1}.
\end{aligned}
$$

Similarly, let $W = X^2$. Then for $-1 \le X \le 1,\ 0 \le W \le 1$, representing a two to one mapping. We keep in mind the factor of 2, noting

$$\left[\Omega_X \to \Omega_W\right] = 21_{0\le w\le 1}.$$

Then $x = w^{\frac{1}{2}}, dx = \frac{1}{2}w^{-\frac{1}{2}}$, and we can conclude

$$
\begin{aligned}
f_W(w) &= f_X(w)\left[dx \to dw\right]\left[\Omega_X \to \Omega_W\right] \\
&= \left(\frac{1}{2}\right)\left(\frac{1}{2}w^{-\frac{1}{2}}\right)\left(21_{0\le w\le 1}\right) \\
&= \frac{1}{2}w^{-\frac{1}{2}}1_{0\le w\le 1}
\end{aligned}
$$

[*] This is true for the region of interest except for the value x=0, but we are not troubled by this since the probability of any one point is zero for this measuring tool.

The measure of the $[0,1]$ reals with this tool is one as it should be if *W* is a proper random variable.

Finally, to consider an alternative mapping, let *X* be a random variable with measuring tool $f_X(x) = 1_{0 \le x \le 1}$, and define the new random variable $Y = \sqrt{X}$. Here each value of *x* is mapped to two different values of *y*. This is one to two mapping, and we reflect that as $[\Omega_X \rightarrow \Omega_Y] = \frac{1}{2} 1_{-1 \le y \le 1}$. to correct for the expansion . To continue, we find that $X = Y^2, dx = 2y\,dy,$ and we compute

$$\begin{aligned} f_Y(y) &= f_X(s)[dx \rightarrow dy][\Omega_X \rightarrow \Omega_Y] \\ &= (2y)\left(\frac{1}{2} 1_{-1 \le y \le 1}\right) = y 1_{-1 \le y \le 1}. \end{aligned}$$

## Generalization to higher dimensions

Commonly we will have circumstances where we are not converting just one variable to another, but two variables to two or three to three. In the cases of showing that the sample mean and variance of a normal distribution are independent, we will need to transform *n* variables to *n* variables.

In these circumstances, we are guided by the same principles that served us well for the one to one transformations. However, the derivative will be the determinant of a matrix of derivatives. In most cases, this will be easy to find.

If we are creating two random variables *V* and *W* from *X* and *Y*, then we write

$$f_{V,W}(v,w) = f_{X,Y}(v,w) J\left[(x,y) \rightarrow (v,w)\right]\left[\Omega(x,y) \rightarrow \Omega(v,w)\right].$$

Where $J\left[(x,y) \rightarrow (v,w)\right]$ is the notation that represents the Jacobian that governs the transformation of $(X,Y) \rightarrow (V,W)$ mapping

We will get much practice with this tool later.

## Convolutions

Determination of the measure of the sums of random variables plays an important role in probability (e.g., the central limit theorem). We have several ways in which to manage this important function.

One of them is the identify directly the measure based on geometry. A second is by using moment generating functions. A third is through the use of the process of random variable transformation.

A fourth tool is through the use of convolutions. A convolution builds the desired probability up from the relationship between the two random variables.

### *Finding the measure of a sum using convolutions*

As an illustration, recall that we have demonstrated that the sum of two random variables that follow binomial measure with the same parameter *p* is also binomial with the same *p* parameter. However, suppose that this simplifying assumption of a common value of $p$ is not the case.

Let's first examine the problem from its simplest form. Assume that the random variable $X_1$ follows binomial measure with parameters $(2, p_1)$ and $X_2$ follows binomial measure with parameters $(2, p_2)$ independent of $X_1$. What is the measure of the random variable $W = X_1 + X_2$?

The possible values of $W$ are 0,1,2,3,or 4. Since this is a relatively small number of possibilities, we can count how each of these may be achieved and from there compute the measure.

For example, $W$ can only be zero if both $X_1$ and $X_2$ are each zero. We write $\mathbf{P}[W=0]=(1-p_1)^2(1-p_2)^2$.

Next, $W = 1$ if either $X_1 = 0, X_2 = 1,$ or $X_1 = 1, X_2 = 0.$ Thus

$$\mathbf{P}[W=1]=(1-p_1)^2\binom{2}{1}p_2(1-p_2)$$
$$+\binom{2}{1}p_1(1-p_1)(1-p_2)^2.$$

We find $\mathbf{P}[W=2]$ analogously. If we let the notation $\langle k, j\rangle$ be the joint event that $X_1 = k$ and $X_2 = j,$ then we write $\mathbf{P}[W=2]=\langle 0,2\rangle+\langle 1,1\rangle+\langle 2,0\rangle.$ Analogously, we find

$$\mathbf{P}[W=3]=\langle 0,3\rangle+\langle 1,2\rangle+\langle 2,1\rangle+\langle 3,0\rangle$$
$$\mathbf{P}[W=4]=\langle 0,4\rangle+\langle 1,3\rangle+\langle 2,2\rangle+\langle 3,1\rangle+\langle 4,0\rangle.$$

Note that the joint events all have the common feature that $k + j = W.$ This is a relationship of which we must take advantage. Begin by writing

$$\mathbf{P}[W=4]=\sum_{k=0}^{4}B(n_1,p_1,k),B(n_2,p_2,4-k).$$

Identifying the relationship between the entries in the bracket is the heart of the convolution. Relying on this development, we conclude that

$$\mathbf{P}[W=m]=\sum_{k=0}^{m}B(n_1,p_1,k),B(n_2,p_2,m-k).$$

We can proceed analogously for the sum of geometric random variables. Let's assume that we have two independent random variables following geometric measure;

$$\mathbf{P}[X_1=k]=q_1p_1^{k-1};\mathbf{P}[X_1=k]=q_2p_2^{k-1}.$$

Then,

$$\mathbf{P}\left[W=m\right]=\sum_{k=0}^{m}\mathbf{P}\left[X_1=k\right]\mathbf{P}\left[X_2=m-k\right]=\sum_{k=0}^{m}q_1p_1^{k-1}q_2p_2^{m-k-1}$$

$$q_1q_2p_2^{m}\sum_{k=0}^{m}\left(\frac{p_1}{p_2}\right)^{k-1} \;=\; q_1q_2p_2^{m}\frac{1-\left(\frac{p_1}{p_2}\right)^{k}}{1-\left(\frac{p_1}{p_2}\right)} \;=q_1q_2p_2^{m+1}\frac{1-\left(\frac{p_1}{p_2}\right)^{k}}{p_2-p_1}$$

### *Convolutions for continuous functions*

The procedure is analogous to that for dichotomous random variables, with the exception that the relevant measure is

$$\mathbf{P}\left[W=X_1+X_2\le z\right]=\iint\limits_{x_1+x_2\le z} f_{X_1,X_2}\left(x_1,x_2\right)d\mathbf{P}$$

$$=\int_{-\infty}^{\infty} f_{X_2}\left(x_2\right)\int_{-\infty}^{z-x_2} f_{X_1}\left(x_1\right)=\int_{-\infty}^{\infty} f_{X_2}\left(x_2\right)\mathbf{F}_{X_1}\left(z-x_2\right).$$

Recommended sections

Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# Survival Measure: Exponential, Gamma, and Related Measures

## Introduction

There are a collection of distributions in biostatistics that receive substantial attention because they have been linked to existence durations, or the time that a system stays in any one state.

The length of time a hospital goes with no patients in its ICU beds, the length of time a stem cell retains its pluripotency, the minutes an ambulance goes without a call, the length of time an individual lives before death each characterize a system without a state change. There are many distributions that characterize this process. Each of them is relatively simple to employ, and harken back to our work in the Poisson process.

Prerequisites
An Introduction to the Concept of Measure
Lebesgue Integration Theory and the Bernoulli Distribution
Conditional Probability
General Poisson Process
Continuous Probability Measure
Variable Transformations

## Development of the thought process

We will begin by returning to the Poisson process. Assume that subjects arrive at a laboratory for a blood draw in accordance with a Poisson process with parameter $\lambda$=5 subjects per hour. Then we know that the arrivals are independent of each other, and can compute the probability that $k$ arrivals occur in time $t$, $\mathbf{P}\left[X(t)=k\right]=\frac{\left(\lambda t\right)^k}{k!}e^{-\lambda t}$. For example, the probability that at least twelve patients arrive in three hours is

$$\mathbf{P}\left[K\geq 12\right]=\sum_{k=12}^{\infty}\frac{15^k}{k!}e^{-15}=0.815.$$

Now, what is the probability that no subjects arrive?

Consider what this means. We are assuming that there was a last arrival some time ago, and we are watching the system, awaiting the next one. We are doing nothing more than counting time.

Now, our intuition also tells us that, given we know the arrival rate, the probability of the next event in a block of time is a function of the size of the time block. For example, the fact that $\lambda$ = 5 arrivals per hour means that we would not really expect an arrival one minute after the last one. However as time passes the probability of at least one arrival in the time interval $t$ should grow. In fact, if we can wait long enough, the likelihood of at least one arrival would be one.

Note how the discussion has turned from one about the number of arrivals for which we have a measure to one about time. The experiment generating this random variable began with a Poisson arrival process. However, now attention has turned to time – time is now the random variable. The experiment is no longer counting the number of arrivals. It is instead seeing how long we must wait until the next arrival.

However, it is the Poisson process that helps us find the probability distribution of this inter-arrival time. The probability that we have to wait time $t$ until the next arrival is $\mathbf{P}\left[T \geq t\right] = \mathbf{P}\left[\, K = 0 \text{ in } (0,t)\,\right] = e^{-\lambda t}$. Continuing with the idea that $T$ is the random variable, then we can also find $\mathbf{P}\left[T \leq t\right] = 1 - \mathbf{P}\left[T \geq t\right] = 1 - e^{-\lambda t} = F_T\left(t\right)$. With this continuous cumulative distribution function in hand, we can write the probability density function for $T$,

$$f_T(t) = \frac{dF_T(t)}{dt} = \lambda e^{-\lambda t} 1_{0 \leq t \leq \infty}.$$

This final formulation is what is known as the negative exponential distribution. Its formulation begins as a consequence of the Poisson process. Its parameter is the same as the parameter from the Poisson distribution. However, since $T$ is the time between arrivals, it sometimes is more useful to focus not on $\lambda$ but instead on $\frac{1}{\lambda}$ or the average time between arrivals.

We can see at once that selecting the range of the random variable $T$ as all nonnegative real numbers is appropriate since $\int_0^\infty \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big]_0^\infty = 1$. Examples reveal the richness of measuring tools from this exponential family

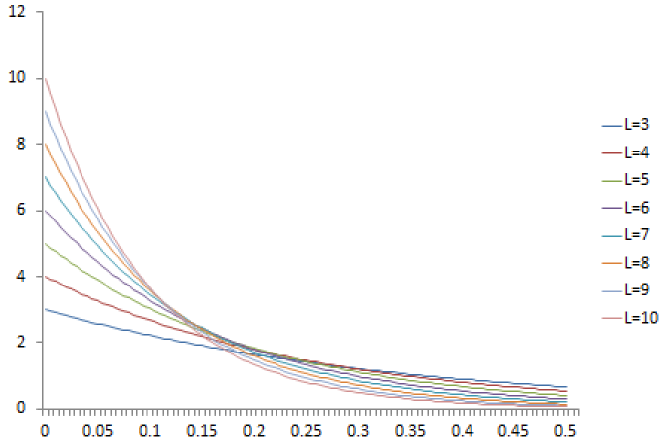


**Figure 1**. Examples from the family of negative exponential distributions with parameter L

## Negative exponential moments

A review of gamma functions reveals the simple result that $\int_0^\infty t\lambda e^{-\lambda t}dt = \frac{1}{\lambda}$, and $\int_0^\infty t^2\lambda e^{-\lambda t}dt = \frac{2}{\lambda^2}$.

Thus $\mathbf{E}[T] = \frac{1}{\lambda}$, and $\mathbf{Var}[T] = \mathbf{E}[T^2] - \mathbf{E}^2[T] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. We can find the moment generating function $\mathbf{M}_X(t)$ directly as

$$\mathbf{M}_X(t) = \mathbf{E}\left[e^{tx}\right] = \int_{\Omega_x} e^{tx}d\mathbf{P} = \int_{\Omega_x} e^{tx}\lambda e^{-\lambda x}1_{0\le x\le\infty} = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx$$

$$= \lambda\int_0^\infty e^{-(\lambda-t)x}dx = \frac{\lambda}{(\lambda-t)}\int_0^\infty(\lambda-t)e^{-(\lambda-t)x}dx = \frac{\lambda}{(\lambda-t)}1_{[\lambda>t]}.$$

## The memoryless property

Consideration of conditional distributions for the exponential random variable provides an unusual finding. Let us suppose that *X* is a random variable following an exp($\lambda$), and we have two specific times, *s* and *t*, such that $0 < s < t < \infty$. What is $\mathbf{P}[X \ge T \mid X \ge s]$?

We know that the unconditional

$$\mathbf{P}[X \ge t] = \mathbf{E}\left[1_{X\ge t}\right] = \int_{\Omega_x} 1_{x\ge t}\,d\mathbf{P} = \int_{\Omega_x} 1_{x\ge t}\lambda e^{-\lambda x}1_{0\le x\le\infty}dx$$

$$= \int_t^\infty \lambda e^{-\lambda x}dx = e^{-\lambda t}.$$

Then

$$\mathbf{P}[X \ge t \mid X \ge s] = \frac{\mathbf{P}[X \ge t \cap X \ge s]}{\mathbf{P}[X \ge s]} = \frac{\mathbf{P}[X \ge t]}{\mathbf{P}[X \ge s]} = \frac{e^{-\lambda t}}{e^{-\lambda s}} = e^{-(t-s)}.$$

If we consider that the time line is comprised of times less than s, times between *s* and *t*, and times greater than *t*, than the $\mathbf{P}[X \ge T \mid X \ge s] = e^{-\lambda(t-s)}$ is wholly related to the time interval between *s* and *t*. Another way to say this is that what happened before time *s* does not count. It is as though the process actually started not at time zero but at time *t* and runs to *t* – *s*.

This feature of a stochastic process (that is, a probability process that involves time), is characterized as memoryless. It also makes the negative exponential distribution one of the easiest distributions to work with because the complications that complex sequences of events that occur early in time can sometimes be ignored.

## Example: Imaging facility

Let's assume that patients arrive to an imaging facility in accordance with a Poisson arrival process, with arrival rate *λ* = 15 patients per hour. Given than 10 subjects have arrived at the cardiac magnetic imaging scanner (cMR) in 60 minutes, what is the likelihood that the next patient will arrive within the next ten minutes?

We use the memoryless property of the negative exponential distribution to see that this is simply the probability that a patient arrives within ten minutes, or $\mathbf{P}[T \le t] = 1 - e^{\lambda t}$. The arrival rate scales to 0.25 patients per minute. We therefore compute

$\mathbf{P}[T \le 10] = 1 - e^{-(0.25)(10)} = 1 - 0.082 = 0.918.$

The timing of the first ten subjects did not matter.

■

## Difference of exponential random variable

A fairly easy result to develop which will permit us to practice with using transformations of random variables is to find the distribution of two independent negative exponential random variables. Let $X_1$ follow a negative exponential distribution with parameter $\lambda_1$ and $X_2$ follow a negative exponential distribution with parameter $\lambda_2$. We seek the distribution of $Y = X_1 - X_2$.

We notice that unlike $X_1$ and $X_2$ that must be on the nonnegative reals, $Y$ can be negative as well as positive. We could attempt to follow the same approach that was taken for identifying the measuring tool for the difference of two uniform random variables, however here we will take a different tack. If, for example we let $W = X_1$ (an admittedly transparent transformation), and we had the joint probability density function for $Y$ and $W$, $f_{Y,W}(y,w)$, we would be able to compute $f_Y(y) = \int_{\Omega_W} f_{Y,W}(y,w)$. We will obtain $f_{Y,W}(y,w)$ from $f_{X_1,X_2}(x_1,x_2)$ using our transformation of variable technique.

We begin by writing

$$f_{Y,W}(y,w) = f_{X_1,X_2}(y,w)\,J\left[(x_1,x_2)\to(y,w)\right]\left[\Omega(x_1,x_2)\to\Omega(y,w)\right].$$

We will manage each of the three operations on the right hand side of this equation. Begin with $Y = X_1 - X_2$ and $W = X_1$ simply means that $X_1 = W$, and $X_2 = W - Y$. We can write

$$J\left[(x_1,x_2)\to(y,w)\right] = \begin{vmatrix} \dfrac{\partial x_1}{\partial w} & \dfrac{\partial x_2}{\partial w} \\ \dfrac{\partial x_1}{\partial y} & \dfrac{\partial x_2}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & -1 \end{vmatrix} = |-1| = 1.$$

The region of positive measure requires close examination. We note as before that $-\infty < Y < \infty$. But also $Y \le W$ and $W \ge 0$. The only way to reconcile these regions is to write this as $-\infty < \max(Y,0) \le W < \infty$. Since ultimately, we will need to integrate out $W$, it looks like it may be wise to consider two cases, the first where $-\infty < y < 0 \le x < \infty$, and the second where $0 \le y \le x < \infty$.

The last component we need to identify the distribution of $Y = X_1 - X_2$ is $f_{X_1,X_2}(x_1,x_2)$ which we write as

$$f_{X_1,X_2}(x_1,x_2) = \lambda_1 e^{-\lambda_1 x_1}\lambda_2 e^{-\lambda_2 x_2} = \lambda_1\lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2}\mathbf{1}_{0\le x_1<\infty,\,0\le x_2<\infty}.$$

Seeing that $-\lambda_1 x_1 - \lambda_2 x_2 = -\lambda_1 w - \lambda_2(w-y) = -(\lambda_1+\lambda_2)w + \lambda_2 y$, we can proceed with

$$f_{Y,W}(y,w) = \lambda_1\lambda_2 e^{-(\lambda_1+\lambda_2)w+\lambda_2 y}\mathbf{1}_{-\infty<\max(Y,0)\,\le\,W<\infty.}$$

Recall there were two cases. For $-\infty < y \le 0$, we write

$$f_Y(y) = \int_{\Omega_W} f_{Y,W}(y,w) = \int_{\Omega_W} \lambda_1\lambda_2 e^{-(\lambda_1+\lambda_2)w+\lambda_2 y} 1_{-\infty<Y\le 0\le W<\infty}$$

$$= \int_0^\infty \lambda_1\lambda_2 e^{-(\lambda_1+\lambda_2)w+\lambda_2 y} 1_{-\infty< Y\le 0.}$$

$$= \lambda_1\lambda_2 e^{\lambda_2 y} 1_{-\infty< Y\le 0.} \int_0^\infty e^{-(\lambda_1+\lambda_2)w} dw$$

$$= \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{\lambda_2 y} 1_{-\infty< Y\le 0}.$$

For $0 \le y < \infty$, we have

$$f_Y(y) = \int_{\Omega_W} f_{Y,W}(y,w) = \int_{\Omega_W} \lambda_1\lambda_2 e^{-(\lambda_1+\lambda_2)w+\lambda_2 y} 1_{0\le Y\le W<\infty.}$$

$$= \lambda_1\lambda_2 e^{\lambda_2 y} 1_{0\le Y\le \infty.} \int_y^\infty e^{-(\lambda_1+\lambda_2)w} dw$$

$$= \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{\lambda_2 y} 1_{0\le Y\le \infty.} \int_y^\infty (\lambda_1+\lambda_2) e^{-(\lambda_1+\lambda_2)w} dw$$

$$= \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{\lambda_2 y} 1_{0\le Y\le \infty.} e^{-(\lambda_1+\lambda_2)y}$$

$$= \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{-\lambda_1 y} 1_{0\le Y\le \infty.}$$

Thus

$$f_Y(y) = \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{\lambda_2 y} 1_{-\infty< Y\le 0} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} e^{-\lambda_1 y} 1_{0\le Y\le \infty}$$

$$= \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\left(e^{\lambda_2 y} 1_{-\infty< Y\le 0} + e^{-\lambda_1 y} 1_{0\le Y\le \infty}\right).$$

Depending on the choice of parameters (and therefore the likely magnitudes of the two random variables whose difference we seek), this distribution has a variety of shapes (Figure 2.)

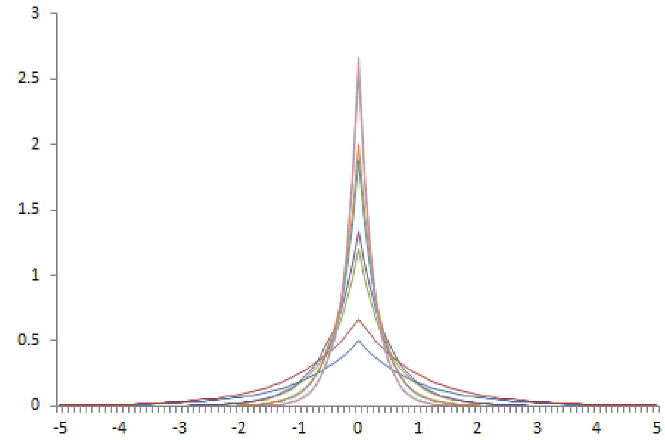


**Figure 2**. Distribution of the difference between two exponential random variables for parameter 1 = 1,2,3,4,and parameter 2 = 1,2, 3,....8

We can find the probability distribution of the linear combination $Y = aX_1 - bX_2$ (*a* and *b* both positive constants) by recognizing that $aX_1 - bX_2 = W - V$ where $W$ follows a negative exponential distribution with parameter $a\lambda_1$ and $V$ follows a negative exponential distribution with parameter $b\lambda_{2.}$ We can apply the previous result to write

$$f_Y(y) = \frac{ab\lambda_1\lambda_2}{a\lambda_1 + b\lambda_2}\left(e^{b\lambda_2}1_{-\infty<y\leq 0} + e^{-a\lambda_1}1_{0\leq y<\infty}\right).$$

## Gamma random variables

Let's revisit the previous example where we considered the time until the first arrival at an imaging center. Having found that this first arrival time is a random variable that follows a negative exponential distribution. We then examined the distribution of the arrival of future events, given a sequence of prior arrival times.

This we saw followed the same negative exponential distribution – what this characteristic we called the memoryless property. This means that once we knew the time of the first arrival, we could find the distribution of the time of the second arrival.

However, suppose we do not know the time of the first arrival. What is the distribution of the time of the second arrival? This would be the sum of two random variables, the first being the time until the initial arrival and the second being the time of the second arrival. We can show formally that the measuring tool or probability density function of this random variable *z* is

$$f_Z(z) = \lambda^2 z e^{-\lambda z}1_{0\leq z<\infty}.$$

A straightforward way to evaluate this is to use the transformation of variables. Let $X_1$ and $X_2$ each follow a negative exponential distribution with parameter $\lambda$. We seek the distribution of $Z = X_1 + X_2$. Allow $W = X_1$ and plan to find $f_{Z,W}(z,w)$, followed by $f_Z(z) = \int_{\Omega_W} f_{Z,W}(z,w)$. We will obtain $f_{Z,W}(z,w)$ from $f_{X_1,X_2}(x_1,x_2)$ using our transformation of variable technique.

We begin by writing

$$f_{Z,W}(z,w) = f_{X_1,X_2}(z,w)J\left[(x_1,x_2)\to(z,w)\right]\left[\Omega(x_1,x_2)\to\Omega(z,w)\right].$$

We will manage each of the three operations on the right hand side of this equation. Beginning with $Z = X_1 + X_2$ and $W = X_1$ simply means that $X_1 = W$, and $X_2 = Z - W$. We can write

$$J\left[\left(x_1, x_2\right) \to (z, w)\right] = \begin{vmatrix} \dfrac{\partial x_1}{\partial w} & \dfrac{\partial x_2}{\partial w} \\ \dfrac{\partial x_1}{\partial z} & \dfrac{\partial x_2}{\partial z} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = |1| = 1.$$

Note that $0 \le w \le z < \infty$.

The last component we need to identify in the distribution of $Z = X_1 + X_2$ is $f_{X_1,X_2}\left(x_1, x_2\right)$ which we write as

$$f_{X_1,X_2}\left(x_1, x_2\right) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} = \lambda^2 e^{-\lambda\left(x_1 + x_2\right)} 1_{0 \le x_1 < \infty, 0 \le x_2 < \infty}.$$

Seeing that $-\lambda\left(x_1 + x_2\right) = -\lambda z$, we can proceed with

$$f_{Z,W}\left(z, w\right) = \lambda^2 e^{-\lambda z} 1_{0 \le w \le z < \infty.}$$

We now simply measure this function on $0 \le w \le z$.

$$\begin{aligned} f_z\left(z\right) &= \int_{\Omega_W} f_{Z,W}\left(z, w\right) = \int_{\Omega_W} \lambda^2 e^{-\lambda z} 1_{0 \le w \le z < \infty.} \\ &= \lambda^2 e^{-\lambda z} 1_{0 \le z < \infty.} \int_0^z dw = \lambda^2 z e^{-\lambda z} 1_{0 \le z < \infty.}. \end{aligned}$$

We note that this is related to a gamma function. We simply let $v = \lambda z$ to see that $f_V(v) = v e^{-v} 1_{0 \le v < \infty}$. and we can write $f_Z(z) = \lambda^2 z e^{-\lambda z} 1_{0 \le z < \infty.} = \dfrac{\lambda^2}{\Gamma(2)} z e^{-\lambda z} 1_{0 \le z < \infty}$.

What about the distribution for the sum of three independent, identically distributed random variables? If we let $Z = X_1 + X_2 + X_3$, and then let $Y = X_1 + X_2$, and $W = X_1$. Following the previous example, we write

$f_Z(z) = \int_{\Omega_{Y,W}} f_{Z,Y,W}(z, y, w)$, and fall back on our transformation formula

$$\begin{aligned} &f_{Z,Y,W}(z, y, w) \\ &= f_{X_1,X_2,X_3}\left(z, y, w\right) J\left[\left(x_1, x_2, x_3\right) \to (z, y, w)\right]\left[\Omega\left(x_1, x_2, x_3\right) \to \Omega(z, y, w)\right]. \end{aligned}$$

We now proceed with the computation. Begin with

$$x_1 = w$$
$$x_2 = y - w$$
$$x_3 = z - y - w$$

$$J\left[\left(x_1,x_2,x_3\right)\to(z,y,w)\right]=\begin{vmatrix}\frac{\partial x_1}{\partial w} & \frac{\partial x_2}{\partial w} & \frac{\partial x_3}{\partial w}\\ \frac{\partial x_1}{\partial y} & \frac{\partial x_2}{\partial y} & \frac{\partial x_3}{\partial y}\\ \frac{\partial x_1}{\partial z} & \frac{\partial x_2}{\partial z} & \frac{\partial x_3}{\partial z}\end{vmatrix}=\begin{vmatrix}1 & -1 & -1\\ 0 & -1 & -1\\ 0 & 0 & 1\end{vmatrix}=\left|-1\right|=1$$

And $f_{X_1,X_2,X_3}\left(x_1,x_2,x_3\right)$
$$=\lambda e^{-\lambda x_1}\lambda e^{-\lambda x_2}\lambda e^{-\lambda x_3}=\lambda^3 e^{-\lambda\left(x_1+x_2+x_3\right)}1_{0\le x_1<\infty,0\le x_2<\infty,0\le x_3<\infty}.$$

Thus $f_{Z,Y,W}(z,y,w)=\lambda^3e^{-\lambda z}1_{0\le w\le y\le z<\infty}$.

The elegance of this approach is the selection of *z*, *y*, and *w*, such that the set of positive measure permits a smooth integration. We write

$$f_Z(z)=\int_{\Omega_{Y,W}} f_{Z,Y,W}(z,y,w)=\int_{\Omega_{Y,W}}\lambda^3e^{-\lambda z}1_{0\le w\le y\le z<\infty}$$
$$=\lambda^3e^{-\lambda z}1_{0\le z<\infty}\int_0^z\int_0^y dwdy=\lambda^3e^{-\lambda z}\frac{z^2}{2}1_{0\le z<\infty}=\frac{\lambda^3}{\Gamma(3)}z^2e^{-\lambda z}1_{0\le z<\infty}$$

In fact, continuing in this manner, one can use an induction argument to find the distribution of *n* i.i.d. exp(λ) random variables.

Another approach to this problem invokes the moment generating function. We know that if *X* follows an exp(*λ*), then its moment generating function $\mathbf{M}_X(t)=\frac{\lambda}{\lambda-t}$. We have also seen from our initial discussion of moment generating functions that the moment generating function of the sum of independent random variables is the product of their individual moment generating functions. We write the random variable *Z* as the sum of *n* i.i.d. exp(λ) random variables. Then $\mathbf{M}_Z(t)=\left(\frac{\lambda}{\lambda-t}\right)^n$. If we write $f_Z(z)=\frac{\lambda^n}{\Gamma(n)}z^{n-1}e^{-\lambda z}1_{0\le z<\infty}$, we have seen that the exponential measure of the real line with respect to this particular measuring tool is one. We find the moment generating function for this variable as

$$\mathbf{M}_Z(t)=\int_{\Omega_Z}e^{tz}d\mathbf{P}$$
$$=\int_{\Omega_Z}e^{tz}\frac{\lambda^n}{\Gamma(n)}z^{n-1}e^{-\lambda z}1_{0\le z<\infty}dz=\frac{\lambda^n}{\Gamma(n)}\int_0^\infty z^{n-1}e^{-(\lambda-t)z}dz.$$

Using what we know of working with gamma functions, we let $v=\left(\lambda-t\right)z$ and can compute

$$\frac{\lambda^n}{\Gamma(n)}\int_0^\infty z^{n-1}e^{-(\lambda-t)z}dz = \frac{\lambda^n}{\Gamma(n)}\int_0^\infty\left(\frac{v}{\lambda-t}\right)^{n-1} e^{-v}\frac{dv}{\lambda-t} = \left(\frac{\lambda}{\lambda-t}\right)^n\frac{\Gamma(n)}{\Gamma(n)}$$
$$=\left(\frac{\lambda}{\lambda-t}\right)^n$$

So, invoking the continuity theorem, $f_Z(z)=\frac{\lambda^n}{\Gamma(n)}z^{n-1}e^{-\lambda z}1_{0\le z<\infty}$, is the density of the sum of $n$ i.i.d. negative exponential random variables with parameter $\lambda$, and is therefore the distribution of the $n^{\text{th}}$ arrival.

The gamma distribution is commonly written as $f_X(x)=\frac{\alpha^r}{\Gamma(r)}x^{r-1}e^{-\alpha x}1_{0\le x<\infty}$. Being a function of two parameters, it has great flexibility of functional form. (Figure 3).

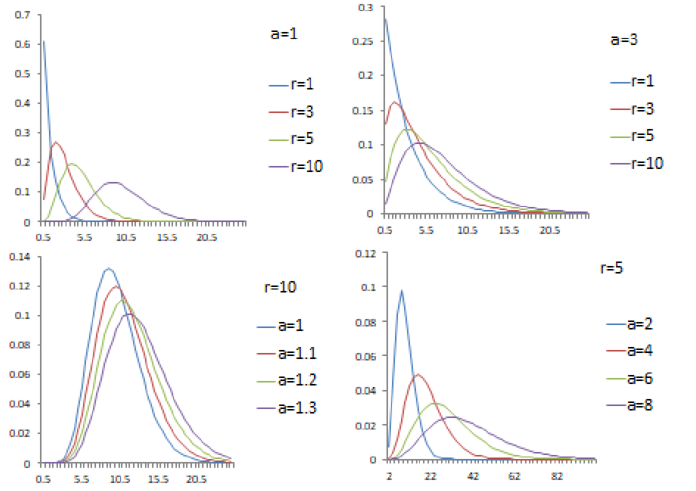


**Figure 3**. Shape of the gamma (a,r) distribution as a function of r and a

## Moments of the gamma distribution

The moments of the gamma distribution with parameters $\alpha$ and $r$ are readily available. We simply write.

$$\mathbf{E}\left[X^k\right]=\int_{\Omega_X}x^k d\mathbf{P}=\int_0^\infty x^k\frac{\alpha^r}{\Gamma(r)}x^{r-1}e^{-\alpha x}dx=\frac{\alpha^r}{\Gamma(r)}\int_0^\infty x^{r+k-1}e^{-\alpha x}dx.$$

We now let $v=\alpha x$, see that that region of positive measure does not change, $dx=\frac{dv}{\alpha}$, and write

$$\frac{\alpha^r}{\Gamma(r)}\int_0^\infty x^{r+k-1}e^{-\alpha x}dx=\frac{\alpha^r}{\Gamma(r)}\int_0^\infty\left(\frac{v}{\alpha}\right)^{r+k-1}e^{-v}\frac{dv}{\alpha}=\frac{\alpha^r}{\alpha^{r+k}\Gamma(r)}\int_0^\infty v^{r+k-1}e^{-v}dv$$
$$=\frac{\alpha^r}{\alpha^{r+k}\Gamma(r)}\frac{\Gamma(r+k)}{1}\int_0^\infty\frac{1}{\Gamma(r+k)}v^{r+k-1}e^{-v}dv \; = \; \frac{\Gamma(r+k)}{\alpha^k\Gamma(r)}$$

Thus $\mathbf{E}\left[X^k\right]=\frac{\Gamma(r+k)}{\alpha^k\Gamma(r)}$. This reveals $\mathbf{E}\left[X\right]=\frac{\Gamma(r+1)}{\alpha\Gamma(r)}=\frac{r}{\alpha}$, $\mathbf{E}\left[X^2\right]=\frac{\Gamma(r+2)}{\alpha^2\Gamma(r)}=\frac{r(r+1)}{\alpha^2}$, and we find

$$\mathbf{Var}\left[X\right]=\mathbf{E}\left[X^2\right]-\mathbf{E}^2\left[X\right]=\frac{(r+1)r}{\alpha^2}-\frac{r^2}{\alpha^2}=\frac{r}{\alpha^2}.$$

## Erlang distribution

The Poisson model has been instrumental in studying waiting times in many systems, e.g., queuing theory, trunk line issues in telephone communications, and electronics. Since waiting times are exponential, and the sum of i.i.d. exponential random variables, follows gamma measure, the gamma distribution is particularly important in complex waiting models. A particular version of the gamma distribution family is when $r$ is an integer *n*. In this case, the gamma distribution is traditionally known as the Erlang distribution, named for the father of the traffic engineering field.

## The Chi-square distribution

We have seen that the negative exponential distribution is related to the gamma distribution, i.e., the negative exponential ($\lambda$) then it is gamma $(1,\lambda)$. Thus the negative exponential distribution is a member of the gamma family. The chi-square distribution is simply a gamma distribution with its parameter *r* equal to an integer and its parameter $\alpha=\frac{1}{2}$. When we say that the random variable *X* follows a chi square $\left(\chi^2\right)$ distribution with *k* degrees of freedom we are really saying that *X* follows a gamma distribution with $r=\frac{k}{2}$ and $\alpha=\frac{1}{2}$. The probability density function is

$f_X(x)=\frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}1_{0\le x<\infty}$. Its moment generating function is

$\mathbf{M}_X(t)=\left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{k}{2}}=\left(\frac{1}{1-2t}\right)^{\frac{k}{2}}=(1-2t)^{-\frac{k}{2}}$. Just as the sum of independent gamma distributions $(\alpha,r)$ with the same parameter $\alpha$ is itself gamma, so the sum of independent random variables that follow a $\chi^2$ distribution with the total number of degrees of freedom being the sum of the degrees of freedom of the summands.

This distribution plays an important role in inferential statistics. In our treatment of normal measure, we will see that the square of a standard normal random variable follows the chi-square distribution.

We will see that the measuring tool for the distribution of the sample variance will be closely related to this distribution.

## Rayleigh distribution

The Rayleigh distribution is related to the gamma distribution. If we want to find the measure of the nonnegative reals using the measuring tool $xe^{-\frac{x^2}{2}}$, we would let $y=\frac{x^2}{2}$. For the non-negative real line, this is one-to-one transformation and the region of measure remains unchanged.

Continuing, we see $dy = 2x\frac{dx}{2} = xdx.$ Thus, $\int_0^\infty xe^{-\frac{x^2}{2}}dx = \int_0^\infty e^{-y}dy = 1.$ The Rayleigh distribution takes advantage of this. It's probability density function $f_X(x)$ is commonly written as $\frac{x}{\beta}e^{-\frac{x^2}{2\beta}}1_{0\le x<\infty}.$ The transformation argument just developed shows that this is equivalent to a negative exponential distribution. For example, to find $\mathbf{P}[X<a],$ we compute $\int_0^a \frac{x}{\beta}e^{-\frac{x^2}{2\beta}}dx.$ We allow $y = \frac{x^2}{2\beta}.$ The interval $0\le x\le a$ maps to $0\le y\le \frac{a^2}{2\beta},$ and $dy = 2x\frac{dx}{2\beta} = \frac{x}{\beta}dx.$ Thus,

$$\mathbf{P}[X<a] = \int_0^a \frac{x}{\beta}e^{-\frac{x^2}{2\beta}}dx = \int_0^{\frac{a^2}{2\beta}} e^{-y}dy = 1-e^{-\frac{a^2}{2\beta}}.$$

## Weibull distribution

We can generalize the Rayleigh distribution by letting our measuring tool be $\frac{k}{\beta^k}x^{k-1}e^{-\left(\frac{x}{\beta}\right)^k}1_{0\le x<\infty}.$ We can find probabilities for this distribution by the transformation related to that discussed for the Rayleigh distribution. For example, to find $\mathbf{P}[X<a],$ we need $\int_0^a \frac{k}{\beta^k}x^{k-1}e^{-\left(\frac{x}{\beta}\right)^k}dx.$ We allow $y = \left(\frac{x}{\beta}\right)^k$. The interval $0\le x\le a$ maps to $0\le y\le \left(\frac{a}{\beta}\right)^k$, and $dy = \frac{k}{\beta^k}x^{k-1}dx.$ Thus,

$$\mathbf{P}[X<a] = \int_0^a \frac{k}{\beta^k}x^{k-1}e^{-\left(\frac{x}{\beta}\right)^k}dx = \int_0^{\frac{a^k}{\beta^k}} e^{-y}dy = 1-e^{-\left(\frac{a}{\beta}\right)^k}.$$

Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# The Measure of Ordering

## Prerequisites

An Introduction to the Concept of Measure
Lebesgue Integration Theory and the Bernoulli Distribution
Conditional Probability
General Poisson Process
Continuous Probability Measure
Variable Transformations
Uniform and Beta Measure
Survival Measure: Exponential, Gamma, and Related

Our work thus far focuses on the measure of a single or collection of random variables. With the exception of examples in using the transformation of variables for distributions such as the gamma distribution, we have not troubled ourselves with the size order of the random variables.

However, an important component of applied probability is the measure associated with the relative order of random variables. When one is formulating dosing strategies for small peptides, for example, then it is not sufficient to focus only on the mean dose; the expected quantity of the minimum concentration is also an important consideration to avoid the likelihood of underdosing.

Similarly, in a study where one focuses on a change in mean diastolic blood pressure between the treatment and the control group, although the mean diastolic pressure may have only decreased three or four millimeters of mercury, the movement in the entire distribution of blood pressures is substantial. By decreasing the mean diastolic blood pressure, the maximum blood pressure is also reduced, and since strokes rates are closely related to blood pressure, decreasing the maximum diastolic blood pressure decreases the overall stroke incidence rate.

The distribution of minimum, maximum, median, and the range of random variables has important implications in the implementation and impact of technology in public health. [1]

## Some initial terminology

We will use some familiar lay language in this discussion of random variables. For example, the smallest of a collection of $n$ observations is the first order statistic, or the minimum.

The next largest random variable is the second order statistic, the third largest is the third order statistic, and so on. The $n^{th}$ order statistic is the maximum. The $n-1^{st}$ order statistic is the penultimate order statistic, and the $n-2^{nd}$ is the antepenultimate order statistic.

The median is the order statistic "in the middle", i.e., it is the observation that is the 50th percentile value in the sample of observations.

## Rank ordering

Our work thus far has focused on a collection of random variables $Y_1, Y_2, Y_3, ..., Y_n$. In these considerations, be they creatinine measures or LDL cholesterol estimates, we have not considered their position based on their magnitude in the sample. For example, no thought has been given to the maximum arrival time of an ambulance, for example; only to the measure of any one of them.

Order statistics focus on not just the random variable, but on the magnitude of the observation. This chapter's focus is on the position of the random variable in the random variable sequence. As we will see, some of these computations we are already able to carry out.

**Example:** Two planes are dispatched to pick up and organ for transplantation. They have the same flight plan, and leave at the same time from dual runways at the same airport. What are the chances that Plane A arrives before Plane B.

Let's assume that the time to arrival of each plane follows a negative exponential distribution with parameter $\lambda$. If $T_A$ is the time until the Plane A arrives, and $T_B$ the time until Plane B arrives, then we need the $\mathbf{P}\left[T_A \leq T_B\right]$.

But, before we begin a detailed probabilistic assessment, lets develop some intuition about the problem. Each plane has the same distance to fly, and same duration of flight.

Given they fly independently of each other, begin their flight simultaneously, and have the same probability measure of flight time, is there any reason that one should arrive ahead of the other? Although the they might not arrive at exactly the same time, wouldn't we expect Plane A to arrive first approximately 50% of the time? Our intuition suggests that

$$\mathbf{P}\left[T_A \leq T_B\right] = \frac{1}{2}.$$

Furthermore, our intuition tells us that there are three possibilities; $T_A < T_B; T_A = T_B; T_A > T_B$. The second possibility we eliminate since, as we saw from our introduction to continuous probability measure, this event occurs with measure zero.

As for the remaining two, should they not have equal probabilities if each event follows the same distribution? Can we show this?

Let's create the function $1_{[T_A > T_B]}$. Then $\mathbf{P}\left[T_A > T_B\right] = \mathbf{E}\left[1_{[T_A > T_B]}\right]$. Thus

$$\mathbf{E}\left[1_{[T_A > T_B]}\right] = \iint\limits_{\Omega} 1_{[T_A \geq T_B]} \mathrm{d}\mathbf{P} = \int\limits_0^{\infty} \lambda e^{-\lambda t_A} \left[\int\limits_0^{t_A} \lambda e^{-\lambda t_B} dt_B\right] dt_A$$

$$= \int\limits_0^{\infty} \left(1 - e^{-\lambda t_A}\right) \lambda e^{-\lambda t_A} dt_A = 1 - \int\limits_0^{\infty} \lambda e^{-2\lambda t_A} dt_A$$

$$= 1 - \frac{1}{2} \int\limits_0^{\infty} 2\lambda e^{-2\lambda t_A} dt_A = \frac{1}{2}.$$

But this computation provides the same solution as for $\mathbf{P}\left[T_A < T_B\right]$, a finding that is ensured by the i.i.d. feature of the two random variables. In fact, from a more general sense, when $T_A$ and $T_B$

are independent and identically distributed, then these two events should be equally likely. Following the previous development, we have

$$\mathbf{E}\left[1_{[T_A<T_B]}\right]=\iint\limits_{-\infty\le T_A\le T_B\le\infty} 1_{[T_A\le T_B]} f_{T_A,T_B}\left(t_A,t_B\right)=\int\limits_{-\infty}^{\infty} f_{T_A}\left(t_A\right)\int\limits_{t_A}^{\infty} f_{T_B}\left(t_B\right)$$

$$=\int\limits_{-\infty}^{\infty}\left[1-F_{T_B}\left(t_A\right)\right]f_{T_A}\left(t_A\right)=\int\limits_{-\infty}^{\infty}\left[1-F_{T_A}\left(t_A\right)\right]f_{T_A}\left(T_A\right).$$

This resembles the continuous analogue to the convolution argument. Continuing,

$$\int\limits_{-\infty}^{\infty}\left[1-F_{T_A}\left(t_A\right)\right]f_{T_A}\left(t_A\right)$$

$$=\int\limits_{-\infty}^{\infty} f_{T_A}\left(t_A\right)-\int\limits_{-\infty}^{\infty} F_{T_A}\left(t_A\right)f_{T_A}\left(t_A\right)=1-\frac{1}{2}=\frac{1}{2}$$

using the probability integral transform to evaluate the last integral in the argument

## Complexity of measure heterogeneity

What simplified the argument above is the assumption concerning the identical nature of the measures. When we drop this assumption, the computations can become much more complicated.

## Example: B cells and viral infections

While there are over one hundred viruses that have been identified that cause the common cold, they each operate the same way. The viral particle, itself many times smaller than a cell, slips through the cell membrane, and makes its way to the cell's nucleus.* Once there, it invades the nucleus, and takes control, commanding the cell to make more viral particles. The cell complies, and, zombie-like, devotes more and more of its energy and resources to viral particles. Finally, the exhausted cell, depleted of resources, its volume taken up with viruses, bursts, releasing thousands of new viral particles that then go on to infect other cells. When enough damage is done to the epithelial cells in the nose, symptoms begin.

The body's specific response to this is antibodies.

Antibodies are tiny proteins whose three dimensional shape conforms to that of the agent against which it is target. This agent can be a virus, bacterium, fungi, protozoa, or foreign body.

Special cells, called B cells reside throughout our body. Their job is to make antibodies; however, each B cell makes one and only one specific antibody against one and only one foreign molecule or antigen. If the virus is recognized by a B cell, that B cell begins its antibody production, and also starts to proliferate, producing other B cells that make the same antibody. These antibodies along with other body responses (e.g. T cell activation, complement production, and cytokine generation) overwhelm the virus, denaturing it before it can enter a cell.

However, if the B cells take too long to make the antibodies (because no one B cell makes the specifically needed antibody and other cells, called undifferentiated B cells have to begin *de novo* to try to make an antibody that matches the viral particle) then the viruses infect thousands, and then hundreds of thousands of cells, producing symptoms.

So, in a sense, the immune system and the viruses are in a race. Should the B cells respond rapidly, the viruses are deactivated before they can generate any symptom-producing

* Some viruses can bypass the nucleus and go directly to the ribosome, where protein construction can take place directly.

injury. However, if the B cell reaction is slow, the viral particles are free to infect more and more nasal cells, producing billions of new viral particles and illness.

Now, let's try to parameterize this problem. Let's assume the random variable $V$ is the time it takes for the immune system to mount a sufficient defense to thwart the virus. Let $W$ be the random variable reflecting how long the virus needs to destroy enough cells to produce symptoms. Then $\mathbf{P}\left[V < W\right]$ is the probability that a sufficient immune response occurs before the virus has an opportunity to produce symptoms and the "cold" is averted.

We will assume that both $V$ and $W$ follow gamma measure. Define

$$f_V\left(v\right)=\frac{\beta^s}{\Gamma\left(s\right)}v^{s-1}e^{-\beta v}1_{0\le v\le\infty};\quad f_W\left(w\right)=\frac{\alpha^n}{\Gamma\left(n\right)}w^{n-1}e^{-\alpha w}1_{0\le w\le\infty}.$$

We also assume that $n$ is a positive integer.

We need the measure of the space $0 \le v \le w \le \infty.$ We begin by writing

$$\mathbf{P}\left[V < W\right] = \iint\limits_{0\le v\le w\le\infty.} f_{V,W}\left(v,w\right) = \iint\limits_{0\le v\le w\le\infty.} f_V\left(v\right)f_W\left(w\right)$$

$$=\int_0^\infty\left[\int_v^\infty f_W\left(w\right)\right]f_V\left(v\right)dv.$$

Note the use of Fubini's theorem to convert a double integral to an iterated one. Let's now evaluate the inner integral.

$$\int_v^\infty f_W\left(w\right)=\int_v^\infty\frac{\alpha^n}{\Gamma\left(n\right)}w^{n-1}e^{-\alpha w}.$$

This is usually considered a scaled version of an incomplete gamma function.* However, we can take advantage of the assumption that $n$ is an integer greater than one.

Let's first transform $T = W - V$: Then, from our consideration of transformation of variables, we can write $w = t + v : dw = dt : v \le w \le \infty \to 0 \le t \le \infty$ for this 1:1 transformation. Thus

$$\int_v^\infty f_W\left(w\right)dw=\int_0^\infty f_T\left(t\right)dt \text{ where } f_T\left(t\right)=\frac{\alpha^n}{\Gamma\left(n\right)}\left(t+v\right)^{n-1}e^{-\alpha(t+v)}.$$

Since $n$ is an integer, we can invoke the binomial theorem to write $\left(t+v\right)^{n-1}=\sum_{k=0}^{n-1}\binom{n-1}{k}v^k t^{n-1-k}.$

Thus,

* An incomplete gamma is commonly written as $\int_a^\infty x^{r-1}e^{-x}dx.$

$$\int_0^\infty f_T(t)\,dt = \int_0^\infty \frac{\alpha^n}{\Gamma(n)}(t+v)^{n-1}e^{-\alpha(t+v)}\,dt$$

$$= \frac{\alpha^n}{\Gamma(n)}e^{-\alpha v}\int_0^\infty (t+v)^{n-1}e^{-\alpha t}\,dt$$

$$= \frac{\alpha^n}{\Gamma(n)}e^{-\alpha v}\int_0^\infty \sum_{k=0}^{n-1}\binom{n-1}{k}v^k t^{n-1-k}e^{-\alpha t}\,dt$$

$$= \frac{\alpha^n}{\Gamma(n)}e^{-\alpha v}\sum_{k=0}^{n-1}\binom{n-1}{k}v^k\int_0^\infty t^{n-1-k}e^{-\alpha t}\,dt.$$

We can write the remaining integral as

$$\int_0^\infty t^{n-1-k}e^{-\alpha t}\,dt = \frac{\Gamma(n-k)}{\alpha^{n-k}}\int_0^\infty \frac{\alpha^{n-k}}{\Gamma(n-k)}t^{n-1-k}e^{-\alpha t}\,dt = \frac{\Gamma(n-k)}{\alpha^{n-k}},$$ [*]

and now have

$$\int_v^\infty f_W(w)\,dw = \frac{\alpha^n}{\Gamma(n)}e^{-\alpha v}\sum_{k=0}^{n-1}\binom{n-1}{k}v^k\,\frac{\Gamma(n-k)}{\alpha^{n-k}} = \sum_{k=0}^{n-1}\frac{(\alpha v)^k}{k!}e^{-\alpha v}.$$

Note that the expression $\frac{(\alpha v)^k}{k!}e^{-\alpha v}$ is Poisson measure.
We can now complete the calculation.

$$\mathbf{P}[V<W] = \int_0^\infty\left[\int_v^\infty f_W(w)\right]f_V(v)\,dv$$

$$= \int_0^\infty\sum_{k=0}^{n-1}\left[\frac{(\alpha v)^k}{k!}e^{-\alpha v}\right]\frac{\beta^s}{\Gamma(s)}v^{s-1}e^{-\beta s}1_{0\le v\le\infty}dv.$$

We begin by reversing the summation and integration procedures and segregate all times not involving the relevant variable $v$ to the left of the integral sign.

$$\frac{\beta^s}{\Gamma(s)}\sum_{k=0}^{n-1}\frac{\alpha^k}{k!}\int_0^\infty v^{s+k-1}e^{-(\alpha+\beta)v}\,dv.$$

We next evaluate the integral $\int_0^\infty v^{s+k-1}e^{-(\alpha+\beta)v}\,dv$ as

[*] This follows from $\binom{n-1}{k}\frac{\Gamma(n-k)}{\Gamma(n)} = \frac{(n-1)!}{k!(n-1-k)!}\frac{(n-k-1)!}{(n-1)!} = \frac{1}{k!}.$

$$\int_0^\infty v^{s+k-1}e^{-(\alpha+\beta)v}\ dv=\frac{\Gamma(s+k)}{(\alpha+\beta)^{s+k}}\int_0^\infty \frac{(\alpha+\beta)^{s+k}}{\Gamma(s+k)}v^{s+k-1}e^{-(\alpha+\beta)v}\ dv$$
$$=\frac{\Gamma(s+k)}{(\alpha+\beta)^{s+k}}.$$

Leaving

$$\mathbf{P}[V<W]=\sum_{k=0}^{n-1}\frac{\Gamma(s+k)}{\Gamma(s)k!}\frac{\alpha^k\beta^s}{(\alpha+\beta)^{s+k}}$$
$$=\sum_{k=0}^{n-1}\frac{\Gamma(s+k)}{\Gamma(s)k!}\left(\frac{\alpha}{\alpha+\beta}\right)^k\left(\frac{\beta}{\alpha+\beta}\right)^s.$$

Further simplification is afforded by assuming that $s$ is a positive integer $m$.

$$\mathbf{P}[V<W]=\sum_{k=0}^{n-1}\frac{\Gamma(m+k)}{\Gamma(m)k!}\left(\frac{\alpha}{\alpha+\beta}\right)^k\left(\frac{\beta}{\alpha+\beta}\right)^m$$
$$\sum_{k=0}^{n-1}\binom{m+k-1}{m-1}\left(\frac{\alpha}{\alpha+\beta}\right)^k\left(\frac{\beta}{\alpha+\beta}\right)^m$$

Thus, the final solution is the sum of negative binomial probabilities. This more complex calculation, was required because $V$ and $W$ were not identically distributed

## Covid-19 screening

During the 2020 Covid-19 pandemic, one of the many critical issues in the US was the ability to test citizens efficiently. Many testing centers were overwhelmed in June and July of that year with people interested in being tested (i.e., screened) for COVID positivity but who had to wait hours in line to be tested (and many days for results).

Let's examine this phenomenon using a simple case.

Assume that we have a single site that tests individuals one at a time. Subjects line up for testing. Let's invoke the Poisson process, and say that patients are arriving at a particular frequency $\lambda t$ and are tested at the frequency $\mu t$.

What can we say about the number of people queued to be screened at a particular point in time $t$?

If $X_t$ is a random variable reflecting arrivals to the system and $Y_t$ reflects screened departures, then define the random variable $W_t=X_t-Y_t$. Can a probability measure be computed for this new random variable? For example, what can say about the $\mathbf{P}[W_t=0]$?

$W_t$ is zero when $X_t=Y_t$, i.e., the number of Poisson arrivals *m* is the same as the number of departures. Recall from the immigration-emigration process that this is

$$\mathbf{P}\left[W_t = 0\right] = \sum_{m=0}^{\infty} \frac{\left(\lambda t\right)^m}{m!} e^{-\lambda t} \frac{\left(\mu t\right)^m}{m!} e^{-ut}$$
$$= \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda}{\lambda+u}\right)^m \left(\frac{\mu}{\lambda+u}\right)^m \frac{\left(\lambda t+\mu t\right)^m}{m!} e^{-(\lambda+u)t}.$$

If the arrival and screening rates were equivalent, we might expect to see average values of $W_t$ hover around zero.[*] However, since its variance increases with $m$ there are times in the system when the line relatively long, and other times in the system when there is no one in line and the screeners are idle.[†]

If $\lambda$ is much greater than $\mu$, there are many more arrivals in a given time then there are subjects who have been tested, and the average line for a test increases.

If $W_t = a$ then there are $a$ more arrivals in the queue than there are departures, the server is not able to keep up, and the average waiting queue lengthens. This situation occurs when no matter how many subjects $m$ are tested, there are $m+a$ arrivals. This probability is

$$\mathbf{P}\left[W_t = a\right] = \sum_{m=0}^{\infty} \frac{\left(\lambda t\right)^{m+a}}{\left(m+a\right)!} e^{-\lambda t} \frac{\left(\mu t\right)^m}{m!} e^{-ut}$$
$$= \sum_{m=0}^{\infty} \binom{2m+a}{m} \left(\frac{\lambda}{\lambda+u}\right)^{m+a} \left(\frac{\mu}{\lambda+u}\right)^m \frac{\left(\lambda t+\mu t\right)^{2m+a}}{\left(2m+a\right)!} e^{-(\lambda+u)t}.$$

We can therefore write that

$$\mathbf{F}_W(w) = \mathbf{P}\left[W_t \le w\right] = \sum_{a=0}^{w} \sum_{m=0}^{\infty} \frac{\left(\lambda t\right)^{m+a}}{\left(m+a\right)!} e^{-\lambda t} \frac{\left(\mu t\right)^m}{m!} e^{-ut}$$
$$= \sum_{a=0}^{w} \sum_{m=0}^{\infty} \binom{2m+a}{m} \left(\frac{\lambda}{\lambda+u}\right)^{m+a} \left(\frac{\mu}{\lambda+u}\right)^m \frac{\left(\lambda t+\mu t\right)^{2m+a}}{\left(2m+a\right)!} e^{-(\lambda+u)t}.$$

While this computation is interesting, it might be more helpful to observe the measure associated with the minimum and maximum value of $W_t$. But, from what perspective does it make sense to talk about the measure of the minimum $\left(m_{W_t}\right)$ and maximum value of $W_t, \left(M_{W_t}\right)$. As we will see, this involves changing our measure's $\left(\Omega, \Sigma\right)$ foundation.

■

## Transformation to rank ordering

[*] Actually "hovering" is somewhat inaccurate. The equality of the arrival and service rates is not a guarantee that line lengths will always be short. Even though the average line length is short, the variance of the queue length increases over time, providing wide swings in the number of subjects in the system

[†] This is commonly observed in grocery stores where customers seem to arrive at the check out counter in bunches, increasing the line length and wait time. This is because while arrival and service times are equivalent they are simply average rates and do reflect how arrivals are packed or spaced at any particular time.

We will need notational foundation to address events such as $\mathbf{P}\left[M_{W_t} > 3\right]$. Specifically, a transformation is required that converts a sequence of random variables that is unstructured by magnitude into one that is so organized.

A very simple maneuver will induce the structure that we need to evaluate the probabilities of more complicated events involving the ordered observations. Given a collection of observations $W_1, W_2, W_3, ...W_n$, how do we convert it to a sequence of rank ordered random variables from smallest to largest. We will denote this new sequence as $W_{[1]}, W_{[2]}, W_{[3]}, ...W_{[n]}$.

This is essentially an $n$ to $n$ transformation that requires us to use one of the most elusive features of the transformation process.

Recall that in our transformation of variable discussion, we wrote

$$f_{V,W}(v,w) = f_{X,Y}(v,w)\, J\left[(x,y) \to (v,w)\right]\left[\Omega(x,y) \to \Omega(v,w)\right].$$

The region is a change from an unordered one where the original measure is applied to each variable, to an ordered one, i.e., $W_{[1]} \le W_{[2]} \le W_{[3]} \le .. \le W_{[n]}$. The Jacobean of the transformation is 1.

However the number of mappings requires careful consideration. If we have $n$ observations, then the maximum $W_{[n]}$ could be created from any one of the original $n$ observations, i.e., we select one observation from $n$ of them. Once this selection is made, there are $n-1$ possibilities for $W_{[n-1]}$. This process is completed when one observation is left for $W_{[1]}$.

Thus there are $n!$ factorial mappings that are required to go from the original collection of random variables $W_1, W_2, W_3, ...W_n$ to the structured collection $W_{[1]}, W_{[2]}, W_{[3]}, ...W_{[n]}$. If we let $f$ be the joint density function of the unordered variables and $g$ the joint density function of the ordered variables, then we write

$$g(w_{[1]}, w_{[2]}, w_{[3]}, ...w_{[n]}) = n!\, f\left(w_{[1]}, w_{[2]}, w_{[3]}, ...w_{[n]}\right) 1_{[w_{[1]} \le w_{[2]}, \le w_{[3]} \le ... \le w_{[n]}]}.$$

## Measure of the minimum order statistic

We have a collection of tools from which to choose from when we are challenged with finding the measure of a function of a random variable. Moment generating functions, convolutions, using the geometry to complete the integration, and transformation of variables all come to mind.

In order to identify the measure of a single order statistic we will use two approaches. The first is a formal approach. The second is a more practical and observational method.

### *Formal approach to finding minimum measure*

The measure for identifying $V$, the minimum of a collection if i.i.d., random variables $\{X_i\}\ i = 1,2,3...n.$ with known cumulative distribution function $\mathbf{F}_X(x)$ and density function $f_X(x)$ begins with a simple observation surrounding this order statistic.

If $V$ is greater than a particular number, then every random variable from which $V$ was created must also be greater.

This gives us a useful jumping off point to begin to identify the cumulative distribution function of $V$, $F_V(v)$, and then, if the derivative exists, differentiate it to obtain $f_V(v)$, the density function as we showed in our introduction to continuous probability measure.

So, for all $v: -\infty \leq v \leq \infty$.

$$1-F_V(v)=P\left[V>v\right]$$
$$=P\left[\left\{X_1 \geq v\right\} \cap \left\{X_2 \geq v\right\} \cap \left\{X_3 \geq v\right\} \cap ... \cap \left\{X_n \geq v\right\}\right].$$

Since the random variables are i.i.d., we can simplify, writing,

$$P\left[\left\{X_1 \geq v\right\} \cap \left\{X_2 \geq v\right\} \cap \left\{X_3 \geq v\right\} \cap ... \cap \left\{X_n \geq v\right\}\right]$$
$$=\prod_{i=1}^{n} P\left[X_i \geq v\right]=\left[1-F_X(v)\right]^n .$$

We can now conclude $1-F_V(v)=\left[1-F_X(v)\right]^n$ or $F_V(v)=1-\left[1-F_X(v)\right]^n$. Finally, assuming that the derivative exist, we can conclude

$$f_V(v)=\frac{dF_V(v)}{dv}=\frac{d\left(1-\left[1-F_X(v)\right]^n\right)}{dv}$$
$$=-n\left[1-F_X(v)\right]^{n-1}\frac{d\left(-F_X(v)\right)}{dv}$$
$$=n\left[1-F_X(v)\right]^{n-1} f_X(v).$$

### *An observational approach to minimum measure*

An experiential approach begins with the view that, as in the previous demonstration, if *v* is the minimum, then $n-1$ observations must be greater than it. So identifying the density problem is not unlike that of identifying the form of the binomial distribution.

Here we select $n-1$ observations from $n$; for each of them compute the probability that they are greater than the minimum $v$ taking advantage of the i.i.d. assumption. The location of $v$ is covered by the density function of $X$ evaluated at the point $x=v.$ That gives us

$$f_V(v)=\binom{n}{n-1} f_X(v)\left(1-\mathbf{F}_X(v)\right)^{n-1}=nf_X(v)\left(1-\mathbf{F}_X(v)\right)^{n-1}.$$

## Maximum measure

We may follow the same approach to compute the measuring tool of the maximum order statistic $W$.

The formal approach requires the observation that if the maximum value of a collection of observations is less than a value $w,$ then all of the observations must also be less than $w.$ Following the preceding development for minimum measure, we proceed.

$$F_W(w)=P\left[W \leq w\right]$$
$$=P\left[\left\{X_1 \leq w\right\} \cap \left\{X_2 \leq w\right\} \cap \left\{X_3 \leq w\right\} \cap ... \cap \left\{X_n \leq w\right\}\right]$$
$$=\prod_{i=1}^{n} F_{X_i}(w) \ =\left(F_X(w)\right)^n .$$

Assuming that $\mathbf{F}_X\left(w\right)$ is differentiable, we can write

$$f_W(w) = \frac{d\mathbf{F}_{X_n}(w)}{dw} = \frac{d\left[\left(\mathbf{F}_X(w)\right)^n\right]}{dw} = n\left(\mathbf{F}_X(w)\right)^{n-1} f_X(w).$$

For the less formal approach, we begin with the location of the $n$ observations as one being the maximum and $n-1$ observations selected from n is less than the maximum.
We can therefore write

$$f_W(w) = \binom{n}{n-1}\left(\mathbf{F}_X(w)\right)^{n-1} f_X(w).$$

Similar approaches can be used to find the distribution of any percentile value, using the median.

### *Example: Cardiac MRI*

Consider an electronic component for a cardiac MRI that is constructed from a sequence of electric elements. Each of the elements must function for the assembled component to function properly. If any single one of the elements fails, a fault indicator lights, and the assembled unit must be replaced. (Figure 1)

Assume that the $i^{\text{th}}$ element of the $n$ elements in the sequence has a lifetime $t_i$ that is a random variable that follows an exponential distribution with parameter λ. We will also assume that the lifetimes of each of the $n$ elements are independent random variables that follow the same probability distribution. Our goal is to identify the expected lifetime of the electronic assembly.

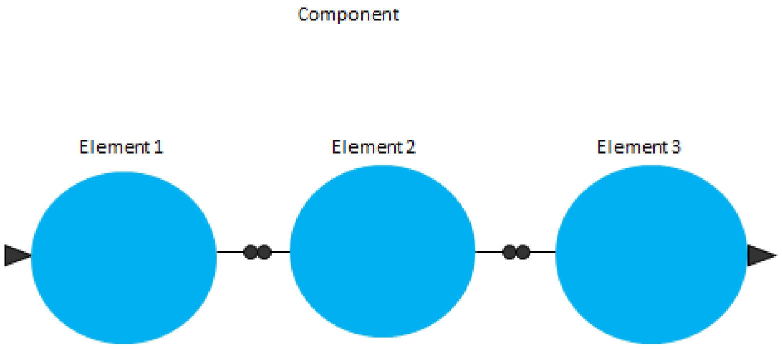


**Figure 1** Three elements that must work in series in order for the system's operation.

Consideration of this problem's construction reveals that the assembly fails when any one of its elements fails. Another way to say this is that the assembly fails with the very first component failure. Thus, the expected lifetime of the unit is the minimum lifetime of the $n$ elements. If $V$ is the minimum lifetime of the component and $X$ is the lifetime of an element in the assembly, then from our earlier work in this chapter

$$\mathbf{F}_V(v) = 1 - \left[1 - \mathbf{F}_V(v)\right]^n.$$

Since $f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x)$, we may write the cumulative probability distribution function of the component is $F_V(v) = 1 - e^{-n\lambda v}$, and write the density function as $f_V(v) = n\lambda e^{-n\lambda v} \mathbf{1}_{[0,\infty)}(v)\,dv$. Thus, the lifetime of the component is a random variable that follows an exponential distribution with parameter *n*λ. The expected lifetime of this component is $\mathbf{E}[v] = \frac{1}{n\lambda}$. Thus the average lifetime of the component is inversely proportional to the number of elements from which it is constructed. If complication is defined by the number of elements that are required to function in sequence (or in series) for the system to function, then the more complicated the system is, the shorter will be its expected lifetime.

## COVID-19 testing revisited

We can now compute the measure of the minimum and maximum number of subjects in a COVID-19 screening center.

However, we do have a paradigm shift that we must manage. When this problem was introduced, we focused on computing the measure of the system, which consisted of one testing station. The provided solution allowed us to compute the likelihood of various queue lengths. In fact we could examine the probability of small queue lengths to obtain an understanding of how short the queue is likely to be, as well as examine probabilities of large queue lengths, attempting to gauge the likely large lengths. This is information about the extreme queue lengths of a kind, and if we are focused on the single testing center, this is sufficient.

However, assume that we have *n* testing stations. Our interest switches from the performance of any particular one of them, to the performance of the system as a whole. What is the minimum time a person can be expected to wait given that they enter the system of *n* testing centers? How efficiently does the system operate?

This is the important perspective of order statistics. They educate us on the performance of the collection of units that make up the system's structure.

In addition, we must keep in mind, that the actual number in the system $W_t$ is a discrete integer, and not a continuous random variable. Therefore we cannot assume that the cumulative distribution function of the derivatives exists. However, we can utilize the heuristic approach to finding the order measure.

So, let's assume that we have *n* clinics operating independently of each other, with the same parameters $\lambda$ and $\mu$.

Recall that if $W_t$ is the number of subjects in a particular testing center, then

$$\mathbf{P}\left[W_t = a\right] = \sum_{m=0}^{\infty} \frac{(\lambda t)^{m+a}}{(m+a)!} e^{-\lambda t} \frac{(\mu t)^m}{m!} e^{-ut}$$

and its cumulative distribution function is

$$\mathbf{F}_W(w) = \mathbf{P}\left[W_t \le w\right] = \sum_{j=0}^{w} \sum_{m=0}^{\infty} \frac{(\lambda t)^{m+j}}{(m+j)!} e^{-\lambda t} \frac{(\mu t)^m}{m!} e^{-ut}.$$

Let's define $H_t$ as the minimum number of people in the system, and $I_t$ as the maximum. Working with the minimum first, we adapt our argument about minimum measure to write $\mathbf{P}\left[H_t = k\right]$.

$$\mathbf{P}\left[H_t = k\right] = nl(k, \lambda, \mu, t)\left(1 - L(k, \lambda, \mu, t)\right)^{n-1}$$

where $l\left(k,\lambda,\mu,t\right)\sum_{m=0}^{\infty}\frac{\left(\lambda t\right)^{m+k}}{\left(m+k\right)!}e^{-\lambda t}\frac{\left(\mu t\right)^{m}}{m!}e^{-ut}$

and $L\left(k,\lambda,\mu,t\right)=\sum_{j=0}^{k}\sum_{m=0}^{\infty}\frac{\left(\lambda t\right)^{m+j}}{\left(m+j\right)!}e^{-\lambda t}\frac{\left(\mu t\right)^{m}}{m!}e^{-ut}$

Similarly for the maximum

$$\mathbf{P}\left[I_t=k\right]=l\left(k,\lambda,\mu,t\right)\left(L\left(k,\lambda,\mu,t\right)\right)^{n-1}.$$

These formulations are the link between observational parameters $\lambda$ and $\mu$, estimates of which can be obtained from the field, and the anticipated overall performance of the clinics. One could also explore the values of the service rate $\mu$ needed to keep the likelihood that the maximum number of people at a clinic is likely to be below some upper bound.

## Example: Respirator construction

An important concern in the early phase of the COVID-19 infection surge was the availability of respirators, requiring consideration of their optimal use. Optimal use means that patients with the greatest need receive the respiratory care required to save their lives, *ceteris paribus*. Patients are placed on a respirator when required, and the respirator is discontinued with the patient has recovered and no longer requires it, or they are dead. Patients who are on a respirator for a short period of time may not have required it at all, but were placed on it in an "abundance of caution"

Let's look at the following A hospital has *n* respirators. The department would like to keep the respirator usage between a minimum and maximum duration of time in their system. We assume that the time of use of a respirator follows a negative exponential distribution with parameter $\lambda$.

If one wanted to look at the minimum and maximum time to constructing a respirator, they would only need to look at the first and the 99th percentile, for example. However, this is absent any experience, and is not informed by the generation of the sample.

Let's assume that patient need and therefore respirator use is i.i.d. What we seek is the measure of the range of use of the system of *n* respirators, where the range is simply the maximum minus the minimum.

We first solve this problem in general, and then contour the solution to the issue of respirator use.

### *General Solution*

Our plan will be to first identify the joint distribution of the minimum and maximum $V$ and $W$. We will then find the distribution of the range $R=W-V$.

We can use our heuristic approach to identify $f_{V,W}(v,w)$. Of the *n* observations in the sample, one must be the minimum $\left(f_X\left(v\right)\right)$, one must be the maximum $\left(f_X(w)\right)$, and the remaining $n-2$ observations fall in the $\left(v,w\right)$ interval on the real line, which occurs with probability $F_X\left(w\right)-F_X\left(v\right)$. Thus, we write

$$f_{V,W}(v,w)=\binom{n}{n-2}f_X(v)\left[F_X(w)-F_X(v)\right]^{n-2}f_X(w)$$
$$=\frac{n(n-1)}{2}f_X(v)\left[F_X(w)-F_X(v)\right]^{n-2}f_X(w).$$

We first need to convert the joint distribution of the minimum and maximum to the distribution of the random variable $R$. We do this by conducting a two variable to two variable conversion, and then integrate out the auxiliary variable.

Begin by defining $R=W-V:S=W$. Then the range of integration $0\le v\le w<\infty$ converts to $0\le r\le s<\infty$. A straightforward examination of the variables reveals that $W=S$, and $V=S-R$. The determinant of the Jacobian is one,

Then using our transformation rule for converting two random variables $V$ and $W$ from $X$ and $Y$,

$$f_{V,W}(v,w)=f_{X,Y}(v,w)J\left[(x,y)\to(v,w)\right]\left[\Omega(x,y)\to\Omega(v,w)\right].$$

We can now write

$$f_{V,W}(v,w)=\frac{n(n-1)}{2}f_X(v)\left[F_X(w)-F_X(v)\right]^{n-2}f_X(w)1_{0\le v\le w<\infty}$$
$$f_{R,S}(r,s)=\frac{n(n-1)}{2}f_X(s-r)\left[F_X(s)-F_X(s-r)\right]^{n-2}f_X(s)1_{0\le r\le s<\infty}.$$

Working on the middle term,

$$F_X(s)-F_X(s-r)=1-e^{-\lambda s}-\left(1-e^{-\lambda(s-r)}\right)$$
$$=e^{-\lambda(s-r)}-e^{-\lambda s}=e^{-\lambda s}\left(e^{\lambda r}-1\right).$$

Thus

$$\left[F_X(s)-F_X(s-r)\right]^{n-2}=\left[e^{-\lambda s}\left(e^{\lambda r}-1\right)\right]^{n-2}$$
$$=e^{-\lambda(n-2)s}\left(e^{\lambda r}-1\right)^{n-2}=e^{-\lambda(n-2)s}\sum_{j=0}^{n-2}\binom{n-2}{j}e^{\lambda jr}(-1)^{n-2-j}.$$

Invoking the binomial theorem for this final step.

Thus, our joint density is

$$f_{R,S}(r,s)$$
$$=\frac{n(n-1)}{2}f_X(s-r)\left[F_X(s)-F_X(s-r)\right]^{n-2}f_X(s)1_{0\le r\le s<\infty}$$
$$=\frac{n(n-1)}{2}\lambda e^{-\lambda(s-r)}e^{-\lambda(n-2)s}\sum_{j=0}^{n-2}\binom{n-2}{j}e^{\lambda jr}(-1)^{n-2-j}\lambda e^{-\lambda s}1_{0\le r\le s<\infty}$$
$$=\frac{n(n-1)}{2}\sum_{j=0}^{n-2}\binom{n-2}{j}e^{\lambda(j+1)r}(-1)^{n-2-j}\lambda^2e^{-n\lambda s}1_{0\le r\le s<\infty}.$$

Now, integrating with respect to s reveals.

$$
\begin{aligned}
f_R(r) &= \int\limits_{\Omega_s} f_{R,S}(r,s) \\
&= \frac{n(n-1)}{2} \int_r^\infty \sum_{j=0}^{n-2} \binom{n-2}{j} e^{\lambda(j+1)r} (-1)^{n-2-j} \lambda^2 e^{-n\lambda s} ds \\
&= \frac{n(n-1)}{2} \sum_{j=0}^{n-2} \binom{n-2}{j} e^{\lambda(j+1)r} (-1)^{n-2-j} \lambda^2 \int_r^\infty e^{-n\lambda s} ds \\
&= \frac{n(n-1)}{2} \sum_{j=0}^{n-2} \binom{n-2}{j} e^{\lambda(j+1)r} (-1)^{n-2-j} \lambda^2 \frac{1}{n\lambda} e^{-n\lambda r} \\
&= \frac{\lambda(n-1)}{2} \sum_{j=0}^{n-2} \binom{n-2}{j} e^{-\lambda(n-(j+1))r} (-1)^{n-2-j}.
\end{aligned}
$$

The measure of the range is a quite tractable sum of exponential functions. One could identify the value of $\lambda$ needed to keep the range as small as possible.

Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

References

1. Kapadia A, Chan W, and Moyė, Mathematical Statistics with Applications. Taylor and Francis.

# Basics of the Normal Distribution

The normal probability distribution is ubiquitous in probability in general and in health care research in particular. Although its mathematical computation requires the use of tables (unlike the binomial or Poisson distributions, many of whose probabilities can be computed directly from their formulas) this requirement has not hindered its widespread adoption, and now when we are surrounded by a universe of personal computing devices that can calculate normal distribution based probabilities (yes, there's an "app" for that), it can be readily computed in the field.

For all of these reasons, the normal distribution is the most commonly used distribution in probability and statistics, and has earned its sobriquet "normal" as it is the distribution "normally" used.

## Prerequisites

Why Probability
From Whence it Came – An Early History of Probability
Probability and the Renaissance
The Random Event
Elementary Set Theory
Properties of Probability

## First Concepts

The first thing to know about the normal distribution is that it does not provide probability for specific numbers but instead for regions of numbers. For example, if $X$ follows a normal distribution, the $\mathbf{P}\left[X=2\right]=0,$ while $\mathbf{P}\left[\frac{3}{2}\leq X\leq 5\right]$ has a nonzero value. For the normal distribution, as for other continuous distributions, we assign probability not to individual numbers but to intervals on the real number line. We will find that it is easy to adapt to this change.

The formula for the normal distribution contains two parameters, $\mu$ its mean, and $\sigma^2$ its variance. When the mean is zero and the variance is one, it becomes the "standard normal distribution" and its probability is governed by the function $f_Z(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$.

Fortunately, we do not have to evaluate this, but instead use a table or a computer application to compute probabilities. (Figure 1)

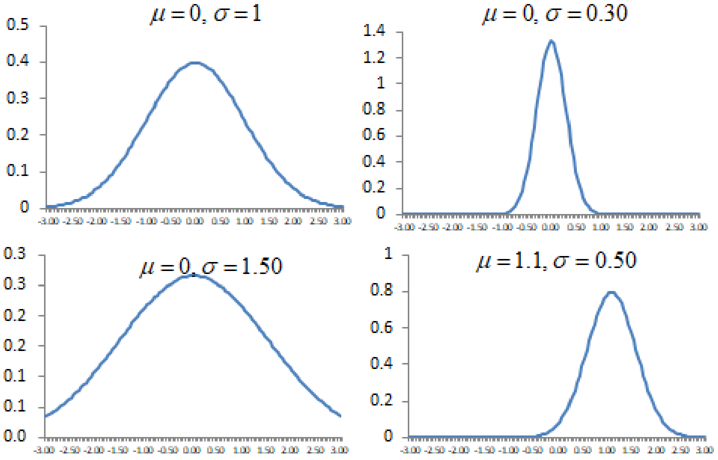


**Figure 1.** Different shapes and locations of the normal distribution.

Figure 1 demonstrates several interesting features. The standard deviation $\sigma$ reflects the spread or dispersion of the measure of the random variable. The mean $\mu$ is the location of the center of the distribution, i.e., its central tendency. In each case however, the distribution is symmetric about the mean.

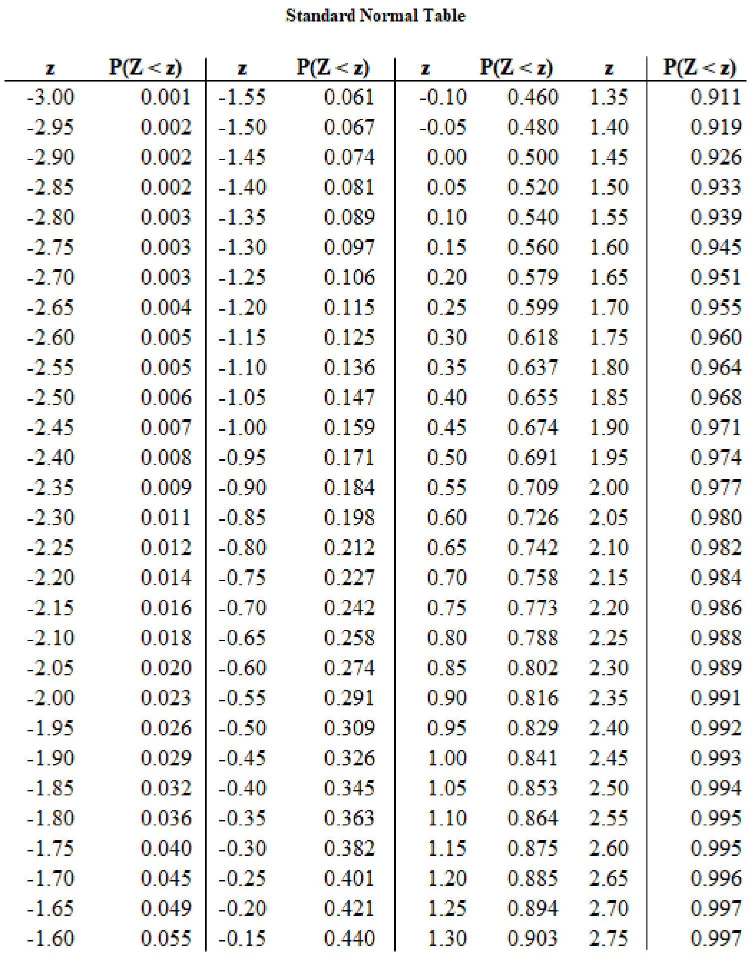
One of the most useful facilities of the standard normal random variable is the finding that its linear functions are also normally distribution. If *X* is normally distributed with mean $\mu$ and variance $\sigma^2$, then a new random variable $Y = aX + b$ is also normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$. We can use this to convert probabilities involving the general normal distribution to probabilities involving the standard normal distribution.

Typically a random variable that follows a standard normal distribution is denoted by *Z*. Let's say that we know that *X* follows a normal distribution with mean 10 and variance 25. We can compute $\mathbf{P}[X \le 15]$ by computing

$$\mathbf{P}[X \le 15] = \mathbf{P}\left[\frac{X-10}{5} \le \frac{15-10}{5}\right] = \mathbf{P}[Z \le 1],$$

And using the standard normal table, we see that $\mathbf{P}[Z \le 1] = 0.841$.

# Standard Normal Tables



Standard Normal Table

| z | P(Z < z) | z | P(Z < z) | z | P(Z < z) | z | P(Z < z) |
|---|---|---|---|---|---|---|---|
| -3.00 | 0.001 | -1.55 | 0.061 | -0.10 | 0.460 | 1.35 | 0.911 |
| -2.95 | 0.002 | -1.50 | 0.067 | -0.05 | 0.480 | 1.40 | 0.919 |
| -2.90 | 0.002 | -1.45 | 0.074 | 0.00 | 0.500 | 1.45 | 0.926 |
| -2.85 | 0.002 | -1.40 | 0.081 | 0.05 | 0.520 | 1.50 | 0.933 |
| -2.80 | 0.003 | -1.35 | 0.089 | 0.10 | 0.540 | 1.55 | 0.939 |
| -2.75 | 0.003 | -1.30 | 0.097 | 0.15 | 0.560 | 1.60 | 0.945 |
| -2.70 | 0.003 | -1.25 | 0.106 | 0.20 | 0.579 | 1.65 | 0.951 |
| -2.65 | 0.004 | -1.20 | 0.115 | 0.25 | 0.599 | 1.70 | 0.955 |
| -2.60 | 0.005 | -1.15 | 0.125 | 0.30 | 0.618 | 1.75 | 0.960 |
| -2.55 | 0.005 | -1.10 | 0.136 | 0.35 | 0.637 | 1.80 | 0.964 |
| -2.50 | 0.006 | -1.05 | 0.147 | 0.40 | 0.655 | 1.85 | 0.968 |
| -2.45 | 0.007 | -1.00 | 0.159 | 0.45 | 0.674 | 1.90 | 0.971 |
| -2.40 | 0.008 | -0.95 | 0.171 | 0.50 | 0.691 | 1.95 | 0.974 |
| -2.35 | 0.009 | -0.90 | 0.184 | 0.55 | 0.709 | 2.00 | 0.977 |
| -2.30 | 0.011 | -0.85 | 0.198 | 0.60 | 0.726 | 2.05 | 0.980 |
| -2.25 | 0.012 | -0.80 | 0.212 | 0.65 | 0.742 | 2.10 | 0.982 |
| -2.20 | 0.014 | -0.75 | 0.227 | 0.70 | 0.758 | 2.15 | 0.984 |
| -2.15 | 0.016 | -0.70 | 0.242 | 0.75 | 0.773 | 2.20 | 0.986 |
| -2.10 | 0.018 | -0.65 | 0.258 | 0.80 | 0.788 | 2.25 | 0.988 |
| -2.05 | 0.020 | -0.60 | 0.274 | 0.85 | 0.802 | 2.30 | 0.989 |
| -2.00 | 0.023 | -0.55 | 0.291 | 0.90 | 0.816 | 2.35 | 0.991 |
| -1.95 | 0.026 | -0.50 | 0.309 | 0.95 | 0.829 | 2.40 | 0.992 |
| -1.90 | 0.029 | -0.45 | 0.326 | 1.00 | 0.841 | 2.45 | 0.993 |
| -1.85 | 0.032 | -0.40 | 0.345 | 1.05 | 0.853 | 2.50 | 0.994 |
| -1.80 | 0.036 | -0.35 | 0.363 | 1.10 | 0.864 | 2.55 | 0.995 |
| -1.75 | 0.040 | -0.30 | 0.382 | 1.15 | 0.875 | 2.60 | 0.995 |
| -1.70 | 0.045 | -0.25 | 0.401 | 1.20 | 0.885 | 2.65 | 0.996 |
| -1.65 | 0.049 | -0.20 | 0.421 | 1.25 | 0.894 | 2.70 | 0.997 |
| -1.60 | 0.055 | -0.15 | 0.440 | 1.30 | 0.903 | 2.75 | 0.997 |

# Normal Measure

Normal measure is ubiquitous in probability in general and healthcare research in particular. Although its mathematical computation requires the use of tables (unlike the distributions e.g. Poisson, uniform, or negative exponential distributions whose formulas can directly provide the measure of events), this requirement has not hindered its widespread adoption. And, now that was are surrounded by a universe of personal computing devices that can calculate normal distribution based probabilities (yes, there's an 'app' for that), the measures can be readily computed in the field.

## Prerequisites

Basic Properties of Probability
Moment and Probability Generating Functions
Continuous Probability Measure
Variable Transformations
An Introduction to the Concept of Measure
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution

## Omnipresence

Normal measure's omnipresence is based in part on the demonstrations that 1) it is the exact distribution for some processes, (e.g., the diffusion of a gas as demonstrated by Einstein), and 2) linear combinations of normal distributions are normal (whether the individual random variables are independent or not).

However, the principal motivation for the widespread use of normal measure is the observation that linear combinations of non-normal random variables under conditions commonly observed in the natural and experimental world act like they are normally distributed. This means that exact probabilities computed on their exact distribution are very close to those computed assuming that the linear combination is normally distributed.

For all of these reasons, normal measure is the most commonly used distribution in probability and statistics, and has earned its sobriquet "normal" as it is the distribution "normally" used.

## Measuring tool

The probability density function for normal measure is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathbf{1}_{-\infty<x<\infty}.$$

It has two parameters, $\mu$ its mean, and $\sigma^2$ its variance. When the mean is zero and the variance is one, the measuring tool reduces to

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathbf{1}_{-\infty<z<\infty}.$$

And we commonly say that this is the "standard normal distribution"

Note that the term in the exponent $z^2$ without a corresponding term $z$ elsewhere in the integrand to help with the integration, differentiates this distribution from the exponential and gamma class of integrals. In fact, its shape is quite different (Figure 1).

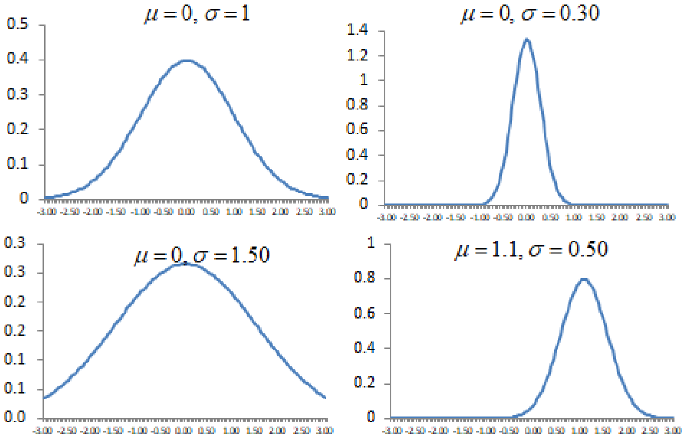


**Figure 1.** Different shapes and locations of the normal distribution.

Figure 1 demonstrates several interesting features. The standard deviation $\sigma$ reflects the spread or dispersion of the measure of the random variable. The mean $\mu$ reflects as expected, the location of the center of the distribution, i.e., its central tendency. In each case however, the distribution is symmetric about the mean. This symmetry is a property of which we will take great advantage.

However, how do we know that this measuring tool is a probability density function? Is the measure of the real line under its use one? It takes a little bit of work, but it can be proven that the measure of the real line using $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathbf{1}_{-\infty<z<\infty}$ is one.

## Linear transforms

One of the most useful facilities of the standard normal random variable is the finding that its linear functions are also normally distribution. Let $z$ follow a standard normal distribution. What is the distribution of $X = \sigma Z + \mu$?. We write $f_X(x) = f_z(x)\left|\frac{dx}{dz}\right|[\Omega_z \to \Omega_x]$. Since $-\infty < z < \infty$, there is no change in the region of measure, $Z = \frac{X-\mu}{\sigma}$, $dz = \frac{dx}{\sigma}$ and we can write

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{\sigma} \mathbf{1}_{-\infty<x<\infty} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathbf{1}_{-\infty<x<\infty}$$

which is the normal probability density function or measuring tool with mean $\mu$, and variance $\sigma^2$. Similarly, we can convert a random variable with any mean and variance into a normal distribution with mean zero and variance one. Thus, the computation of normal probabilities which is table based, can be condensed to probabilities for this latter distribution, which is commonly known as the standard normal distribution.

## Probability integral transforms revisited

Another example of interest that exemplifies the concept of the cumulative distribution function of a random variable as a random variable itself derives from the concept of expectation.

For example, suppose we are interested in computing $\mathbf{E}\left[F_Z(Z)\right]$ where $Z$ is a standard normal random variable. Recall that we have already demonstrated that the random variable $Y = F_Z(Z)$ follows a uniform distribution on the [0,1] interval, so we know at once that

$$\mathbf{E}\left[F_Z(Z)\right] = \mathbf{E}[Y] = \frac{1}{2}.$$

We can use this device to evaluate many expectations that involve the cumulative distribution function of the random variable of interest. For example, in order to find the $\mathbf{E}\left[1-\mathbf{F}_Z(Z+A)\right]$ where $Z$ is a standard normal random variable, consider $\mathbf{P}[X-Y>A]$ where $X$ and $Y$ are independent standard normal random variables. Then we can write

$$\begin{aligned}\mathbf{P}\left[X-Y>A\right] &= \iint_{x-y>A} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{y+A}^{\infty} f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty}\left(1-\mathbf{F}_X(Y+A)\right)f_Y(y)dy = \int_{-\infty}^{\infty}\left(1-\mathbf{F}_Y(Y+A)\right)f_Y(y)dy \\ &= \mathbf{E}\left[1-\mathbf{F}_Y(Y+A)\right] = \mathbf{E}\left[1-\mathbf{F}_Z(Z+A)\right].\end{aligned}$$

Alternative, we also know that

$$\begin{aligned}\mathbf{P}[X-Y>A] &= \mathbf{P}\left[W>A\right] = \mathbf{P}\left[X-Y>A\right] = \mathbf{P}\left[N(0,2)>A\right] \\ &= \mathbf{P}\left[N(0,1)>\frac{A}{\sqrt{2}}\right] = 1-\mathbf{F}_Z\left(\frac{A}{\sqrt{2}}\right).\end{aligned}$$

Combining these two results, we find that $\mathbf{E}\left[1-\mathbf{F}_Z(Z+A)\right] = 1-\mathbf{F}_Z\left(\frac{A}{\sqrt{2}}\right)$, or

$$\mathbf{E}\left[\mathbf{F}_Z(Z+A)\right] = \mathbf{F}_Z\left(\frac{A}{\sqrt{2}}\right).$$

## Moments and MGFs.

We have described the parameters of normal measure as mean and variance. Direct integration finds the mean for us. We compute

$$\mathbf{E}\left[X\right] = \int_{\Omega_x} xd\mathbf{P} = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Let $Z=\dfrac{X-\mu}{\sigma}$. Then there is no change in the region of integration, $X=\sigma Z+\mu$, $dx=\sigma dz$, and

$$\begin{aligned}\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx &= \int_{-\infty}^{\infty}(\sigma z+\mu)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{z^2}{2}}dz\\ &= \int_{-\infty}^{\infty}\frac{\sigma z}{\sqrt{2\pi\sigma^2}}e^{-\frac{z^2}{2}}dz+\mu\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{z^2}{2}}dz\\ &= \int_{-\infty}^{\infty}\frac{\sigma z}{\sqrt{2\pi\sigma^2}}e^{-\frac{z^2}{2}}dz+\mu\end{aligned}$$

The remaining integral becomes

$$\int_{-\infty}^{\infty}\frac{\sigma z}{\sqrt{2\pi\sigma^2}}e^{-\frac{z^2}{2}}dz=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ze^{-\frac{z^2}{2}}dz=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0}ze^{-\frac{z^2}{2}}dz+\int_{0}^{\infty}ze^{-\frac{z^2}{2}}dz$$
$$=\frac{1}{\sqrt{2\pi}}\left[\int_{0}^{-\infty}-ze^{-\frac{z^2}{2}}dz+\int_{0}^{\infty}ze^{-\frac{z^2}{2}}dz\right]=0.$$

Thus $\mathbf{E}[X]=\mu.$ We can use an integration by parts argument to compute the variance, demonstrating $\mathbf{Var}[X]=\sigma^2.$

### *MGF of normal measure*

Computing $\mathbf{E}\left[e^{tz}\right]$ when Z is a standard normal distribution is worthy of a calculation here because the sequence of computations will help us in our discussion of compounding normal distributions. We begin by writing

$$\mathbf{E}\left[e^{tz}\right]=\int_{\Omega}e^{tz}d\mathbf{P}=\int_{-\infty}^{\infty}e^{tz}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+tz}dz$$

We now develop the exponent

$$\begin{aligned}-\frac{z^2}{2}+tz &= -\frac{1}{2}\left(z^2-2tz\right)=-\frac{1}{2}\left(z^2-2tz+t^2-t^2\right)\\ &= -\frac{1}{2}\left(z^2-2tz+t^2\right)+\frac{t^2}{2}=-\frac{1}{2}(z-t)^2+\frac{t^2}{2}.\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{E}\left[e^{tz}\right]=\int_{\Omega}e^{tz}d\mathbf{P}=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+tz}dz &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2+\frac{t^2}{2}}dz\\ =e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2}dz &= e^{\frac{t^2}{2}},\end{aligned}$$

recognizing that $\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2}dz$ is the measure of the real line using a normal mean *t*, variance one measuring tool.

To compute the moment generating function of a random variable $X$ that is normal with mean $\mu$ and variance $\sigma^2$, we know that a standard normal random variable $X = Z\sigma + \mu$, and

$$\mathbf{E}\left[e^{tx}\right] = \mathbf{E}\left[e^{t(\sigma z+\mu)}\right]$$

$$= e^{t\mu}\mathbf{E}\left[e^{t\sigma z}\right] = e^{\frac{(t\sigma)^2}{2}+t\mu}.$$

## Explanation of the central limit theorem

We will later have a discussion about the mathematics behind the central limit theorem. However, it will serve us well at this point to make some simple observations.*

First, the measuring tool for normal measure is non-intuitive. At first glance, one can be forgiven for thinking that it is too complicated to serve as the distribution of any but artificially generated random variables. While the same might be said of the gamma distribution, at least we showed that it could be built up from sums of random variables that followed the negative exponential distribution. So far we have only shown that a normally distributed random variable is only the sum or difference of other normally distributed random variables.

Yet, its symmetry and central tendency seem to serve as an attractor for the distribution not of other individual random variables which are not normal, but their sums.

Recall for example the uniform distribution. In that discussion, we saw that while $X$, which is U(0,1) has no central tendency (in fact, equal intervals are equally likely, regardless of where they are located on the $[0,1]$ line, the sum of two such independent variables does exhibit a tendency to centralization of much of its measure (Figure 2)

As another example, consider a discrete random variable with the measuring tool $\mathbf{P}[X = x] = 0.50\,1_{x=0.01} + 0.50\,1_{x=0.99}$. This distribution divides its probability between two extreme values on $[0,1]$.

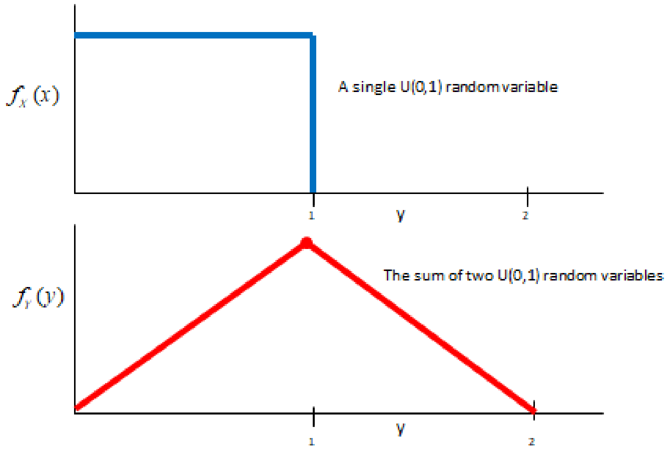


**Figure 2.** Note the movement to central tendency when one sums two U(0,1) random variables.

* This dicussion is adapted from Kapadia A, Chen W, Moyé L.Mathematical Statistics with Applications.

Yet if we accumulate sums of independently and identically distributed random variables that follow this distribution, we see this undeniable movement to centrality (Figure 3). Normality was not part of the summands, yet it appears to arise from their sum.

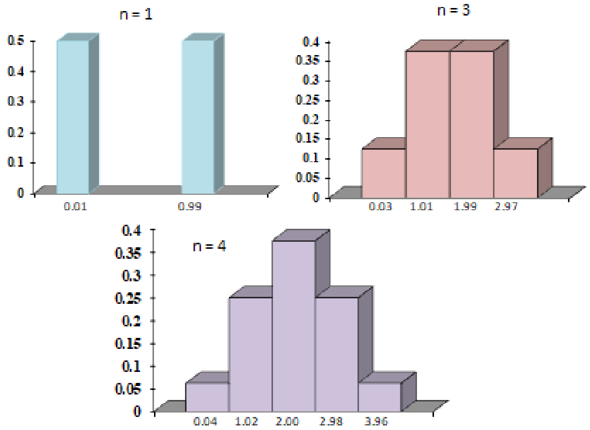


**Figure 3.** Movement to normality of the sums of dichotomous variables

That this movement to centrality and normality occurs is ineluctable. Why it should be so is worthy of exploration.

Consider for example two random variables each independent with probability mass function that places probability $\frac{1}{2}$ on each of 0 and 1. Note there is no central tendency here. We can easily examine the joint distribution of $X$ and $Y$ (Table 1).

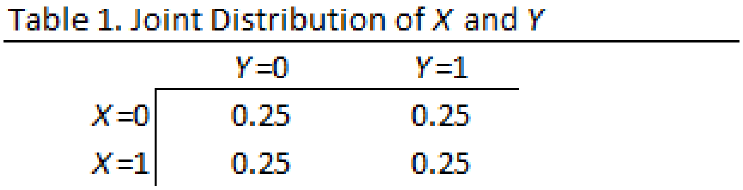


Table 1. Joint Distribution of $X$ and $Y$

| | $Y$=0 | $Y$=1 |
|---|---|---|
| $X$=0 | 0.25 | 0.25 |
| $X$=1 | 0.25 | 0.25 |

Each combination is equally likely as we would expect. However, if we now ask the question what is the distribution of the new random variable $Z = X + Y$, then centrality emerges (Table 2).

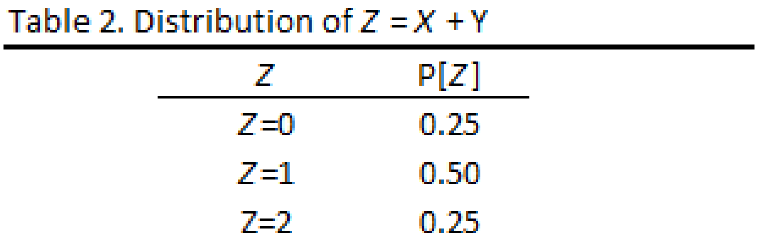


Table 2. Distribution of $Z = X + Y$

| $Z$ | P[$Z$] |
|---|---|
| $Z$=0 | 0.25 |
| $Z$=1 | 0.50 |
| Z=2 | 0.25 |

Examination of Table 1 shows us the source. In taking the sums of $(X,Y)$, we see that there were two ways to produce the intermediate value one ((0,1) and (1,0)), yet only one way to produce the extreme values (0,0), and (2,2). The increase in the number of ways to produce these intermediate values was the driving force to centrality.

Furthermore, the different possibilities for the summands were independent of the probabilities.

We can conclude that the number of intermediate values in such summands will always aggregate. How fast the probability aggregates depends on the probabilities themselves, but ultimately, they will follow the force produced by the increasing numbers of ways to produce the intermediate values. This is the explanation for Figure 3; even though the only nonzero probabilities began for the extreme events, the momentum toward centrality was irresistible.

As a final example, let's consider the sum of two random variables which not only show no inclination to central tendency (as one the case with the uniform random variable) but instead are the antithesis of central tendency. Consider the random variable whose measure is defined as

$$f_X(x) = 2x1_{0\le x\le 1}$$

In this case, most of the probability concentrates on the right portion of the region (Figure 4).

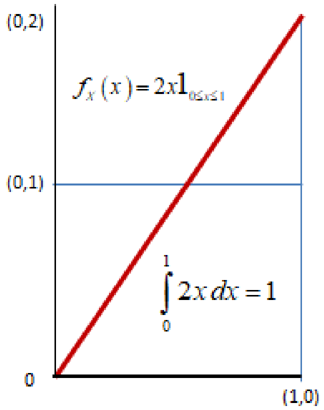


**Figure 4.** A random variable whose probability mass concentrates at the extreme of its region

If we define a second random variable $Y$ independent and with the same distribution as $X$, then define $Z = X + Y$ then what appearance would we expect for the probability density function of $Z, f_Z(z)$?

Recognizing that the range of Z is on interval $\left[0, 2\right]$ we would expect that there would be little concentration of measure close to zero, since the probability that each would produce a small value such that Z was also small is very low. We would expect most of the summand's measure to be close to two. The exact probability density function can be computed.

## Solution

It is

$$f_Z(z) = \frac{2}{3}z^3 1_{0\le z\le 1} + g(z)1_{1\le z\le 2}$$

Where

$g(z) = \dfrac{12z - 8 + 12\left(1 - z^2\right)\left(z-1\right) + 12z\left(z-1\right)^2 - 4\left(z-1\right)^3}{6}$. Its appearance shows a clear presence of central tendency (Figure 5).

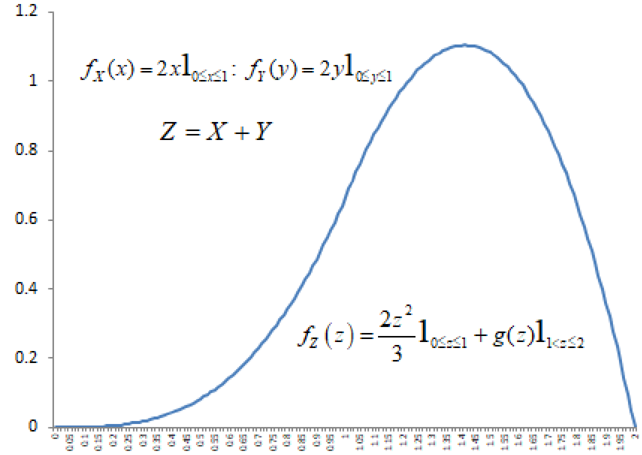


**Figure 5** Central tendency from the sum of two non-uniform random variables

Note that the probability of extremely large values is very small. While not yet symmetric based on the sum of two random variables the high likelihood of intermediate values of $z$ demonstrates the generation of central tendency that we might not expect from the sum on only two random variables both of which concentrate probability at their larger values.

A formal proof of this movement to central tendency will be discussed later when we review and prove the central limit theorem. This key finding states that when suitably normalized, the probability of events evolving sums can be approximated by the use of normal measure. Essentially, for our purposes here if we have a collection of i.i.d. random variables $X_i$, $i$ = 1, 2, 3, …, $n$ where $\mathbf{E}\left[X_i\right] = \mu$, and $\mathbf{Var}\left[X_i\right] = \sigma^2$, then for "large" $n$, we can approximate the probability distribution of $\dfrac{S_n - n\mu}{\sqrt{n\sigma^2}}$ by a standard normal distribution. Thus, probabilities of events involving $S_n$ which may be cumbersome to compute exactly, can be nicely and easily approximated.

### *Example: Clinical trial recruitment*

A clinical trial has seventy centers. Each recruits patients into the trial at a rate in accordance with a Poisson distribution with parameter $\lambda$ = 4 patients per month. What is the probability that the trial will reach its recruitment goal of 3200 recruited patients in a year.

We know that four patients per month translates to forty-eight patients per year for one center and (48)(70)=3360 patients per year on average. However, to compute the exact probability that this Poisson random variable is greater than 3200 is to compute $\sum_{k=3200}^{\infty} \dfrac{3360^k}{k!}e^{-3360}$, or $1 - \sum_{k=0}^{3200} \dfrac{3360^k}{k!}e^{-3360}$, either of which can be time consuming.

Alternatively, we could recognize that $S_n$ follows a Poisson distribution with mean 3200 and variance 3200. We then compute using the central limit theorem that

$$\begin{aligned}\mathbf{P}\left[S_{\mathrm{n}} \geq 3200\right] &= \mathbf{P}\left[\frac{S_n - 3360}{\sqrt{3360}} > \frac{3200-3360}{\sqrt{3360}}\right] \\ &= \mathbf{P}\left[N(0,1) > \frac{3200-3360}{\sqrt{3360}}\right] \\ &= \mathbf{P}\left[N(0,1) > -2.76\right] = 0.997.\end{aligned}$$

## "Within normal limits"

In health care, one of the most commonly used expressions among physicians, nurses, and other health care providers in the process of studying the characteristics and findings of their patients is "within normal limits".

This term is applied to heart measurements, blood sugar levels, plasma hormone assessments, serum cholesterol evaluations, and blood pressure measurements to name just a few. Essentially, a measurement is considered to be within normal limits if it falls within a range of values commonly seen in healthy people. Generally, normal measure is used to determine the values of these normal limits.

However, just how applicable is normal measure in this setting? As an example, let's consider one of the parameters that is assumed to follow a normal distribution, an individual's white blood cell count. White cells are particular cells that inhabit the blood and the lymphatic system.

These cells are part of the body's line of defense against invasion. Unlike red blood cells (erythrocytes) which are carried along passively by the currents and eddies of the blood stream, white blood cells are capable of independent motion, freely "choosing" their own directions of movement. However, they are especially attracted to toxins released by invading organisms and to compounds that are released by damaged cells.

When sensitized by these substances, these white blood cells react aggressively, attaching and moving through blood vessel walls, as they leave the blood stream and negotiate their way to the region of injury. Once they arrive at the site of cellular disruption, they produce substances that kill the invading organism (e.g., bacteria ), or destroy the foreign body (e.g., a wooden splinter).

White blood cells are short-lived, typically not surviving for more than 48 hours. However, their counts can dramatically increase with important systemic infections occur (e.g. pneumonia). White blood cells can also be produced in astonishing huge and damaging numbers when they are the product of cancer.

Clearly, there are many factors that affect the white blood cell count, and it would be difficult to see why the precise probability distribution of this count would be normal. Nevertheless, this is the probability distribution that is used to describe the white blood cell count.

While the use of normal measure is a natural consequence of the central limit theorem, the applicability of this theorem can be examined from another perspective in this example. There are many factors that influence the white blood cell count.

While we can think of a few (e.g. presence of infection, foreign bodies, cancer producing substances, hormone levels, compounds that are elaborated by other white cells) there are undoubtedly many, many more of these influences.

By and large, the impact of any single one of these influence is to either increase or decrease the white blood cell count by a small amount. Thus the white blood cell count is the result of the combined effect of all of these factors, each of which exerts only a small effect.

This is essentially what the central limit theorem states. The impact of the sum of many independent influences individually have a small effect, when suitably normalized, follows a standard normal distribution.

We can go one step further. Although the assumption of a normal distribution for the white blood cell count admits a wide range of possible values, 95% of the population whose white blood cell counts are healthy will have their count fall within 1.96σ of the population mean. Therefore, one could compute the mean μ and standard deviation σ in a population of subjects who have healthy white blood cell counts.

From this computation, the lower 2.5 percentile value (μ – 1.96σ) and the upper 97.5 percentile value (μ + 1.96σ) could be calculated. This region ($\mu$– 1.96σ , $\mu$ + 1.96σ ), commonly described as the 95% confidence interval, is the range of white blood cell counts that are "within normal limits".

The construction and use of this region represents an attempt to incorporate the observation that, while variability is a routine occurrence in nature, too much variability, while possibly normal, is the hallmark of an abnormality.

## Example: Ejection fraction

One measure of heart function is left ventricular ejection fraction (LVEF), which measures the percent of blood ejected from the heart. Ideally, the left ventricle which is the main pumping chamber of the heart pushes out at the end of each beat most all of the blood that it contains. Healthy individuals typically have ejection fractions of 80% or more.

Those patients who have had a heart attack can see their ejection fraction fall to 45% or less. Subjects who have heart failure can have ejection fractions as low as 10% to 15%, a number so low that they must have either a left ventricular assist device or a heart transplant to sustain their lives.

In hospitals, LVEF is most commonly measured by two modalities. One is magnetic resonance imaging. The outcome is continuous random variable which we will assume follows a normal distribution. However, sometimes patients undergo a bedside echocardiographic produced ejection fraction.

While one might ideally expect that this "bedside echo" is also normally distributed, characteristically, it is read in increments of 5 units. In addition, on average the echo based LVEF is seven absolute percentage points less than the MR echo.

Assume the chances that a patient who has just had a heart attack has a bedside echocardiogram is 60%, with a 40% probability that they will have an MR based LVEF determination. What is the probability that they will have an ejection fraction of 40 or less?

We begin by asking what can we say about the distribution of LVEF. The answer is of course easy if one has an MR based LVEF. We simply need to know its mean and variance. Similarly, if one knows the LVEF will be echo based, a measurement which tends to be in increments of five, the distribution of the discrete probabilities is also simple. However, what is one does not know whether the LVEF is either, but an individual simply has the measurement?

In this case, we must maintain flexibility with our measuring tool. In some regions we use the Riemann integral or area under the curve. In other regions, we use a "probability as mass" tool. We might write the probability function for the ejection fraction $X$ as

$$f_X(x) = 0.40\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} + 0.60\sum_{k=1}^{5} p_k 1_{x=a_k}.$$

This function has two "parts" revealing the two measuring tools that must be implemented to compute probabilities (Figure 6).

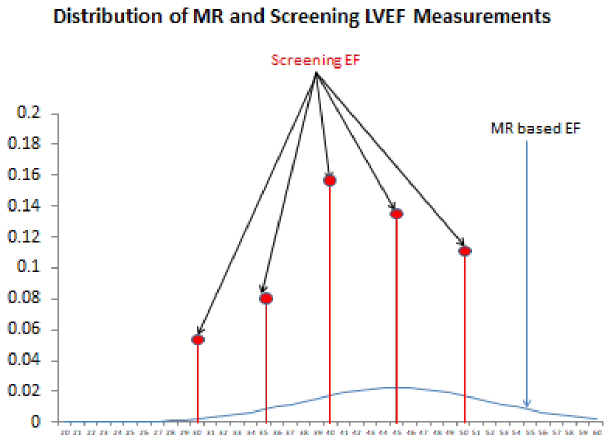


**Figure 6.** Distribution of LVEF by MR and screening echo modalities

We satisfy ourselves that the measure is one over the entire space (i.e., the entire real line) by seeing that the normal density integrates to one, and using point mass measure, probability over the discrete point measure integrates to one, the norming constants 0.40 and 0.60 being necessary for the measure to scale down to one over the entire real number line.

To compute $\mathbf{P}\left[X \leq 40\right]$, we simply accumulate measure over the real line, switching back and forth between the two measuring tools. This is

$$\int_{-\infty}^{40} f_X(x) = \int_{-\infty}^{40} \left( 0.40 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx + 0.60 \sum_{k=1}^{5} p_k \mathbf{1}_{x=a_k} \right)$$

$$= \int_{-\infty}^{40} 0.40 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx + 0.60 \sum_{k=1}^{5} p_k \mathbf{1}_{x=a_k \cap x \leq 40}$$

$$= 0.40\mathbf{P}\left[N\left(\mu,\sigma^2\right) \leq 40\right] + 0.60 \sum_{k=1}^{5} p_k \mathbf{1}_{x=a_k \cap x<40}.$$

If we let $\mu = 45, \sigma^2 = 10,$ we can complete the first part of this computation at once. To complete the screening component we need the of screening echocardiographic LVEF values and their probabilities. (Table 3).

<<Table 3>>

Thus we compute

$$\mathbf{P}\left[X \leq 40\right] = 0.60(0.309) + 0.40\left[0.10 + 0.15 + 0.30\right]$$
$$= 0.185 + 0.40\left[0.55\right] = 0.405.$$

## Chi-square distribution and $s^2$

Not only is normal measure of interest in its own right, but it produces other measures with their own applications. A good way to begin is with the distribution of the mean of the normal distribution.

We begin with taking a sample of observations $X_1, X_2, X_3, ...X_n$ which are independently and identically distribution $N\left(\mu,\sigma^2\right)$. We are seeking the distribution of the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, and what we will call the sample variance $s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$, which estimation theory tells us is the best estimate for the variance. If we were to take, say *m* samples of size *n* from the population, each sample producing an $\left(\overline{X}, s^2\right)$ pair, what would the probability distribution function look like?

Since normal measure has been so "user friendly" thus far, perhaps we are not so surprised to learn that in fact the distribution of the sample mean is independent of the sample variance. We will review here how that is demonstrated, and discuss its implications.

We are interested in determining the probability density function for $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, and $s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$, We will first write

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} \\ (n-1)s^2 &= \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \end{aligned}$$

And using some of the rules of sigma notation, we write

$$\begin{aligned} \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 &= \sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) \\ &= \sum_{i=1}^{n}\left(X_i^2\right) - 2\bar{X}\sum_{i=1}^{n}\left(X_i\right) + \sum_{i=1}^{n}\left(\bar{X}^2\right) \\ &= \sum_{i=1}^{n}\left(X_i^2\right) - 2\bar{X}n\bar{X} + n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \end{aligned}$$

Thus $(n-1)s^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$ or $\sum_{i=1}^{n} X_i^2 = (n-1)s^2 + n\bar{X}^2$.

Now, we can see that

$$\begin{aligned} \sum_{i=1}^{n}\left(X_i - \mu\right)^2 &= \sum_{i=1}^{n}\left(X_i^2 - 2X_i\mu + \mu^2\right) = \sum_{i=1}^{n} X_i^2 - 2\mu\sum_{i=1}^{n}\left(X_i\right) + n\mu^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2\mu n\bar{X} + n\mu^2. \end{aligned}$$

Now, knowing that $\sum_{i=1}^{n} X_i^2 = (n-1)s^2 + n\overline{X}^2$, we continue

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n} X_i^2 - 2\mu n\overline{X} + n\mu^2 = (n-1)s^2 + n\overline{X}^2 - 2\mu n\overline{X} + n\mu^2$$

$$= (n-1)s^2 + n(\overline{X} - \mu)^2$$

This we will need to establish that the probability distribution functions of the sample mean and sample variance are independent.

## Independent sample mean and variance

Our path will be to begin with the joint distribution of a sample of observations $X_1, X_2, X_3, \ldots X_n$ which are independently and identically distribution $N(\mu,\sigma^2)$. After a transformation of variables, we will use the results from the previous section to identify the desired result.

This is a daunting task, so we will first begin with a smaller one. Let $z$ be a standard normal random variable. What is the distribution of $w = z^2$?

We see at once that there is a change in the region of integration. While $-\infty \le z \le \infty$, $w$ must be nonnegative. This implies an integration factor of 2.*

From our discussion of transformation of variables we know that if we wish to create a new variable $y$ from $x$ we write

$$f_Y(y) = f_X(y)[dx \to dy][\Omega_x \to \Omega_y]$$

which for us in this case may be written

$$f_W(w) = f_Z(z)[dz \to dw][\Omega_z \to \Omega_w].$$

We know $\Omega_z \to \Omega_w$ is a mapping of $1_{-\infty < z < \infty}$ to $1_{0 \le w < \infty}$. Defining $w = z^2$ implies that $z = w^{\frac{1}{2}}, dz = \frac{1}{2} w^{-\frac{1}{2}}$, giving us all of the ingredients that we need. Since $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} 1_{-\infty<z<\infty}$ we can now write

$$f_W(w) = f_Z(w)[dz \to dw][\Omega_z \to \Omega_w]$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{w}{2}} \frac{1}{2} w^{-\frac{1}{2}} (2) 1_{0 \le w < \infty}$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\sqrt{\pi}} w^{-\frac{1}{2}} e^{-\frac{w}{2}} 1_{0 \le w < \infty}$$

* If $z$ were positive, the mapping would be 1 to 1. However, since $z$ can also be negative, the map is 2 to one, implying a multiplicative factor of 2.

It remains to show that $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$, which may seem alien on its face, but can be directly proved. Thus, the square of a standard normal variable follows a chi-squared distribution with 1 degree of freedom.

Returning to the original issue of demonstrating the independence of the sample mean and variance from a normal distribution. We will begin with the following many-to-many transformation of variables.
$Y_1 = \overline{X}, Y_2 = X_1 - \overline{X},\ Y_3 = X_2 - \overline{X},\ Y_4 = X_3 - \overline{X}, \ldots Y_n = X_{n-1} - \overline{X},$ and process through our transformation of variable operation

$$\begin{aligned}
&f_{Y_1,Y_2,Y_3,\ldots Y_n}(y_1,y_2,y_3,\ldots y_n) = \\
&f_{X_1,X_2,X_3,\ldots X_n}(y_1,y_2,y_3,\ldots y_n)J\left[(x_1,x_2,x_3,\ldots,x_n) \to (y_1,y_2,y_3,\ldots y_n)\right] \\
&\Omega(x_1,x_2,x_3,\ldots,x_n) \to \Omega(y_1,y_2,y_3,\ldots y_n).
\end{aligned}$$

This transformation does not change the region of measure, which remains over the entire real number line. The joint density of the original random variables $X_1, X_2, X_3, \ldots X_n$ is

$$f_{X_1,X_2,X_3,\ldots X_n}(x_1,x_2,x_3,\ldots x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Proceeding to scaling the space correctly, we can demonstrate that the Jacobean $J\left[(x_1,x_2,x_3,\ldots,x_n) \to (y_1,y_2,y_3,\ldots y_n)\right]$ is $n$. Finally, we have just calculated $\sum_{i=1}^{n}(x_i-\mu)^2 = (n-1)s^2 + n\left(\overline{X}-\mu\right)^2$. Thus,

$$\begin{aligned}
f_{Y_1,Y_2,Y_3,\ldots Y_n}(y_1,y_2,y_3,\ldots y_n) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-(n-1)\frac{s^2}{2\sigma^2}+\frac{n\left(\overline{X}-\mu\right)^2}{2\sigma^2}} n \\
&= \left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} e^{-(n-1)\frac{s^2}{2\sigma^2}}\right]\left[\frac{1}{\sqrt{2\pi\sigma^2/n}}e^{-\frac{n\left(\overline{X}-\mu\right)^2}{2\sigma^2/n}}\right].
\end{aligned}$$

We recognize $\frac{1}{\sqrt{2\pi\sigma^2/n}}e^{-\frac{n\left(\overline{X}-\mu\right)^2}{2\sigma^2/n}}$ as a normal distribution with mean $\mu$ and variance $\frac{\sigma^2}{n}$. The fact that the sample variance is in terms separate and apart from the newly found probability density function for the sample means permits us to conclude that they are independent, and we can pursue its distribution separately.

We know that if a random variable X follows a normal distribution with, then $\mathbf{E}\left[\overline{X}\right] = \mu.$ The distribution of $Z = \frac{X-\overline{X}}{\sigma}$ follows a standard normal distribution, and we have shown that $Z^2 = \left(\frac{X-\overline{X}}{\sigma}\right)^2$ follows a $\chi^2$ distribution with one degree of freedom. It follows from the development above that the individual deviations $X_1-\overline{X},\ X_2-\overline{X},\ X_3-\overline{X},\ ...X_n-\overline{X}$ are each independent of each other. Then we know that $\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{(n-1)s^2}{\sigma^2}$ follows a $\chi^2$ distribution with $n-1$ degrees of freedom. This finding of the independence of the sample mean and variance from a collection of independent normally distributed random variables is the foundation of many useful inferential procedures in statistics.

## Predicting future variance from the past

Let's assume that we have a sequence of observations $X_1, X_2, X_3, ..., X_m, ..., X_n...$ that are normal with mean $\mu$ and variance $\sigma^2$. We may compute a sequence of sample variances $s_2^2, s_3^2, s_4^2, ..., s_m^2, ..., s_n^2...$ Is there any way that we can compute the probability of values of the future variance $s_n^2$, based on knowledge of the variance at current point based on *m* observations, $s_m^2$ were $1<m<n$?

What we seek is $\mathbf{P}\left[s_n^2 \geq b | s_m^2 = a\right]$. Clearly they are related. Let's assume for a moment that $n=100$ and $m=99$. Then the distribution of $s_{100}^2$ will be well informed by the value of $s_{99}^2$ since they include the same measurements except the single observation $X_{100}$.

Alternatively, if $m=2$, then we might expect that the value of $s_{100}^2$ is not well informed by $s_2^2$ since there are many more observations included in $s_{100}^2$ driving its value either toward or away from the value of $s_2^2$.

Let's assume for a moment that the $\mu=0$, so $s_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, and $s_m^2 = \frac{1}{m}\sum_{i=1}^{m} X_i^2$. Then to compute $\mathbf{P}\left[s_n^2 \geq b | s_m^2 = a\right]$ we simply write

$$\begin{aligned}
\mathbf{P}\left[s_n^2 \geq b | s_m^2 = a\right] &= \mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 > b \;\Big|\; \frac{1}{m}\sum_{i=1}^{m} X_i^2 = a\right] \\
&= \mathbf{P}\left[\sum_{i=1}^{n} X_i^2 > nb \;\Big|\; \sum_{i=1}^{m} X_i^2 = ma\right] \\
&= \mathbf{P}\left[\sum_{i=1}^{m} X_i^2 + \sum_{i=m+1}^{n} X_i^2 > nb \;\Big|\; \sum_{i=1}^{m} X_i^2 \quad = ma\right]
\end{aligned}$$

This becomes

$$= \mathbf{P}\left[ma + \sum_{i=m+1}^{n} X_i^2 > nb\right] = \mathbf{P}\left[\sum_{i=m+1}^{n} X_i^2 > nb - ma\right]$$

Finally

$$\mathbf{P}\left[\frac{1}{\sigma^2}\sum_{i=m+1}^{n} X_i^2 > \frac{nb-ma}{\sigma^2}\right] = \mathbf{P}\left[\sum_{i=m+1}^{n}\left(\frac{X_i}{\sigma}\right)^2 > \frac{nb-ma}{\sigma^2}\right].$$

But $\sum_{i=m+1}^{n}\left(\frac{X_i}{\sigma}\right)^2$ follows a $\chi^2_{n-m-1}$ distribution, so $\mathbf{P}\left[s_n^2 \geq b | s_m^2 = a\right] = 1 - \mathbf{F}_{\chi^2_{n-m-1}}\left[\frac{nb-ma}{\sigma^2}\right]$.

However, how can we manage the case where $\mu \neq 0$? We know that the sample mean and variance are independent for a normal distribution, thus the sample variance is not a function of the mean. We can then compute $W_i = X_i - \overline{X}$ for $i$= 1 to $n$. Then each $W_i$ is normally distributed, $\mathbf{Var}\left[W_i\right] = \mathbf{Var}\left[X_i\right]$, and we can apply the derivation from above.

## Introduction to the F and T measure.

Another example of the probability distribution of the use of sample measures is the comparison of the ratio of sample variances. Assume that the random variables $V$ and $W$ are each independent $\chi^2$ distributions with $k$ and $m$ degrees of freedom respectively. Then through a process of transformations we can find the distribution of the random variable the random variables $X = \frac{V}{V+W}$, $G = \frac{V/k}{W/m}$, and $T = \sqrt{G}$. These derivations require some facility with transformations of random variables, and reveal that

$$f_X(x) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1}(1-x)^{\frac{m}{2}-1} 1_{0 \leq x \leq 1}$$

or a beta distribution that we recognize from previous discussions. The derivational result for the random variable $G$ is new, derived as

$$f_G(g) = \frac{\Gamma\left(\frac{m+k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{m}{k}\right)^{\frac{k}{2}} g^{\frac{k}{2}-1} \left(\frac{1}{\frac{m}{k}g+1}\right)^{\frac{m+k}{2}} 1_{0 \leq x \leq \infty}$$

This is an $F$ distribution with degrees of freedom $k$ and $m$. The square root of this random variable we call $T$ and it follows a Student's $t$ distribution whose measuring tool is

$$f_T(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{\pi}} m^{\frac{1}{2}} \left(\frac{1}{mt^2+1}\right)^{\frac{m+1}{2}} 1_{-\infty \leq t \leq \infty}$$

# Integrating the Normal Measure

Our task is to show that the measure of the real line using the measuring tool $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathbf{1}_{-\infty<x<\infty}$ is one.

## Using polar coordinates

The solution requires the use of a helpful transformation, polar coordinates. The problem is to demonstrate

$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1,$ equivalent to $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}.$ Let's begin by letting $A = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx.$ Then

$$A^2 = \left[\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx\right]\left[\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx\right] = \left[\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx\right]\left[\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{\left(x^2+y^2\right)}{2}} dxdy$$

This use of Fubini's theorem, allowing us to convert an iterated integral into a double integral is what permits the required transformation to polar coordinates.

We now conduct a transformation of variables to polar coordinates. Let $x = r\cos(\theta) : y = r\sin(\theta).$ Then, using our transformation of variables approach

$$f_{R,W}(r,\theta) = f_{X,Y}(r,\theta) J\left[(x,y) \to (r,\theta)\right]\left[\Omega(x,y) \to \Omega(r,\theta)\right].$$

The range of the measure is quite different. As we saw in the introduction to polar coordinates the range $-\infty < x < \infty; -\infty < y < \infty$ is transformed to $0 \le \theta \le 2\pi; 0 \le r < \infty.$ The Jacobian of the transformation is

$$\begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r\cos^2(\theta) + r\sin^2(\theta) = r.$$

Finally, since $x^2 + y^2 = r^2\cos^2(\theta) + r^2\sin^2(\theta) = r^2\left(\cos^2(\theta) + \sin^2(\theta)\right) = r^2,$ we can write $e^{-\frac{\left(x^2+y^2\right)}{2}}dx\,dy = re^{-\frac{r^2}{2}}dr\,d\theta.$ Now, completing the integration, we have

$A^2 = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} e^{-\frac{\left(x^2+y^2\right)}{2}}dxdy = \int\limits_{0}^{\infty}\int\limits_{0}^{2\pi} re^{-\frac{r^2}{2}}dr\,d\theta = \int\limits_{0}^{2\pi} d\theta \int\limits_{0}^{\infty} re^{-\frac{r^2}{2}}dr = 2\pi.$ Thus $A = \sqrt{2\pi},$ $\int\limits_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx = \sqrt{2\pi}$ , and $\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}dx = 1.$

# Deriving the Variance of Normal Measure

To compute the $\mathbf{Var}\left[X\right]$ where the measuring tool or probability density function is $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathbf{1}_{-\infty<x<\infty}$ we begin with $\mathbf{Var}\left[X\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}^2\left[X\right] = \mathbf{E}\left[X^2\right] - \mu^2$, demonstrating that our efforts must concentrate on finding $\mathbf{E}\left[X^2\right]$. We note

$$\mathbf{E}\left[X^2\right] = \int_{\Omega_x} x^2 d\mathbf{P} = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Let $Z = \frac{X-\mu}{\sigma}$. Then there is no change in the region of integration, $X = \sigma Z + \mu$, $dx = \sigma dz$, and

$$\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \left(\sigma z + \mu\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2}} \sigma dz$$
$$= \int_{-\infty}^{\infty} \left(\sigma^2 z^2 + 2\mu\sigma z + \mu\right)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$
$$= \int_{-\infty}^{\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz + \int_{-\infty}^{\infty} 2\mu\sigma z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz + \int_{-\infty}^{\infty} \mu^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

The second term becomes

$\int_{-\infty}^{\infty} 2\mu\sigma z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 2\mu\sigma \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0,$ since the mean of the standard normal distribution is 0. Thus we have

$$\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz \ + \mu^2$$

And the computation comes down to the evaluation of $= \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz.$ For this, we use integration by parts. Let

$$u = z; \ du = dz$$

$$dv = ze^{-\frac{z^2}{2}}; \ v = -e^{-\frac{z^2}{2}}$$

So $= \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz = \left[ ze^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = 0 + \sqrt{2\pi} = \sqrt{2\pi}.$

Thus $\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz \ + \mu^2 = \frac{\sigma^2}{\sqrt{2\pi}} \ \sqrt{2\pi} + \mu^2$

$$= \sigma^2 + \mu^2.$$

Thus $\mathbf{Var}\left[X\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}^2\left[X\right] = \mathbf{E}\left[X^2\right] - \mu^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$

# Sum of Two Linear Random Variables

Here, we are interested in deriving the sum of two specific, independent, and identically distributed random variables in order to demonstrate the generation of central tendency. We define each of $X$ and $Y$ to have positive measure on the $[0, 1]$ interval, where

$$f_X(x) = 2x\mathbf{1}_{0 \le x \le 1} : f_Y(y) = 2y\mathbf{1}_{0 \le y \le 1}$$

## Goal and methods

We desire the measure of their sum $Z = X + Y$. We first note that the range for Z is on $[0, 2]$. As we saw for the sum of two U(0,1) random variables the shape of the region of interest depends on the value of Z (Figure 1.)

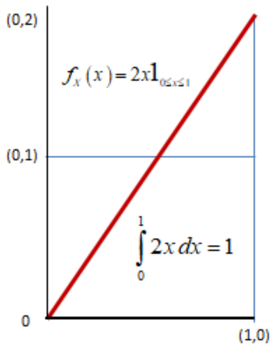


**Figure 1**. A random variable whose probability mass concentrates at the extreme of its region

As before, we note that there are two cases, depending of the value of $z$. We will manage each of these two cases separately

Case 1: $0 \le z < 1$. (Figure 2).

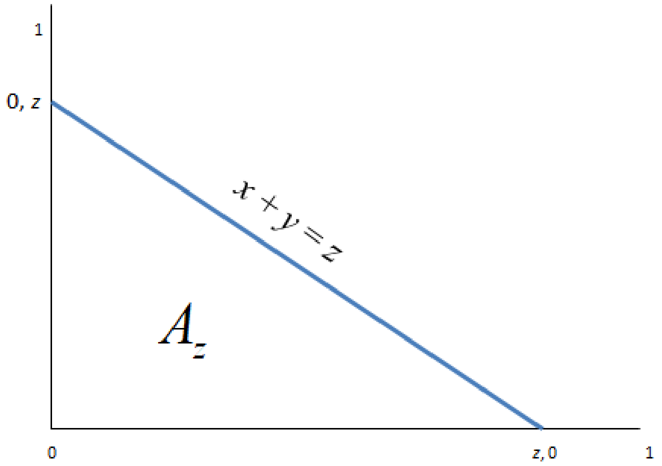


**Figure 2**. The sum of two random variables. Case 1. $Z \le 1$

Here, we compute for $0 \le z < 1$ $\mathbf{F}_Z(z) = \mathbf{P}[Z \le z] = \mathbf{P}[X + Y \le z]$. Examining the region from Figure 2 permits us to write

$$\mathbf{F}_Z(z) = \iint_{A_z} f_{X,Y}(x, y)dxdy = \int_0^z 2ydy \int_0^{z-y} 2xdx.$$

The second integral is $\int_0^{z-y} 2xdx = x^2 |_0^{z-y} = (z-y)^2$. Continuing

$$\begin{aligned}\mathbf{F}_Z(z) &= \int_0^z 2ydy \int_0^{z-y} 2xdx = \int_0^z 2y(z-y)^2 dy \\ &= 2\int_0^z y\left(z^2 - 2zy + y^2\right)dy = 2\int_0^z z^2 y - 2zy^2 + y^3 dy \\ &= 2\left[z^2 \frac{y^2}{2} - 2z\frac{y^3}{3} + \frac{y^4}{4}\right]_0^z = 2\left[\frac{z^4}{2} - \frac{2z^4}{3} + \frac{z^4}{4}\right] \\ &= 2\left[\frac{z^4}{2} - \frac{2z^4}{3} + \frac{z^4}{4}\right] = \frac{1}{6}\left[6z^4 - 8z^4 + 3z^4\right] = \frac{z^4}{6}.\end{aligned}$$

This is illumination. The fact that $\mathbf{F}_Z(1) = \frac{1}{6}$ tells us that most of the measure is to the right of one. We now proceed with Case 2.

Case 2: $1 \le z \le 2$. (Figure 3)

Here we must compute $\mathbf{F}_Z(z) = 1 - \mathbf{P}[B_z]$ $\mathbf{F}_Z(z) = 1 - \mathbf{P}[B_z] = 1 - \iint_{B_z} f_{X,Y}(x,y)dxdy$. And from an examination of Figure 3 we can write $\iint_{B_z} f_{X,Y}(x,y)dxdy = \int_{z-1}^{1} 2xdx \int_{z-x}^{1} 2ydy.$ Note here that $z \geq 1$. Now,

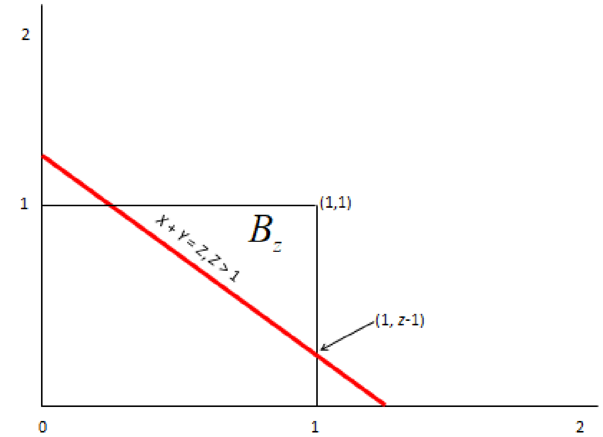


**Figure 3**. The sum of two random variables on [0,1] Case 2. Z > 1

$$\int_{z-x}^{1} 2ydy = y^2 |_{z-x}^{1} = \left(1 - z^2\right) + 2zx - x^2 .$$

Continuing,

$$\mathbf{P}[B_z] = \int_{z-1}^{1} 2\left[x\left(\left(1-z^2\right) + 2zx - x^2\right)\right] dx$$

$$= 2\int_{z-1}^{1} \left(1-z^2\right)x + 2zx^2 - x^3 dx$$

$$= 2\left[\frac{\left(1-z^2\right)x^2}{2} + \frac{2zx^3}{3} - \frac{x^4}{4}\right]_{z-1}^{1}$$

$$= \frac{1}{6}\left[6\left(1-z^2\right)x^2 + 8zx^3 - 3x^4\right]_{z-1}^{1}$$

This reduces to

$$\frac{1}{6}\left[6\left(1-z^2\right)x^2 + 8z - 3 - 6\left(1-z^2\right)\left(z-1\right)^2 - 8z\left(z-1\right)^3 + 3\left(z-1\right)^4\right]$$

Recalling that $\mathbf{F}_Z\left(z\right)=1-\mathbf{P}\left[B_z\right]$ we can see that $\mathbf{F}_Z\left(1\right)=1-B_Z(1)=1-\frac{5}{6}=\frac{1}{6}$ which is the solution we found for case 1. In addition $\mathbf{F}_Z\left(2\right)=1-B_Z(2)=1-0=1$. We can now take derivatives to compute

$$f_Z(z)=\frac{2}{3}z^3 1_{0\le z\le 1}+g(z)1_{1\le z\le 2}$$

where

$g(z)=\dfrac{12z-8+12\left(1-z^2\right)\left(z-1\right)+12z\left(z-1\right)^2-4\left(z-1\right)^3}{6}$. Its appearance shows a clear presence of central tendency.

# Determinant for an *n* to *n* Transformation

From the demonstration of the independence of the sample mean and variance of normal measure, we started with a sample of observations $X_1, X_2, X_3, \ldots X_n$ which are independently and identically distribution $N\left(\mu, \sigma^2\right)$, transforming to $Y_1 = \overline{X}, Y_2 = X_1 - \overline{X},\ Y_3 = X_2 - \overline{X},\ Y_4 = X_3 - \overline{X}, \ldots Y_n = X_{n-1} - \overline{X}.$

Examining this transformation in terms of the $X_1, X_2, X_3, \ldots X_n$, we have $X_2 = Y_2 + Y_1,\ X_3 = Y_3 + Y_1,\ X_4 = Y_4 + Y_1, \ldots X_n = Y_n + Y_1,$ and

$$X_1 = Y_2 + Y_3 + \ldots + Y_n - Y_1.$$

Our goal is to identify the Jacobian of this transformation. For the simple case of *n* = 2, we have

$$J\left[\left(x_1, x_2\right) \to (y_1, y_2)\right] = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 1 & 1 \end{vmatrix} = 2.$$

Expanding to *n* = 3 produces

$$J\left[\left(x_1, x_2, x_3\right) \to (y_1, y_2, y_3)\right] = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \frac{\partial x_3}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_3}{\partial y_2} \\ \frac{\partial x_1}{\partial y_3} & \frac{\partial x_2}{\partial y_3} & \frac{\partial x_3}{\partial y_3} \end{vmatrix} = \begin{vmatrix} -1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{vmatrix}$$

$$= \left|(-1)(1) - (1)(1) + (1)(-1)\right| = 3$$

We can take advantage of this pattern matrix. If we apply the ACBD lemma for determinant computation,

$$\left\|\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix}\right\| = |\mathbf{A}||\,\mathbf{A}-\mathbf{BDB}'|$$

Where **A** is an invertible *p* x *p* matrix and **D** is *r* x *r*.

Letting $\mathbf{A}=1,\ \mathbf{B}=\begin{bmatrix}1 & 1\end{bmatrix},$ and $\mathbf{D}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, we see

$$\begin{vmatrix} -1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{vmatrix} = |-1| - \left|\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix}\right| = \| -1-2 \| = 3.$$

One worked for us in this case is that D was the identity matrix. This converted $\mathbf{BDB}'$ to simply $\mathbf{BB}'$ which translates to 2, permitting the calculation $\|-1-2\| = 3.$

One more example

$$J\left[(x_1,x_2,x_3,x_4)\rightarrow(y_1,y_2,y_3,y_4)\right] = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \frac{\partial x_3}{\partial y_1} & \frac{\partial x_4}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_3}{\partial y_2} & \frac{\partial x_4}{\partial y_2} \\ \frac{\partial x_1}{\partial y_3} & \frac{\partial x_2}{\partial y_3} & \frac{\partial x_3}{\partial y_3} & \frac{\partial x_4}{\partial y_3} \\ \frac{\partial x_4}{\partial y_4} & \frac{\partial x_2}{\partial y_4} & \frac{\partial x_3}{\partial y_4} & \frac{\partial x_4}{\partial y_4} \end{vmatrix}$$

$$= \begin{vmatrix} -1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{vmatrix}$$

$$= |(-1)(1)-(1)(1)+(1)(-1)| = 3$$

Letting $\mathbf{A}=1,\ \mathbf{B}=\begin{bmatrix}1 & 1 & 1\end{bmatrix},$ and $\mathbf{D}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$, we see applying the ABCD lemma that

$$\begin{vmatrix} -1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{vmatrix} = 4.$$

and, when there is a sample of observations $X_1, X_2, X_3,...X_n$ the determinant of our transformation is $n$.

# Derivation of *F and T Measure*

## Initial comments and calculations

Here we will derive two of the most frequently used probability distributions in applied statistics. The derivation of these distributions can appear complex, but they are fine examples, allowing us to use the tools of transformation of variables that we discussed for the gamma, and beta distributions.

## Distribution of the normal square

We begin with finding the distribution of the square of a standard normal distribution function. Let *Z* follow a standard normal distribution. Then what is the measuring tool of $Y = Z^2$?

We first note that this function is mapping the $(-\infty, \infty)$ reals to $[0, \infty)$. So this is 2 to 1 mapping, an observation that we will have to take into account during the mathematics of this transformation. Also, since, most of the probability for the standard normal distribution is closer to rather than further away from zero, we would expect the same for the measuring tool of its square.

We formally begin by applying

$$f_X(x) = f_Z(x)\mathbf{D}\left[z \to x\right]\left[\Omega_Z \to \Omega_X\right]$$

The last expression, $\left[\Omega_Z \to \Omega_X\right]$just reflects the relative mapping spaces. Here we are mapping the $(-\infty, \infty)$ reals to $[0, \infty)$. This double mapping to the nonnegative reals means that we will need to multiply the resulting density by $2$. The function $z = x^{\frac{1}{2}}$ produces $dz = \frac{1}{2}x^{-\frac{1}{2}}dx.$ So we are ready to write

$$\begin{aligned}
f_X(x) &= f_Z(x)\mathbf{D}\left[z \to x\right]\left[\Omega_Z \to \Omega_X\right] \\
&= \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x}{2}}\right)\left(\frac{1}{2}x^{-\frac{1}{2}}\right)\left(2\,1_{0 \le x < \infty}\right) \\
&= \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)}x^{-\frac{1}{2}}e^{-\frac{x}{2}}1_{0 \le x < \infty}.
\end{aligned}$$

This recalls that $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$. This we recognize as a chi-square distribution with one degree of freedom. Since we know the moment generating function $\mathbf{M}_X\left[t\right]=\left(1-2t\right)^{-\frac{1}{2}}$, then $W=\sum_{i=1}^{n}X_i$ has moment generating function $\mathbf{M}_W\left[t\right]=\mathbf{E}\left[e^{wt}\right]=\mathbf{E}\left[e^{\sum_{i=1}^{n}Xt}\right]=\prod_{i=1}^{n}\mathbf{E}\left[e^{Xt}\right]=\left(1-2t\right)^{-\frac{n}{2}}$ is a $\chi_n^2$ distribution. Thus the measuring tool for the sum of the squares of $n$ i.i.d. standard normal random variables is a chi-square distribution with *n* degrees of freedom.

## Derivation of the F Distribution

Assume that the random variables *V* and *W* are each independent $\chi^2$ distributions with *k* and *m* degrees of freedom respectively. Then through a process of transformations we can find the distribution of the random variable the random variables $\frac{V/k}{W/m}$. In addition, it is sometimes of value to have the distribution of $\frac{V}{V+W}$. we will begin by identifying the probability density function of this second measuring tool.

Begin by letting $X=\frac{V}{V+W}$, $Y=V+W$. Our goal is to transform $(V,W)$ to $\left(X,Y\right)$ and then integrate out *Y*. (We know that *Y* will follow a $\chi^2$ distribution with *k* + *m* degrees of freedom, so we know what to watch for). We turn to our formula for transformation of variables:

$$f_{X,Y}(x,y)=f_{V,W}\left(x,y\right)J\left[\left(v,w\right)\rightarrow(x,y)\right]\left[\Omega\left(v,w\right)\rightarrow\Omega(x,y)\right].$$

We first attend to the region of positive measure. For $0\leq w\leq\infty, 0\leq v\leq\infty,$ we have $0\leq x\leq 1,\ 0\leq y\leq\infty.$ We find that $V=XY, W=Y\left(1-X\right)$

$$J\left[\left(v,w\right)\rightarrow(x,y)\right]=\begin{vmatrix}\frac{\partial v}{\partial x} & \frac{\partial w}{\partial x}\\ \frac{\partial v}{\partial y} & \frac{\partial w}{\partial y}\end{vmatrix}=\begin{vmatrix}y & -y\\ x & 1-x\end{vmatrix}=y\left(1-x\right)+xy=y$$

The joint density of *V* and *W* is

$$\frac{\left(\frac{1}{2}\right)^{k}}{\Gamma\left(\frac{k}{2}\right)} v^{\frac{k}{2}-1} e^{-\frac{v}{2}} 1_{0\le v<\infty} \frac{\left(\frac{1}{2}\right)^{m}}{\Gamma\left(\frac{m}{2}\right)} w^{\frac{m}{2}-1} e^{-\frac{w}{2}} 1_{0\le w<\infty}$$

$$=\frac{\left(\frac{1}{2}\right)^{k}\left(\frac{1}{2}\right)^{m}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} v^{\frac{k}{2}-1} w^{\frac{m}{2}-1} e^{-\frac{(v+w)}{2}} 1_{0\le v<\infty} 1_{0\le w<\infty},$$

We can now write

$$f_{X,Y}(x,y)=f_{V,W}(x,y)J\left[(v,w)\rightarrow(x,y)\right]\left[\Omega(v,w)\rightarrow\Omega(x,y)\right]$$

$$=\frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} y^{\frac{k}{2}-1} y^{\frac{m}{2}-1} (1-x)^{\frac{m}{2}-1} e^{-\frac{(y)}{2}} y 1_{0\le x\le 1} 1_{0<y<\infty}$$

$$=\frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} (1-x)^{\frac{m}{2}-1} y^{\frac{k}{2}+\frac{m}{2}-1} e^{-\frac{(y)}{2}} 1_{0\le x\le 1} 1_{0<y<\infty}$$

We can now proceed with our plan to integrate over $y$ to find

$$f_X(x)=\int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} (1-x)^{\frac{m}{2}-1} y^{\frac{k+m}{2}-1} e^{-\frac{y}{2}} 1_{0\le x\le 1}$$

$$=\frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} (1-x)^{\frac{m}{2}-1} 1_{0\le x\le 1} \int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{k+m}{2}}}{\Gamma\left(\frac{k+m}{2}\right)} y^{\frac{k+m}{2}-1} e^{-\frac{(y)}{2}}$$

$$=\frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} (1-x)^{\frac{m}{2}-1} 1_{0\le x\le 1}$$

Which is the measuring tool for the beta distribution.

In order to find the probability density function for the random variable variables $F=\frac{V/k}{W/m}$, we will find the joint distribution for $X=\frac{V}{W}, Y=W$, planning to integrate over the entire range of the random variable $Y$ to obtain the measuring tool of $X$, Once we have that, we can define $F = \frac{k}{m}X.$ Thus, our goal is to transform $(V,W)$ to $\left(X,Y\right)$ and then integrate out *Y*. Using $f_{X,Y}(x,y) = f_{V,W}\left(x,y\right)J\left[\left(v,w\right)\to(x,y)\right]\left[\Omega\left(v,w\right)\to\Omega(x,y)\right],$

we know that $0 \le x \le \infty, 0 \le y \le \infty,$ and $W = Y$ and $V = XY.$ In addition

$$J\left[\left(v,w\right)\to(x,y)\right] = \begin{vmatrix} \frac{\partial v}{\partial x} & \frac{\partial w}{\partial x} \\ \frac{\partial v}{\partial y} & \frac{\partial w}{\partial y} \end{vmatrix} = \begin{vmatrix} y & 0 \\ x & 1 \end{vmatrix} = y.$$

Thus

$$f_{X,Y}(x,y) = f_{V,W}\left(x,y\right)J\left[\left(v,w\right)\to(x,y)\right]\left[\Omega\left(v,w\right)\to\Omega(x,y)\right]$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}(xy)^{\frac{k}{2}-1}y^{\frac{m}{2}-1}e^{-\frac{y(x+1)}{2}}y1_{0\le x\le\infty}1_{0<y<\infty}$$

Continuing,

$$= \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}x^{\frac{k}{2}-1}y^{\frac{m+k}{2}-1}e^{-\frac{y(x+1)}{2}}1_{0\le x\le\infty}1_{0<y<\infty}$$

We now must integrate out *y*.

$$f_X(x) = \int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}x^{\frac{k}{2}-1}y^{\frac{m+k}{2}-1}e^{-\frac{y(x+1)}{2}}1_{0\le x\le\infty}$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}x^{\frac{k}{2}-1}1_{0\le x\le\infty}\int_0^\infty y^{\frac{m+k}{2}-1}e^{-\frac{y(x+1)}{2}}dy.$$

To integrate $\int_0^\infty y^{\frac{m+k}{2}-1}e^{-\frac{y(x+1)}{2}}dy,$ we let $u = y\frac{x+1}{2}, y = \frac{2u}{x+1}, dy = \frac{2}{x+1}du$ and no change in the region of integration. Thus

$$\int_0^\infty y^{\frac{m+k}{2}-1}e^{-\frac{y(x+1)}{2}}dy = \int_0^\infty \left(\frac{2u}{x+1}\right)^{\frac{m+k}{2}-1}e^{-u}\frac{2}{x+1}du$$

$$= \left(\frac{2}{x+1}\right)^{\frac{m+k}{2}}\int_0^\infty u^{\frac{m+k}{2}-1}e^{-u}du$$

$$= \left(\frac{2}{x+1}\right)^{\frac{m+k}{2}}\Gamma\left(\frac{m+k}{2}\right)$$

Thus

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} 1_{0\le x\le\infty} \left(\frac{2}{x+1}\right)^{\frac{m+k}{2}} \Gamma\left(\frac{m+k}{2}\right)$$

$$= \frac{\Gamma\left(\frac{m+k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{k}{2}-1} \left(\frac{1}{x+1}\right)^{\frac{m+k}{2}} 1_{0\le x\le\infty}$$

To conclude, we set $g = \frac{k}{m}x,\ x = \frac{m}{k}g,\ dx = \frac{m}{k}dg,$ note that there is no change in the region of integration and write

$$f_G(g) = \frac{\Gamma\left(\frac{m+k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{m}{k}g\right)^{\frac{k}{2}-1} \left(\frac{1}{\frac{m}{k}g+1}\right)^{\frac{m+k}{2}} \frac{m}{k} 1_{0\le x\le\infty}$$

$$= \frac{\Gamma\left(\frac{m+k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{m}{k}\right)^{\frac{k}{2}} g^{\frac{k}{2}-1} \left(\frac{1}{\frac{m}{k}g+1}\right)^{\frac{m+k}{2}} 1_{0\le x\le\infty}$$

## Deriving the T-distribution

Let $t = \sqrt{g}$, where $k = 1$. Here we have a change in the region of integration, where now $-\infty \le t \le \infty$, and a one to two mapping. We proceed, by noting $g = t^2,\ dg = 2t\,dt,$ and

$$f_T(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{\pi}} m^{\frac{1}{2}} \left(t^2\right)^{-\frac{1}{2}} \left(\frac{1}{mt^2+1}\right)^{\frac{m+1}{2}} 2t\frac{1}{2} 1_{-\infty\le t\le\infty}$$

$$= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{\pi}} m^{\frac{1}{2}} \left(\frac{1}{mt^2+1}\right)^{\frac{m+1}{2}} 1_{-\infty\le t\le\infty}$$

Cauchy, Laplace, and Double Exponential
Ordering Random Variables
Normal Measure
Compounding
F and T Measure
Asymptotics
Tail Event Measure

# Cauchy and Laplace Distributions

Three distributions of continuous measure that are not used as heavily as the normal, or gamma families are the Cauchy, Laplace, and double exponential distributions. Each has revealing properties and uses that reveal something interesting about the nature of probability and the use of density functions.

## Prerequisites

Basic Properties of Probability
Moment and Probability Generating Functions
Continuous Probability Measure
Variable Transformations
An Introduction to the Concept of Measure
Working with Measure
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution

## Cauchy distribution

The Cauchy distribution, named for Augustin-Louis Cauchy is one of the most interesting distributions considered in probability theory. As is the standard normal distribution, the standard Cauchy is symmetric with a median or 50th percentile of zero. Its measuring tool is

$$f_X(x) = \frac{1}{\pi\left(1+x^2\right)} 1_{-\infty<x<\infty}$$

And probabilities over the set are found integrating this function recognizing that $\int \frac{1}{1+x^2}\,dx = \arctan\left(x\right)$. Thus probabilities are closed form and as straightforward to find as those of, for example, the exponential distribution.

A graph of this probability density function demonstrates that this distribution is symmetric around $x=0$ (Figure 1).

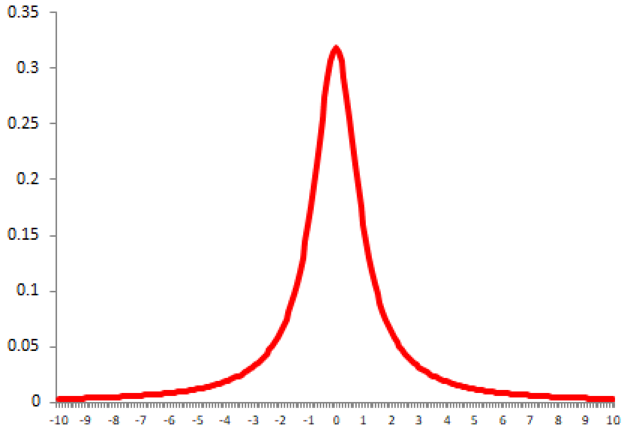


**Figure 1**. The standard Cauchy distribution.

However, one feature that is not discernable from Figure 1 is that while the Cauchy distribution has a median (which is zero), it has no mean. This is a direct reflection of the fact that $\mathbf{E}[X]=\int_{\Omega_x} x\,d\mathbf{P}=\int_{-\infty}^{\infty}\frac{x}{\pi\left(1+x^2\right)}dx=\infty.$ This follows as

$$\int_{-\infty}^{\infty}\frac{x}{\pi\left(1+x^2\right)}dx=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{2x}{\left(1+x^2\right)}dx=\frac{1}{2\pi}\ln\left[\left(1+x^2\right)\right]_{-\infty}^{\infty}$$

$$=\frac{1}{2\pi}\left[\lim_{x\to\infty}\ln\left(1+x^2\right)-\lim_{x\to-\infty}\ln\left(1+x^2\right)\right]$$

and each of these limits diverges. As we might expect, higher order moments are also nonextant.

This behavior is unusual, and would not be expected from the graph. However, the comparison with the standard normal distribution provides an interesting contrast (Figure 2).

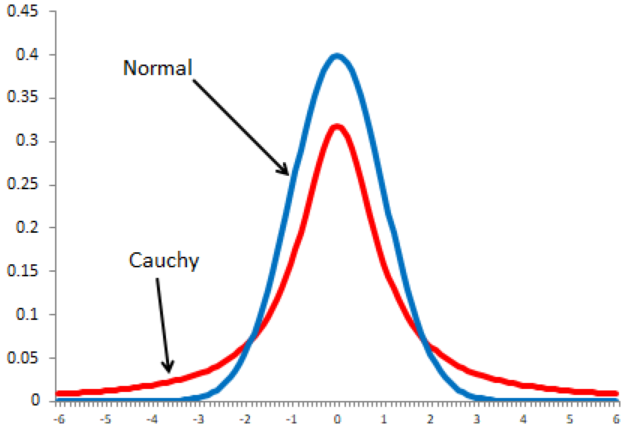


**Figure 2**. Comparison of the standard Cauchy and standard normal measuring tools.

We see that while both distributions are symmetric, the Cauchy distribution is more dispersed at the extremes. These "fatter tails" are precisely the feature that denies the Cauchy distribution its finite mean. How this occurs is revealed in a comparison of the curves that must be integrated in order to compute the mean. For this demonstration, we focus on the positive real number line (Figure 3).

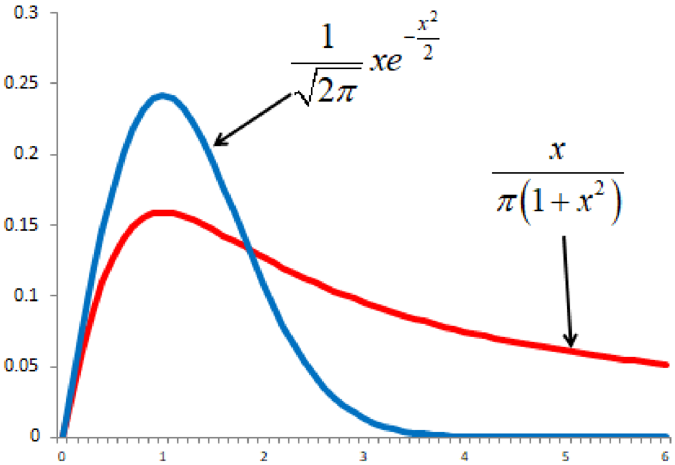


**Figure 3.** The normal "mean" function converges to zero faster than the Cauchy.

The function that we must integrate to find the mean of normal measure, $\frac{1}{\sqrt{2\pi}}xe^{-\frac{x^2}{2}}$, an expression dominated by $e^{-\frac{x^2}{2}}$ rapidly approaches zero. However, the term that we must integrate to find the mean of the Cauchy distribution $\frac{x}{\pi(1+x^2)}$, while still approaching zero, does so at a substantially slower rate. In fact, the rate is so slow, that its measure of the real line is infinity.

## Laplace distribution

The Laplace distribution is related to the negative exponential distribution. However, while the exponential distribution defined as $f_X(x) = \lambda e^{-\lambda x} 1_{0 \le x < \infty}$ is defined for the nonnegative reals, a random variable that follows the Laplace distribution takes positive measure over the entire real line. A random variable $X$ that follows the Laplace distribution has the measuring tool

$$f_X(x) = \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} 1_{-\infty < x < \infty}.$$

From what we know of the exponential distribution, we expect that the parameter is a location parameter, identifying where the center of the measuring tool resides. The parameter $\beta$ functions as a scale parameter, controlling the dispersal of the distribution over the real line (Figure 4).

Working with this distribution poses no obstacle as long as we are facile with absolute values. For $\mu = 0$ and $\beta = 1$ this reduces to $f_X(x) = \frac{1}{2}e^{-|x|}1_{-\infty<x<\infty}$, which is commonly referred to as the double exponential distribution. In the case of the double exponential we can show fairly easily that the measure over the entire real number line is one:

$$\int_{\Omega_x} d\mathbf{P} = \int_{-\infty}^{\infty}\frac{1}{2}e^{-|x|}dx = \frac{1}{2}\left[\int_{-\infty}^{0} e^{-|x|}dx + \int_{0}^{\infty} e^{-|x|}dx\right]$$

$$= \frac{1}{2}\left[\int_{-\infty}^{0} e^{x}dx + \int_{0}^{\infty} e^{-x}dx\right].$$

Recognizing that $\int_{-\infty}^{0} e^{x}dx = \int_{0}^{\infty} e^{-x}dx,$ we can continue

$$\int_{\Omega_x} d\mathbf{P} = \frac{1}{2}\left[\int_{-\infty}^{0} e^{x}dx + \int_{0}^{\infty} e^{-x}dx\right] = \frac{1}{2}2\int_{0}^{\infty} e^{-x}dx = 1.$$

For other values of $\beta$ and $\mu$ we can show that $\mathbf{E}\left[X\right] = \mu$ and $\mathbf{Var}\left[X\right] = \beta^2$. The moment generating function is also available and is $\mathbf{M}_X\left(t\right) = \frac{1}{1-t^2}$.

## Determinant of a normal matrix

And interesting use of the concept of double expectation can be seen from the following question. Let each of the elements of a two by two matrix be independent and follow a standard normal distribution. What is the probability density function of the determinant?

We can write this determinant as the function of four i.i.d, standard normal distributions $X, Y, V, W$. We will write the determinant as

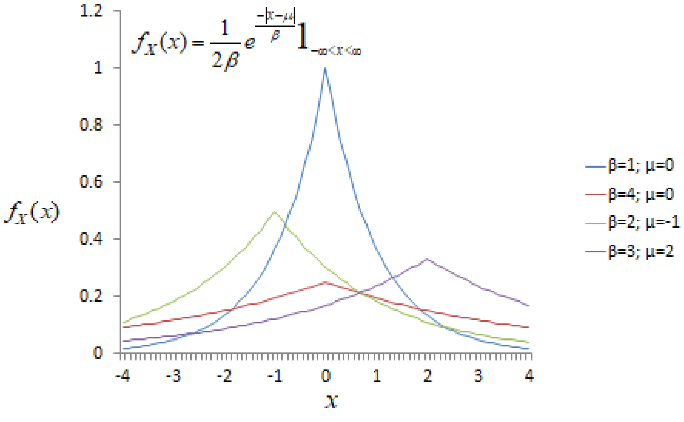


**Figure 4.** The Laplace distribution.

$$U = XY - VW.$$

Certainly $XY$ and $VW$ are themselves independent and identically distributed.

We will first identify the moment generating function of the product of the random variables $XY$. We will use the principle of double expectation to find the moment generating function of this product.

$$\mathbf{E}\left[e^{xyt}\right]=\mathbf{E}_X\left[\mathbf{E}_Y\left[e^{xyt}\right]\right].$$

We will find the inner expectation first.

$$\mathbf{E}_Y\left[e^{xyt}\right]=\mathbf{E}_Y\left[e^{(xt)y}\right]=e^{\frac{t^2x^2}{2}},$$

Since we know the moment generating function of a standard normal distribution. We now compute

$$\mathbf{E}_X\left[\mathbf{E}_Y\left[e^{xyt}\right]\right]=\mathbf{E}_X\left[e^{\frac{t^2x^2}{2}}\right]=\int_{-\infty}^{\infty}e^{\frac{t^2x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2\left(1-t^2\right)}{2}}dx$$

$$=\sqrt{\frac{1}{\left(1-t^2\right)}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\frac{1}{\left(1-t^2\right)}}}e^{-\frac{x^2}{2\frac{1}{\left(t^2-1\right)}}}dx$$

$$=\sqrt{\frac{1}{\left(1-t^2\right)}}$$

Thus, the moment generating function of the product $XY$ is $\sqrt{\frac{1}{\left(1-t^2\right)}}$.

The moment generating function of $U=XY-VW$, is

$$\mathbf{E}_U\left[e^{tu}\right]=\mathbf{E}\left[e^{t((xy)-(uv))}\right]=\mathbf{E}\left[e^{txy-tuv}\right]=\mathbf{E}\left[e^{txy}\right]\mathbf{E}\left[e^{(-t)uv}\right]$$

$$=\sqrt{\frac{1}{\left(1-t^2\right)}}\sqrt{\frac{1}{\left(1-t^2\right)}}\;=\frac{1}{1-t^2}.$$

The determinant follows a double exponential distribution.

Compounding

Asymptotics
Tail Event Measure

# Features of the Laplace Distribution

The Laplace distribution has measuring tool

$$f_X(x) = \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} 1_{-\infty<x<\infty}.$$

## Working with the Laplace distribution

In order to demonstrate $\int_{\Omega_x} d\mathbf{P} = 1$ by breaking the integral into two mutually exclusive and exhaustive ranges based on the absolute value of $|x-\mu|$. Begin by writing

$$\int_{\Omega_x} d\mathbf{P} = \int_{-\infty}^{\infty} \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} dx = \int_{-\infty}^{\mu} \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} dx \quad + \quad \int_{\mu}^{\infty} \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} dx.$$

Let’s take the second integral first.

$$\int_{\mu}^{\infty} \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} dx = \frac{1}{2}\int_{\mu}^{\infty} \frac{1}{\beta} e^{\frac{-(x-\mu)}{\beta}} dx.$$

$y = x - \mu.$ Then $\mu < x < \infty$ implies $0 \le y < \infty,$ $dx = dy,$ and

$$\frac{1}{2}\int_{\mu}^{\infty} \frac{1}{\beta} e^{\frac{-(x-\mu)}{\beta}} dx = \frac{1}{2}\int_{0}^{\infty} \frac{1}{\beta} e^{\frac{-y}{\beta}} dy = \frac{1}{2}.$$

The first integral $\int\limits_{-\infty}^{\mu} \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} dx = \frac{1}{2}\int\limits_{-\infty}^{\mu} \frac{1}{\beta} e^{\frac{(x-\mu)}{\beta}} dx.$ Carry out the same transformation to see

$\frac{1}{2}\int\limits_{-\infty}^{\mu} \frac{1}{\beta} e^{\frac{(x-\mu)}{\beta}} dx = \frac{1}{2}\int\limits_{-\infty}^{0} \frac{1}{\beta} e^{\frac{y}{\beta}} dx,$ then let $w = -y$ to compute $\frac{1}{2}\int\limits_{-\infty}^{0} \frac{1}{\beta} e^{\frac{y}{\beta}} dx = \frac{1}{2}\int\limits_{0}^{\infty} \frac{1}{\beta} e^{-\frac{y}{\beta}} dx = \frac{1}{2},$

and so $\int\limits_{\Omega_x} d\mathbf{P} = \frac{1}{2} + \frac{1}{2} = 1.$

## Moments of the Laplace distribution

We can take advantage of this symmetry in the Laplace distribution to find the higher order moments. For *k* a non-negative integer, we can find $\mathbf{E}\left[X^k\right]$ as

$$\mathbf{E}\left[X^k\right] = \int X^k d\mathbf{P} = \int X^k \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} 1_{-\infty<x<\infty} = \int\limits_{-\infty}^{\infty} X^k \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} dx$$

$$= \int\limits_{-\infty}^{\mu} X^k \frac{1}{2\beta} e^{\frac{-(x-\mu)}{\beta}} dx \quad + \int\limits_{\mu}^{\infty} X^k \frac{1}{2\beta} e^{\frac{-(x-\mu)}{\beta}} dx$$

First, evaluate $\int\limits_{\mu}^{\infty} X^k \frac{1}{2\beta} e^{\frac{-(x-\mu)}{\beta}} dx.$ We allow $y = x - \mu.$ Then $\mu < x < \infty$ implies $0 \le y < \infty,\ dx = dy,$ to see

$\int\limits_{u}^{\infty} x^k \frac{1}{\beta} e^{\frac{-(x-\mu)}{\beta}} dx = \int\limits_{0}^{\infty} \left(y+\mu\right)^k \frac{1}{\beta} e^{\frac{-y}{\beta}} dx.$ So, now we invoke the binomial theorem to write

$\left(y+\mu\right)^k = \sum\limits_{m=0}^{k} \binom{k}{m} y^m \mu^{k\text{-}m},$ permitting

$$\int\limits_{0}^{\infty} \left(y+\mu\right)^k \frac{1}{\beta} e^{\frac{-y}{\beta}} dx = \int\limits_{0}^{\infty} \sum_{m=0}^{k} \binom{k}{m} y^m \mu^{k-m} \frac{1}{\beta} e^{\frac{-y}{\beta}} dx$$

$$= \sum_{m=0}^{k} \binom{k}{m} \mu^{k-m} \int\limits_{0}^{\infty} y^m \frac{1}{\beta} e^{\frac{-y}{\beta}} dx.$$

In order to evaluate $\int\limits_{0}^{\infty} y^m \frac{1}{\beta} e^{\frac{-y}{\beta}} dx,$ we simply let $w = \frac{y}{\beta}, y = \beta w, dy = \beta dw$ and

$\int\limits_{0}^{\infty} y^m \frac{1}{\beta} e^{\frac{-y}{\beta}} dx = \int\limits_{0}^{\infty} w^m \beta^m \frac{1}{\beta} e^{-w} \beta dw = \beta^m \Gamma\left(m+1\right) = \beta^m m!.$ So

We may compute

$$\int\limits_{u}^{\infty} x^k \frac{1}{2\beta} e^{\frac{-(x-\mu)}{\beta}} dx = \frac{1}{2}\sum_{m=0}^{k} \binom{k}{m} \mu^{k-m} \int\limits_{0}^{\infty} y^m \frac{1}{\beta} e^{\frac{-y}{\beta}} dx$$

$$= \frac{1}{2}\sum_{m=0}^{k} \binom{k}{m} \mu^{k-m} \beta^m m! \; = \; \frac{\mu^k}{2} \sum_{m=0}^{k} \frac{k!}{\left(k-m\right)!} \left(\frac{\beta}{\mu}\right)^m .$$

Analogously,

$\int_{-\infty}^{\mu} X^k \frac{1}{2\beta} e^{\frac{-(x-\mu)}{\beta}} dx = \int_{-\infty}^{0} (y+\mu)^k \frac{1}{2\beta} e^{\frac{y}{\beta}} dy.$ Now, let $z=-y$ to compute

$$\begin{aligned}
\int_{-\infty}^{0} (y+\mu)^k \frac{1}{2\beta} e^{\frac{y}{\beta}} dy \quad &= \int_{0}^{\infty} (u-z)^k \frac{1}{2\beta} e^{\frac{-z}{\beta}} dy \\
&= \sum_{m=0}^{k} \binom{k}{m} (-1)^{k-m} \mu^m \int_{0}^{\infty} z^{k-m} \frac{1}{2\beta} e^{\frac{-z}{\beta}} dy \\
&= \frac{1}{2} \sum_{m=0}^{k} \binom{k}{m} (-1)^{k-m} \mu^m \beta^{k-m} (k-m)!
\end{aligned}$$

Thus $\mathbf{E}\left[X^k\right] = \frac{\mu^k}{2} \sum_{m=0}^{k} \frac{k!}{(k-m)!} \left(\frac{\beta}{\mu}\right)^m + \frac{\beta^k}{2} \sum_{m=0}^{k} \frac{k!}{m!} (-1)^{k-m} \left(\frac{\mu}{\beta}\right)^m$

For the mean, we compute that for $k=1$, $\mathbf{E}[X] = \frac{\mu}{2} + \frac{\beta}{2} - \frac{\beta}{2} + \frac{\mu}{2} \quad = \mu.$ We can also find

$$\begin{aligned}
\mathbf{E}\left[X^2\right] = &\frac{\mu^2}{2} \left[ \frac{2!}{(2-0)!} \left(\frac{\beta}{\mu}\right)^0 + \frac{2!}{(2-1)!} \left(\frac{\beta}{\mu}\right)^1 + \frac{2!}{(2-2)!} \left(\frac{\beta}{\mu}\right)^2 \right] \\
&+ \frac{\beta^2}{2} \left[ \frac{2!}{0!} (-1)^2 \left(\frac{\mu}{\beta}\right)^0 - \frac{2!}{1!} \left(\frac{\mu}{\beta}\right) + \frac{2!}{2!} \left(\frac{\mu}{\beta}\right)^2 \right]
\end{aligned}$$

Continuing

$$\begin{aligned}
&= \frac{\mu^2}{2} \left[ 1 + 2\frac{\beta}{\mu} + \left(\frac{\beta}{\mu}\right)^2 \right] + \frac{\beta^2}{2} \left[ 1 - 2\frac{\mu}{\beta} + \left(\frac{\mu}{\beta}\right)^2 \right] \\
&= \frac{1}{2} (\mu+\beta)^2 + \frac{1}{2} (\mu-\beta)^2 = \mu^2 + \beta^2.
\end{aligned}$$

Thus $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}^2[X] = \mu^2 + \beta^2 - \mu^2 = \beta^2.$

## Moment generating function

For $\mu = 0, \beta = 1$ we can find the moment generating function for the Laplace distribution. We write

$$
\begin{aligned}
\mathbf{E}\left[e^{tx}\right] &= \int_{\Omega_x} e^{tx}\, d\mathbf{P} = \int_{-\infty}^{\infty} e^{tx}\,\frac{1}{2}e^{-|x|}dx = \frac{1}{2}\int_{-\infty}^{\infty} e^{tx-|x|}dx \\
&= \frac{1}{2}\left[\int_{-\infty}^{0} e^{tx+x}dx + \int_{0}^{\infty} e^{tx-x}dx\right] = \frac{1}{2}\left[\int_{-\infty}^{0} e^{(t+1)x}dx + \int_{0}^{\infty} e^{(t-1)x}dx\right] \\
&= \frac{1}{2}\left[\int_{0}^{\infty} e^{-(t+1)x}dx + \int_{0}^{\infty} e^{-(1-t)x}dx\right] = \frac{1}{2}\left[\frac{1}{t+1} + \frac{1}{1-t}\right] \\
&= \frac{1}{2}\left[\frac{2}{1-t^2}\right] = \frac{1}{1-t^2}.
\end{aligned}
$$

# Compounding

Working in probability commonly involves working with multiple probability density functions. For example, there are circumstances where a random variable $X$ follows a normal distribution with parameter $\mu$ and $\sigma^2$. However, how do we manage computing the probability of events when the "parameter" $\mu$ has its own probability distribution.

This is a set of circumstances that we will be able to manage with the help of the law of total probability.

Prerequisites
An Introduction to the Concept of Measure
Elementary Set Theory
Measure Based Integration
Lebesgue Integration Theory and the Bernoulli Distribution
Working with Measure
Properties of Probability
Conditional Probability
Bernoulli Distribution – In Depth Discussion
Moment and Probability Generating Functions
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Continuous Probability Measure
Variable Transformations
Uniform and Beta Measure
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Normal Measure

## Restatement of the law

The Law of Total Probability is quite simple. It only states that, if we have the joint probability of two random variables $X$ and W, then we can find the probability of $X$ by summing over the probabilities of $W$. Thus.

$$\boldsymbol{P}\left[X=x\right]=\sum_{W}\boldsymbol{P}\left[X=x,\,W=w\right]:\ \boldsymbol{P}\left[W=w\right]=\sum_{X}\boldsymbol{P}\left[X=x,\,W=w\right],$$

or, more generally,

$$f_X(x)=\int_{\Omega_Y} f_{X,Y}\left(x,y\right)dy:\ \ f_Y(y)=\int_{\Omega_X} f_{X,Y}\left(x,y\right)dx$$

This result is self-evident for independent random variables, since

$$\begin{aligned}f_X(x)&=\int_{\Omega_y} f_{X,Y}\left(x,y\right)dy=\int_{\Omega_y} f_X\left(x\right)f_Y\left(y\right)dy\\&=f_X\left(x\right)\int_{\Omega_y} f_Y\left(y\right)dy=f_X\left(x\right).\end{aligned}$$

While this is useful, it also holds for dependent random variables as well. In this, commonly, it is more helpful to write the joint density in the form of a conditional distribution. Here, we write

$$f_{X|Y}\left(x,y\right)=\frac{f_{X,Y}\left(x,y\right)}{f_Y\left(y\right)}$$

or

$$f_{X,Y}\left(x,y\right)=f_{X|Y}\left(x,y\right)f_Y\left(y\right)$$

Thus, we can rewrite the law of total probability as

$$f_X(x)=\int_{\Omega_y} f_{X,Y}\left(x,y\right)dy=\int_{\Omega_y} f_{X|Y}\left(x,y\right)f_Y\left(y\right)dy.$$

Of course, it is also true that,

$$f_Y(y)=\int_{\Omega_x} f_{X,Y}\left(x,y\right)dy=\int_{\Omega_x} f_{Y|X}\left(x,y\right)f_X\left(x\right)dx.$$

And given our penchant for using the integral sign $\int$ as merely a statement of our intent to accumulate measure, these formulas will apply to discrete probability distributions as well as continuous ones.

### *Example: Natural killer T-cells*

Natural killer or NK cells are T-cells that have a special role in the immune system. One of the functions that these lymphocytes carry out is immunosurveillance. They survey the receptors on other cells, looking for evidence that these other cells, by either absorbing an unusual compound or the creation of some intracellular disruption, may have been compromised.

If the NK cell sees evidence of this compromised cell (through a particular molecular configuration that appears on the examined cell's membrane), the NK cell, suitable excited, will transmit a chemical signal to the signaling cell to kill itself through a process known as apoptosis.

This is one process by which the body identifies and destroys cancerous cells before they can metastasize or spread to other organ systems.

Assume that the probability that out of $n$ cells expressing this signal the probability that $k$ of them, $0 \leq k \leq n,$ are destroyed follows a binomial distribution with probability of a successful cell destruction $p$ where $0 \leq p \leq 1.$ However, suppose that $n$, the number of expressing cells, follows a Poisson distribution with parameter $\lambda.$ What now is the unconditional probability that $k$ cells are destroyed by the natural killer cell?

Here, we are given the conditional probability, $\mathbf{P}\left[X = k \mid N = n\right].$ We are tasked with finding the unconditional probability $\mathbf{P}\left[X = k\right].$ We turn to the law of probability that permits us to write

$$\mathbf{P}\left[X = k\right] = \int_{\Omega_n} \mathbf{P}\left[X = k, N = n\right] = \int_{\Omega_n} \mathbf{P}\left[X = k \mid N = n\right]\mathbf{P}\left[N = n\right].$$

Begin this computation by writing

$$\mathbf{P}\left[X = k\right] = \int_{\Omega_n} \binom{n}{k} p^k \left(1-p\right)^{n-k} 1_{k \in I(0,n)} \frac{\lambda^n}{n!} e^{-\lambda} 1_{n \in I(k,\infty)}.$$

The expression $n \in I\left(k,\infty\right)$ simply means that $n$ is an integer greater than or equal to $k$. Note that the index function for $n$ has a lower bound of $k$ since the event $\left\{X = k\right\}$ presumes $n$ cannot be less than $k$. What remains before us simplification.

$$\begin{aligned}
\mathbf{P}\left[X = k\right] &= \int_{\Omega_n} \binom{n}{k} p^k \left(1-p\right)^{n-k} 1_{k \in I(0,n)} \frac{\lambda^n}{n!} e^{-\lambda} 1_{n \in I(k,\infty)} \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k \left(1-p\right)^{n-k} 1_{k \in I(0,n)} \frac{\lambda^n}{n!} e^{-\lambda} \\
&= \sum_{n=k}^{\infty} \frac{n!}{k!\left(n-k\right)!} p^k \left(1-p\right)^{n-k} 1_{k \in I(0,n)} \frac{\lambda^n}{n!} e^{-\lambda} \\
&= \frac{1}{k!} e^{-\lambda} 1_{k \in I(0,\infty)} \sum_{n=k}^{\infty} \frac{p^k \left(1-p\right)^{n-k}}{\left(n-k\right)!} \lambda^n \\
&= \frac{1}{k!} e^{-\lambda} 1_{k \in I(0,\infty)} \sum_{n=k}^{\infty} \frac{\left(\lambda p\right)^k \left(\lambda\left(1-p\right)\right)^{n-k}}{\left(n-k\right)!} \\
&= \frac{\left(\lambda p\right)^k}{k!} e^{-\lambda} 1_{k \in I(0,\infty)} \sum_{n=k}^{\infty} \frac{\left(\lambda\left(1-p\right)\right)^{n-k}}{\left(n-k\right)!}.
\end{aligned}$$

We now only have to let $m = n - k,$ to helpfully adjust the index governing the range of the summand

$$\begin{aligned}\mathbf{P}[X=k]&=\frac{(\lambda p)^k}{k!}e^{-\lambda}1_{k\in I(0,\infty)}\sum_{m=0}^{\infty}\frac{(\lambda(1-p))^m}{m!}\\&=\frac{(\lambda p)^k}{k!}e^{-\lambda}1_{k\in I(0,\infty)}e^{\lambda(1-p)}\\&=\frac{(\lambda p)^k}{k!}e^{-\lambda p}1_{k\in I(0,\infty)}.\end{aligned}$$

The unconditional distribution of the random variable $X$ is Poisson with parameter $\lambda p$, a parameter whose components are taken from the unconditional Poisson distribution for the number of signaling cells and the conditional binomial distribution, respectively.

This process of finding unconditional distributions based on conditional distributions is historically known as *compounding*.

We can use this to provide more realism to a problem developed in the basic introduction to the binomial distribution involving hurricanes. There we compute the probability of in at least five of ten years, there were at least three hurricanes per year. In order to compute this with the tools that we had at the time, we had to assume that there were fifteen storms per year.

However, now we can simply assume that the number of storms $n$ follows a Poisson distribution with parameter $\lambda$. Let's assume that $\lambda = 15$. Then the probability there are $k$ hurricanes in that year is binomial with parameters $n$ and $p = 0.28$. We now know that the distribution of the number hurricanes in a given year is Poisson with parameter $(15)(0.28) = 4.2$. The probability of at least three hurricanes in a given year is $\sum_{k=3}^{\infty}\frac{4.2^k}{k!}e^{-4.2} = 0.790$, and the probability that in ten years there are at least five with more than three hurricanes each is $\sum_{k=5}^{10}\binom{10}{k}(0.790)^k(0.210)^{n-k} = 0.992$.

## Binomial- negative exponential compounding

Suppose we are working with a death process. Consider an outcome $X$ that follows a binomial distribution where the death parameter $\mu$ is a constant. Then the probability that there are $k$ deaths in the system by time $t$ given there are $n$ subjects in the system at time 0 is

$$\boldsymbol{P}[X=k\mid\mu]=\binom{n}{k}e^{-k\mu t}\left(1-e^{-\mu t}\right)^{n-k}.$$

However, assume that $\mu$ is not constant, but follows an exponential distribution with parameter $\lambda$. We are interested in computing the unconditional distribution of $X$.

Using the law of total probability, we write

$$\boldsymbol{P}[X=k]=\int_{\Omega_\upsilon}\boldsymbol{P}[X\cap\mu]=\int_{\Omega_\upsilon}\boldsymbol{P}[X=k\mid\mu]\boldsymbol{P}[\mu].$$

and substitute the binomial distribution for $\boldsymbol{P}[X=k\mid\mu]$, and the exponential distribution with parameter $\lambda$ for $\mathbf{P}[\mu]$.

$$\begin{aligned}\boldsymbol{P}\left[X=k\right] &= \int_{\Omega_\upsilon} \boldsymbol{P}\left[X=k \mid \mu\right]\boldsymbol{P}\left[\mu\right] \\ &= \int_{\Omega_\upsilon} \binom{n}{k} e^{-k\mu t}\left(1-e^{-\mu t}\right)^{n-k} \lambda e^{-\lambda\mu} d\mu\end{aligned}$$

Since $n-k$ is a non-negative integer, we may write

$$\left(1-e^{-\mu t}\right)^{n-k} = \sum_{i=0}^{n-k} \binom{n-k}{i} \left(-1\right)^i e^{-i\mu t}.$$

Substituting this expression, we have

$$\begin{aligned}\boldsymbol{P}\left[X=k\right] &= \int_{\upsilon} \binom{n}{k} e^{-k\mu t} \sum_{i=0}^{n-k} \binom{n-k}{i} \left(-1\right)^i e^{-i\mu t} \lambda e^{-\lambda\mu} d\mu \\ &= \binom{n}{k} \sum_{\boldsymbol{i=0}}^{\boldsymbol{n-k}} \binom{n-k}{i} \left(-1\right)^i \int_{\upsilon} \lambda e^{-\left(\left(k+i\right)t+\lambda\right)\mu} d\mu\end{aligned}$$

The remaining integral is simply $\int_0^\infty \lambda e^{-\left(\left(k+i\right)t+\lambda\right)\mu} d\mu = \frac{\lambda}{\left(k+i\right)t+\lambda}$.

Thus

$$\boldsymbol{P}\left[X=k\right] = \binom{n}{k} \sum_{\boldsymbol{i=0}}^{\boldsymbol{n-k}} \binom{n-k}{i} \left(-1\right)^i \frac{\lambda}{\left(k+i\right)t+\lambda},$$

which is an easily computable, finite sum.

### *Compounding a Poisson-gamma*

Consider an outcome $X_t$ that represents the number of arrivals to an emergency room through time *t*. If we assume that $X_t$ follows a Poisson distribution with parameter *λ*, we may write

$$\boldsymbol{P}\left[\boldsymbol{X_t} = k \mid \lambda\right] = \frac{\left(\lambda t\right)^k}{k!} e^{-\lambda t}$$

for $k = 0, 1, 2, 3, \ldots$ Thus $X_t$ is the cumulative number of arrivals from time 0 to time *t*. However, let's also assume that the parameter *λ* is not constant, but follows a gamma distribution with parameters *α* and *r*.

$$\boldsymbol{P}\left[\lambda\right] = \frac{\alpha^r}{\Gamma\left(r\right)} \lambda^{r-1} e^{-\alpha\lambda} 1_{0\le\lambda\le\infty}.$$

We are interesting in identifying the unconditional probability distribution of $X_t$. Using the Law of Total Probability, we write

$$\boldsymbol{P}[X_t = k] = \int_{\Omega_\lambda} \boldsymbol{P}[X_t \cap \lambda] = \int_{\Omega_\lambda} \boldsymbol{P}[X_t = k \mid \lambda]\boldsymbol{P}[\lambda]$$
$$= \int_0^\infty \frac{(\lambda t)^k}{k!} e^{\lambda t} \frac{\alpha^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda} d\lambda.$$

Removing all terms not involving the variable λ outside the integral, we can rewrite the second line of expression as

$$\frac{\alpha^r}{\Gamma(r)k!} t^k \int_0^\infty \lambda^k e^{\lambda t} \lambda^{r-1} e^{-\alpha\lambda} d\lambda \;=\; \frac{\alpha^r}{\Gamma(r)k!} t^k \int_0^\infty \lambda^{k+r-1} e^{-(\alpha+t)\lambda} d\lambda.$$

The integral on the right side of this equation needs only a constant to allow it to be one. We therefore write

$$\frac{\alpha^r}{\Gamma(r)k!} t^k \int_0^\infty \lambda^{k+r-1} e^{-(\alpha+t)\lambda} d\lambda$$
$$= \frac{\alpha^r}{\Gamma(r)k!} t^k \frac{\Gamma(k+r)}{(\alpha+t)^{k+r}} \int_0^\infty \frac{(\alpha+t)^{k+r}}{\Gamma(k+r)} \lambda^{k+r-1} e^{-(\alpha+t)\lambda} d\lambda.$$

The integral on the right-hand side of equation is that of a variable that follows a gamma distribution parameters $\alpha + t$ and $k + r$. This integrates to one over the entire range of $\lambda$. Thus, we are left with

$$\boldsymbol{P}[X_t = k] = \frac{\alpha^r}{\Gamma(r)k!} t^k \frac{\Gamma(k+r)}{(\alpha+t)^{k+r}}$$
$$= \frac{\Gamma(k+r)}{\Gamma(r)k!} \frac{\alpha^r}{(\alpha+t)^r} \frac{t^k}{(\alpha+t)^k}$$
$$= \binom{k+r-1}{r-1} \left(\frac{\alpha}{\alpha+t}\right)^r \left(\frac{t}{\alpha+t}\right)^k.$$

We recognize this last expression as the probability of $k$ failures before the $r^{\text{th}}$ success when the probability of a success is $\alpha/(\alpha+t)$, and the probability of failure is $t/(\alpha+t)$.

It is of interest that we began with Poisson measure, and ended with that of the negative binomial. Each provides positive probability over the non-negative integers, and it can sometimes be difficult to differentiate which of these two ubiquitous distributions should be selected.

Some guidance can be provided by the mean and variance of the data collected from the experiment. If they are close to each other, than one might start with a Poisson distribution. However, if one expects from the nature of the experiment that the process should be governed by the Poisson distribution, but the mean and variance are not similar, consideration of the compound Poisson derived above may be warranted.[*]

Consider an outcome $X$ that follows a gamma distribution with parameters $\alpha$ and $n$ where $n$ is an integer. Let $n$ itself follows negative binomial measure. with parameters $r$ and $p$. What is the unconditional distribution of $X$?

Using the Law of Total Probability we write

$$\begin{aligned} f_X(x) &= \int_{\Omega_n} f(x \cap n) = \int_{\Omega_n} f(x \mid n)\mathbf{P}[N = n] \\ &= \sum_{n=r}^{\infty} \frac{\alpha^n}{\Gamma(n)} x^{n-1} e^{-\alpha x} \binom{n-1}{r-1} p^r (1-p)^{n-r} \\ &= e^{-\alpha x} \frac{1}{x} \left(\frac{p}{1-p}\right)^r \sum_{n=r}^{\infty} \frac{\alpha^n}{\Gamma(n)} x^n \binom{n-1}{r-1} (1-p)^n . \end{aligned}$$

Recognizing that $\binom{n-1}{r-1} = \frac{\Gamma(n)}{\Gamma(r)(n-r)!}$ permits cancellation of the $\Gamma(n)$ term, allowing us to write

$$f_X(x) = e^{-\alpha x} \frac{1}{\Gamma(r)} \frac{1}{x} \left(\frac{p}{1-p}\right)^r \sum_{n=r}^{\infty} \frac{\alpha^n}{(n-r)!} x^n (1-p)^n .$$

The summand may be written as

$$\begin{aligned} \sum_{n=r}^{\infty} \frac{\alpha^n}{(n-r)!} x^n (1-p)^n &= \alpha^r x^r (1-p)^r \sum_{n=r}^{\infty} \frac{\left[\alpha x (1-p)\right]^{n-r}}{(n-r)!} \\ &= \alpha^r x^r (1-p)^r e^{\alpha x (1-p)} . \end{aligned}$$

Allowing us to write

$$\begin{aligned} f_X(x) &= e^{-\alpha x} \frac{1}{\Gamma(r)} \frac{1}{x} \left(\frac{p}{1-p}\right)^r \alpha^r x^r (1-p)^r e^{\alpha x (1-p)} \\ &= \frac{(\alpha p)^r}{\Gamma(r)} x^{r-1} e^{-\alpha p x} \end{aligned}$$

which is a gamma density with parameters $\alpha p$ and $r$.

### Gamma-gamma compounding

Consider an outcome $x$ that follows a gamma distribution with parameters $\alpha$ and $m$ where $m$ is an integer greater than zero.

$$f_X(x) = \frac{\lambda^m}{\Gamma(m)} x^{m-1} e^{-\lambda x} 1_{0 \le x \le \infty} .$$

However, here, the parameter $\lambda$ has its own probability distribution

[*] From. Dr. John P. Young, 1973, the Johns Hopkins University.

$$f_\lambda(\lambda) = \frac{\alpha^r}{\Gamma(r)}\lambda^{r-1}e^{-\alpha\lambda}1_{0\le\lambda\le\infty} \ ,$$

where we will assume that *r* is also a positive integer. We are interested in finding the marginal distribution of *x*. Using the Law of Total Probability we write

$$\begin{aligned} f_X(x) &= \int_{\Omega_\lambda} f(x,\lambda) = \int_{\Omega_\lambda} f(x\,|\,\lambda)f(\lambda) \\ &= \int_0^\infty \frac{\lambda^m}{\Gamma(m)}x^{m-1}e^{-\lambda x}\frac{\alpha^r}{\Gamma(r)}\lambda^{r-1}e^{-\alpha\lambda}d\lambda. \end{aligned}$$

Removing terms involving $\lambda$, the second line of expression  becomes

$$f_X(x) = \frac{\alpha^r}{\Gamma(m)\Gamma(r)}x^{m-1}\int_0^\infty \lambda^{r+m-1}e^{-(\alpha+x)\lambda}d\lambda.$$

Recognizing that the integrand in expression  is related to that of a variable that follows a gamma distribution, we include the appropriate constant so that this integral's value is one.

$$\begin{aligned} f_X(x) &= \frac{\alpha^r}{\Gamma(m)\Gamma(r)}x^{m-1}\frac{\Gamma(r+m)}{(\alpha+x)^{r+m}}\int_0^\infty \frac{(\alpha+x)^{r+m}}{\Gamma(r+m)}\lambda^{r+m-1}e^{-(\alpha+x)\lambda}d\lambda \\ &= \frac{\alpha^r}{\Gamma(m)\Gamma(r)}x^{m-1}\frac{\Gamma(r+m)}{(\alpha+x)^{r+m}}. \end{aligned}$$

We arranging terms, this becomes

$$\begin{aligned} f_X(x) &= \frac{\Gamma(r+m)}{\Gamma(m)\Gamma(r)}\frac{\alpha^r}{(\alpha+x)^{r+m}}x^{m-1} \\ &= \frac{(r+m-1)!}{(m-1)!(r-1)!}\frac{\alpha^r}{(\alpha+x)^{r+1}}\left(\frac{x}{\alpha+x}\right)^{m-1} \\ &= \frac{1}{\alpha}\frac{(r+1)r}{r+m}\frac{(r+m)(r+m-1)!}{(m-1)!(r+1)(r)(r-1)!}\left(\frac{\alpha}{\alpha+x}\right)^{r+1}\left(\frac{x}{\alpha+x}\right)^{m-1} \\ &= \frac{(r+1)r}{a(r+m)}\binom{r+m}{r+1}\left(\frac{\alpha}{\alpha+x}\right)^{r+1}\left(\frac{x}{\alpha+x}\right)^{m-1}. \end{aligned}$$

## Compounding two normal distributions

Let *X* be a random variable following a normal distribution with mean $\theta$ and variance $\sigma^2$. However, in this case, $\theta$ is itself normally distributed with mean $\mu$, and variance $\upsilon^2$. We seek the unconditional or marginal distribution of *X*. Using the Law of Total Probability we write

$$f_X(x) = \int_{\Omega_\theta} f\left(x,\theta\right) = \int_{\Omega_\theta} f(x\,|\,\theta) f(\theta)$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\upsilon^2}} e^{-\frac{(\theta-\mu)^2}{2\upsilon^2}} d\theta.$$

Our goal is to carry out the integration in the last line of this expression with respect to $\theta$. This expression can be rewritten as

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\upsilon^2}} e^{-\frac{(\theta-\mu)^2}{2\upsilon^2}} d\theta$$

$$= \frac{1}{2\pi\sigma\upsilon} \int_{-\infty}^{\infty} e^{-\left[\frac{(x-\theta)^2}{2\sigma^2} + \frac{(\theta-\mu)^2}{2\upsilon^2}\right]} d\theta$$

$$= \frac{1}{2\pi\sigma\upsilon} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\upsilon^2}\right]} d\theta \;=\; \frac{1}{2\pi\sigma\upsilon} \int_{-\infty}^{\infty} e^{K(x,\theta)} d\theta.$$

And our attention turns to simplifying the exponent in this integral. The process we will follow is one of completing the square. Begin by

$$K\left(x,\theta\right) = -\frac{1}{2}\left[\frac{\left(x-\theta\right)^2}{\sigma^2} + \frac{\left(\theta-\mu\right)^2}{\upsilon^2}\right]$$

$$= -\frac{1}{2}\left[\frac{x^2 - 2\theta x + \theta^2}{\sigma^2} + \frac{\theta^2 - 2\theta\mu + \mu^2}{\upsilon^2}\right].$$

Continuing

$K\left(x,\theta\right) = -\frac{1}{2}\left(\left[\frac{1}{\sigma^2} + \frac{1}{\upsilon^2}\right]\theta^2 - 2\left(\frac{x}{\sigma^2} + \frac{\mu}{\upsilon^2}\right)\theta + \frac{x^2}{\sigma^2} + \frac{\mu^2}{\upsilon^2}\right)$ and if $H(x) = \frac{x^2}{\sigma^2} + \frac{\mu^2}{\upsilon^2}$, we can write

$$K(x,\theta) = -\frac{1}{2}\left(\left[\frac{\sigma^2+\upsilon^2}{\sigma^2\upsilon^2}\right]\theta^2 - 2\left(\frac{x}{\sigma^2} + \frac{\mu}{\upsilon^2}\right)\theta\right) - \frac{1}{2}H(x).$$

$$= -\frac{\left[\frac{\sigma^2+\upsilon^2}{\sigma^2\upsilon^2}\right]}{2}\left(\theta^2 - 2\frac{1}{\left[\frac{\sigma^2+\upsilon^2}{\sigma^2\upsilon^2}\right]}\left(\frac{x}{\sigma^2} + \frac{\mu}{\upsilon^2}\right)\theta\right) - \frac{1}{2}H(x)$$

$$= -\frac{1}{2\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]}\left(\theta^2 - \frac{2}{\sigma^2+\upsilon^2}\left(\upsilon^2 x + \sigma^2\mu\right)\theta\right) - \frac{1}{2}H(x).$$

It now remains to complete the square in $\theta$.

$$
\begin{aligned}
&\theta^2 - \frac{2}{\sigma^2+\upsilon^2}\left(\upsilon^2 x+\sigma^2\mu\right)\theta \\
&=\theta^2 - \frac{2}{\sigma^2+\upsilon^2}\left(\upsilon^2 x+\sigma^2\mu\right)\theta + \left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2 - \left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2 \\
&=\left(\theta-\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right)^2 - \left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2.
\end{aligned}
$$

We can now incorporate the completed square term.

$$
\begin{aligned}
K(x,\theta) &= \frac{-1}{2\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]}\left\{\left(\theta-\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right)^2 - \left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2\right\} - \frac{1}{2}H(x) \\
&= \frac{-\left(\theta-\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right)^2}{2\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]} + \frac{1}{2}\frac{\left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2}{\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]} - \frac{1}{2}H(x)
\end{aligned}
$$

Continuing

$$
= \frac{-\left(\theta-\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right)^2}{2\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]} - \frac{1}{2}G(x) + \frac{1}{2}H(x).
$$

We can now write

$$
\begin{aligned}
f(x) &= \frac{1}{2\pi\sigma\upsilon}\int_{-\infty}^{\infty} e^{K(x,\theta)}d\theta \\
&= \frac{\sqrt{\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}}}{\sqrt{2\pi}\sigma\upsilon}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}}} e^{-\frac{\left(\theta-\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right)^2}{2\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]}} e^{-\frac{1}{2}\left(-G(x)+H(x)\right)}d\theta \\
&= \frac{\sqrt{\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}}}{\sqrt{2\pi}\sigma\upsilon} e^{-\frac{1}{2}\left(-G(x)+H(x)\right)}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}}} e^{-\frac{\left(\theta-\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right)^2}{2\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]}} d\theta.
\end{aligned}
$$

The integral in the last line of expression is one, and we are left with

$$f(x) = \frac{1}{\sqrt{2\pi\left(\sigma^2+\upsilon^2\right)}} e^{-\frac{1}{2}\left(-G(x)+H(x)\right)}$$

and it remains to simply $\frac{1}{2}\left(-G(x)+H(x)\right)$. We begin by writing

$$\begin{aligned}
-G(x)+H(x) &= -\frac{\left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2}{\left[\frac{\sigma^2\upsilon^2}{\sigma^2+\upsilon^2}\right]} + \frac{x^2}{\sigma^2}+\frac{\mu^2}{\upsilon^2} \\
&= -\left(\frac{\sigma^2+\upsilon^2}{\sigma^2\upsilon^2}\left[\frac{\left(\upsilon^2 x+\sigma^2\mu\right)}{\sigma^2+\upsilon^2}\right]^2\right)+\frac{x^2}{\sigma^2}+\frac{\mu^2}{\upsilon^2} \\
&= \frac{-\left(\upsilon^2 x+\sigma^2\mu\right)^2}{\sigma^2\upsilon^2\left(\sigma^2+\upsilon^2\right)}+\frac{x^2}{\sigma^2}+\frac{\mu^2}{\upsilon^2} \\
&= \frac{-\left(\upsilon^2 x+\sigma^2\mu\right)^2+\upsilon^2\left(\sigma^2+\upsilon^2\right)x^2+\sigma^2\left(\sigma^2+\upsilon^2\right)\mu^2}{\sigma^2\upsilon^2\left(\sigma^2+\upsilon^2\right)} \\
&= \frac{-\upsilon^4x^2-2\sigma^2\upsilon^2\mu x-\sigma^4\mu^2+\upsilon^2\sigma^2x^2+\upsilon^4x^2+\sigma^4\mu^2+\sigma^2\upsilon^2\mu^2}{\sigma^2\upsilon^2\left(\sigma^2+\upsilon^2\right)}.
\end{aligned}$$

Cancellation reveals

$$\begin{aligned}
-G(x)+H(x) &= \frac{-2\sigma^2\upsilon^2\mu x+\upsilon^2\sigma^2x^2+\sigma^2\upsilon^2\mu^2}{\sigma^2\upsilon^2\left(\sigma^2+\upsilon^2\right)} \\
&= \frac{\sigma^2\upsilon^2\left(x^2-2\mu x+\mu^2\right)}{\sigma^2\upsilon^2\left(\sigma^2+\upsilon^2\right)} \\
&= \frac{1}{\left(\sigma^2+\upsilon^2\right)}\left(x-\mu\right)^2.
\end{aligned}$$

We can now write $f_X(x)$ as

$$f_X(x) = \frac{1}{\sqrt{2\pi\left(\sigma^2+\upsilon^2\right)}} e^{-\frac{1}{2\left(\sigma^2+\upsilon^2\right)}\left(x-\mu\right)^2}$$

which is the function of a normal distribution with mean $\mu$ and variance $\sigma^2+\upsilon^2$.

Asymptotics
Tail Event Measure

# Asymptotics

## Prerequisites

The Concept of the Limit
Convergent Series
Cauchy Sequences
Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Moment and Probability Generating Functions
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Continuous Probability Measure
Uniform and Beta Measure
Survival Measure
Normal Measure

One of the most challenging topics in probability for students is the application of limit theory. Yet this theory has a fine motivation and can be mastered if introduced helpfully.

Much of the work in applied statistics deals with drawing a sample from a much larger population. We would like to believe that the result that we obtain from that sample provides insight into the population from which the sample was drawn. Yet, is that insight there? Is that really a principal relationship between the sample and the population? How do we know?

Another way to say that is that as the samples become larger and larger, we would like our estimators or functions of the data to become closer and closer to the actual parameters that can only be known if the entire population could be studied.

For example, we understand that we will never actually know with infinite precision the population parameter $\sigma^2$ from normal measure. However, we can be assured that the estimate

$\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}$ of $\sigma^2$ that is constructed from a sample of data will get closer and closer to $\sigma^2$ as $n$ gets larger. Without this confidence, the estimate is of diminished value to applied scientists.

How can we provide them this assurance?

## Probability and convergence

Asymptotic theory combines probability and the limiting process from calculus. This combination requires new thought, and has produced four types of convergence available to us.

1- Convergence in distribution (in law)
2- Convergence in probability
3- Convergence with probability one
4- Convergence in mean

This list reveals that there is not just one way, but several ways to combine the principals of probability with those of convergence, and the number of them gives students pause. However, like many things in mathematics, this is quite manageable if we take one step at a time.

Let's get started.

## Convergence in distribution

How then might random variables converge? We understand the convergence process from our review of limits, developed by Augustin Cauchy. We begin with $\{X_n\}$, an infinite sequence of random variables, for which each random variable has a cumulative distribution function $\mathbf{F}_{X_n}(x)$.

For convergence in distribution, we require that (and only that) the cdf's of the random variables in the sequence converge to a function that is itself a cdf.

Formally this means that if there is a sequence of random variables $\{X_n\}$ each of which has its own cumulative distribution $\mathbf{F}_{X_n}(x)$, then if $\mathbf{F}_{X_n}(x) \to \mathbf{F}_X(x)$, then the random variables $X_n \to X$, assuming that $\mathbf{F}_X(x)$ is itself a cumulative distribution function.

Convergence in distribution (otherwise known as convergence in law) is simply a statement about the convergence of cumulative distribution functions.

As an example, let's consider a collection of *n* random variables each of which follows the $\mathbf{U}(0,1)$ distribution. We are interested in the measure of its sample minimum $V_n$.

We have worked with $V_n$ in our discussion of the measure of order. Can we apply the notion of convergence in distribution to $V_n$?

Our goal is to identify the cumulative distribution function $\mathbf{F}_n(x)$ and then look for any limiting behavior of these cumulative distribution functions.

But first, let's consider what we might expect. We know that while the absolute minimum of a $\mathbf{U}(0,1)$ must be zero, we would not expect to actually obtain that minimum value from any sample.

Yet, as the sample sizes increase, intervals closer and closer to zero, however small, will be likely to contain observations as well. Thus, we will expect that the minimum of these larger and larger samples "should" become close to zero. Thus, as sample size increases, it becomes more likely that the minimum would become smaller; the limiting cumulative distribution function would reflect this.

Let $\mathbf{F}_{V_n}(v) = \mathbf{P}\left[V_n \leq v\right]$. Recalling our work on order statistics, we find that

$1-\mathbf{F}_{V_n}(v) = \mathbf{P}\left[V_n > v\right] = \left(1-\mathbf{F}_X(v)\right)^n = \left(1-v\right)^n$, and

$\mathbf{F}_{V_n}(v) = \left[1-\left(1-v\right)^n\right]\mathbf{1}_{0\leq v\leq 1}$. We now evaluate

$$\lim_{n\to\infty}\mathbf{F}_{V_n}(v) = \lim_{n\to\infty}\left[1-\left(1-v\right)^n\right]\mathbf{1}_{0\leq v\leq 1} = \lim_{n\to\infty} v^n\mathbf{1}_{0\leq v\leq 1} = \mathbf{F}_V(v).$$

$\mathbf{F}_V(v)$ is the cumulative distribution function of the random variable $\mathbf{P}\left[V=0\right]=1$, and we have demonstrated our result. We conclude that the random variable $V_n$ converges in distribution to zero.

Formally, we say that a sequence of random variables $\left\{X_n\right\}$ with associated cumulative distribution functions $\mathbf{F}_{X_n}\left(x\right)$ converges in distribution (or in law) to the random variable $X$ with a cumulative distribution function identified as $\mathbf{F}_X\left(x\right)$ if $\lim_{n\to\infty}\mathbf{F}_{X_n}\left(x\right) = \mathbf{F}_{\mathbf{X}}\left(x\right)$ for every $x$ at which the cumulative distribution function $\mathbf{F}_X\left(x\right)$ is continuous.

It is important to see how the introduction of probability has altered our definition of convergence. In our introductory discussion of the limiting process, the elements of the sequence $\left\{X_n\right\}$ became closer to one another as $n$ increases. However, in this current setting, the elements of the sequence $\left\{X_n\right\}$ are random — they are unknown. What converges is not the random variable values themselves, but the distributions which govern the values that these random variables can assume.*

This convergence is termed weak convergence. It is weak because it is not the random variables that are converging to a particular value, but the distribution functions that are converging. In weak convergence the occurrence of random variable values quite far from the limiting value denoted by $\mathbf{F}_X\left(x\right)$ although unlikely, are possible.

This is evident in the minimum example of this section. No matter how large the sample, there is a probability $\left(1-\varepsilon\right)^n$ that the minimum observation in that sample will be greater than some positive number $\varepsilon$. Although $\lim_{n\to\infty}\left(1-\varepsilon\right)^n = 0$, there is never a guarantee that all of the minimums will be within $\varepsilon$ of zero, for any large value of $n$.

## Convergence in distribution and $M_n(t)$

We defined the concept of convergence in distribution as one that involves a sequence of random variables $\left\{X_n\right\}$ and a corresponding sequence of cumulative distribution functions, $\left\{\mathbf{F}_{X_n}(x)\right\}$ which, for the sake of notational brevity, we will describe as $\left\{\mathbf{F}_n(x)\right\}$ We said that the sequence of random variables $\left\{X_n\right\}$ converged in distribution to the random variable $X$ if $\lim_{n\to\infty}\mathbf{F}_n(x) = \mathbf{F}(x)$.

* The importance of $\lim_{n\to\infty}\mathbf{F}_{X_n}\left(x\right) = \mathbf{F}_{\mathrm{X}}(x)$ for only those points $x$ for which the cumulative distribution function is continuous is discussed in detail elsewhere(Rao, 1984).

However, we might also use the link between moment generating functions and cumulative distribution functions to discuss convergence in distribution in terms of the MFG's.

Recall that a cumulative distribution function has one and only one moment generating function. Therefore, since the moment generating functions are unique, the convergence of $\mathbf{F}_n(x)$ to $\mathbf{F}(x)$ suggests that if $\mathbf{M}_n(t)$ is the moment generating function of the random variable $X_n$, then $\mathbf{M}_n(t)$ must converge to $\mathbf{M}(t)$ where $\mathbf{M}_X(t) = \mathbf{E}\left[e^{tx}\right]$. This is called Lévy's Continuity Theorem for Moment Generating Functions, which is present without proof.

## Lévy's continuity theorem

Let $\{X_n\}$ be a sequence of random variables. Denote the cumulative distribution function of $X_n$ as $\mathbf{F}_n(t)$, and the moment generating function of $X_n$ by $\mathbf{M}_n(t)$. Then for every point *t* such that $|t| < 1$ and $\lim_{t\to 0}\mathbf{M}(t) = 1$, then $\lim_{n\to\infty}\mathbf{F}_n(x) = \mathbf{F}(x)$ *iff* $\lim_{n\to\infty}\mathbf{M}_n(t) = \mathbf{M}(t)$.

## MGF's and convergence in distribution

The Lévy Continuity Theorem is the basis of the argument for the use of moment generating functions as a vehicle to obtain convergence in distribution. We will find that, in general, the moment generating function is a useful tool in demonstrating the convergence of random variables. However it must be recognized, that working with moment generating functions can at first be frustrating.

For example, the examination of the manner in which the sum of uniformly distributed random variables on the unit interval produced probability mass in the center of the resulting distribution provided intuition into why central tendency emerges from sums of variables which by themselves presented no such tendency.

However, frequently there is no such insight when one works with moment generating functions. This is not to suggest that arguments based solely on moment generating functions are impenetrable — only that the intuition for the underlying argument can sometimes be lost in the deeper consideration of limits and continuity.

We will find in our approach to moment generating functions and asymptotics that a useful tool on which moment generating functions commonly rely is the following easily demonstrated limit statement;

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$$

### *Poisson random variables and normal measure*

We have already seen how the binomial distribution converges to the Poisson distribution using probability generating functions. Out study of the use of normal measure as an approximation to the measure of other non-normal random variables begins with an examination of the manner in which the Poisson distribution produces probabilities that are approximately Gaussian.

We rely on moment generating functions in this demonstration. Recall that,

$$M_X(t) = E\left[e^{tx}\right] = \mathbf{E}\left[\sum_{k=0}^{\infty} \frac{t^k X^k}{k!}\right] = \mathbf{E}\left[\sum_{k=0}^{\infty} X^k \frac{t^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{\mathbf{E}\left[X^k\right]}{k!} t^k$$

$$= 1 + \frac{\mathbf{E}\left[X\right]}{1} t + \frac{\mathbf{E}\left[X^2\right]}{2} t^2 + \frac{\mathbf{E}\left[X^3\right]}{3!} t^3 + \ldots + \frac{\mathbf{E}\left[X^k\right]}{k!} t^k + o\left(t^k\right)$$

where $o\left(t^k\right)$ reflects additional terms that contain powers of *t* that are greater than *k* (i.e., terms that are of order $t^{k+1}$ or greater). A careful analysis of the expansion of the moment generating function reveals that $\lim_{t \to 0} o\left(t^k\right) = 0$.

Our goal is to demonstrate the utility of this expansion in proving that the probability distribution of a random variable converges to the probability distribution of another. Specifically, we will show that the expansion of the moment generating function of the Poisson distribution, will converge to that of normal measure. Recall that if *X* is a Poisson random variable with mean λ, then

$$\mathbf{M}_X\left(t\right) = e^{\lambda\left(e^t - 1\right)}$$

Now consider the random variable *Z* defined by $Z = \frac{X - \lambda}{\sqrt{\lambda}}$, which is the Poisson random variable with its mean subtracted and divided by its standard deviation (i.e., a "standardized" Poisson random variable). Then the mgf of *Z* is

$$\mathbf{M}_Z(t) = \mathbf{E}\left[e^{\frac{(X-\lambda)t}{\sqrt{\lambda}}}\right] = e^{-t\sqrt{\lambda}}\mathbf{E}\left[e^{\frac{t}{\sqrt{\lambda}}X}\right].$$

Since, $\mathbf{M}_X\left(t\right) = e^{\lambda\left(e^t - 1\right)}$ we may write $\mathbf{E}\left[e^{\frac{t}{\sqrt{\lambda}}X}\right] = e^{\lambda\left(e^{\frac{t}{\sqrt{\lambda}}} - 1\right)}$, and continue as

$\mathbf{M}_Z\left(t\right) = e^{-t\sqrt{\lambda}}\left[e^{\lambda\left(e^{t/\sqrt{\lambda}} - 1\right)}\right]$. By expanding $e^{\frac{t}{\sqrt{\lambda}}}$ we can write

$$\lambda\left(e^{\frac{t}{\sqrt{\lambda}}} - 1\right) = \lambda\left[1 + \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2!\lambda} + \frac{t^3}{3!\lambda^{3/2}} + \ldots + \frac{t^n}{n!\lambda^{n/2}} + \ldots - 1\right]$$

$$= \lambda\left[1 + \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2!\lambda} + o(t) - 1\right] = \lambda\left[\frac{t}{\sqrt{\lambda}} + \frac{t^2}{2!\lambda} + o(t)\right]$$

Thus

$$\mathbf{M}_Z\left(t\right) = \mathbf{E}\left[e^{\frac{(X-\lambda)t}{\sqrt{\lambda}}}\right] = e^{-t\sqrt{\lambda}}\left[e^{\lambda\left(e^{t/\sqrt{\lambda}} - 1\right)}\right] = e^{-t\sqrt{\lambda}}\left[e^{t\sqrt{\lambda} + \frac{1}{2}t^2 + o(t)}\right] = e^{\frac{1}{2}t^2 + o(t)}$$

And $\lim_{\lambda \to \infty} \mathbf{M}_Z\left(t\right) = e^{\frac{1}{2}t^2}$.

Since this is the moment generating function of the standard normal distribution, we have convergence in law of the Poisson distribution to normal measure. Another demonstration of this result is also available.

This finding does not imply that Poisson random variables become normal. This is not true for many reasons, beginning with that the measure of the Poisson random variables is only positive on the nonnegative integers, while normal measure is positive over the entire real number line.

The appropriate conclusion is that, suitably normalized, normal measure can be used to approximate Poisson measure.

## Introduction to the central limit theorem

The central limit theorem is sometimes called the Lindburg-Lévy Theorem. Its proof is based on a property of convergence in distribution (law). The observation that the Central Limit Theorem is ubiquitous in its use in biostatistics speaks to the fact that convergence in distribution, however "weak" that it may be in theory, can have a wide range of practical implications.

Our goal is to prove the central limit theorem by relying on Lévy's criteria for the convergence of moment generating functions. With our understanding of the Lévy Continuity Theorem, and the prior example of the Poisson distribution converging in distribution to normal measure, the proof of the central limit theorem will be surprisingly easy. We begin with its statement.

## Central limit theorem

If there is a sequence of independently distributed random variables $X_i$, $i = 1,2,3,\ldots, n$, each with a moment generating function $\mathbf{M}_{X_i}(t)$, mean $\mu$, and variance $\sigma^2$, then the probability distribution of the random variable $\dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$ converges in distribution to $\mathbf{N}(0,1)$ measure as $n$ goes to infinity.

With the monitory that convergence in distribution does not mean that the random variable $\dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$ itself becomes a normal random variable, but instead that its distribution can be approximated by normal measure,[*] we proceed.

We can proceed directly to the proof. Let us start as we did in the previous discussion concerning the weak convergence of a Poisson distribution to a normal distribution by beginning with the observation that

$$T_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n\sigma^2}}$$

Thus, the random variable of interest is the sum of $n$ independent, normed random variables. We may use this fact to write

$$\mathbf{M}_{T_n}(t) = \left( \mathbf{M}_{\frac{X_i - \mu}{\sqrt{n\sigma^2}}}(t) \right)^n$$

[*] The CLT does suggest however that however skewed the measure of the random variable $X_i$ may be, the behavior of $\dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$ does exhibit central tendency.

We now examine the moment generating function $\mathbf{M}_{\frac{X_i-u}{\sqrt{n\sigma^2}}}(t)$. We can start by writing

$$\mathbf{M}_{\frac{X_i-\mu}{\sqrt{n\sigma^2}}}(t) \;=\; \sum_{k=0}^{\infty}\frac{E\left[\frac{X_i-\mu}{\sqrt{n\sigma^2}}\right]^k}{k!}t^k \;=\; \sum_{k=0}^{\infty}\frac{E\left[X_i-\mu\right]^k}{k!}\left(\frac{t}{\sqrt{n\sigma^2}}\right)^k$$

$$= 1 + \frac{E\left[X_i-\mu\right]}{1}\left(\frac{t}{\sqrt{n\sigma^2}}\right) + \frac{E\left[X_i-\mu\right]^2}{2}\left(\frac{t}{\sqrt{n\sigma^2}}\right)^2 + \left(o\left(\frac{t}{\sqrt{n\sigma^2}}\right)^2\right)$$

Since $\mathbf{E}\left[x_1\right]=\mu,$ and $\mathbf{E}\left[x_1-\mu\right]^2=\sigma^2,$ we may substitute these quantities to find that

$$\mathbf{M}_{\frac{X_i-\mu}{\sqrt{n\sigma^2}}}(t) = 1 + \frac{\sigma^2}{2}\frac{t^2}{n\sigma^2} + \left(o\left(\frac{t}{\sqrt{n\sigma^2}}\right)^2\right)$$

$$=1 + \frac{t^2}{2n} + o\left(\frac{t}{\sqrt{n\sigma^2}}\right)^2 \;=\; 1 + \frac{\frac{t^2}{2}+no\left(\frac{t^2}{n\sigma^2}\right)}{n}$$

Now it remains for us to take advantage of the fact that the moment generating function of the sum of independent random variables is the product of the moment generating function of the summands. Thus

$$\mathbf{M}_{T_n}(t) = \left[\mathbf{M}_{\frac{X_i-\mu}{\sqrt{n\sigma^2}}}(t)\right]^n = \left(1+\frac{\frac{t^2}{2}+no\left(\frac{t^2}{n\sigma^2}\right)}{n}\right)^n$$

and, taking limits, we conclude

$$\lim_{n\to\infty}\mathbf{M}_{T_n}(t) = \lim_{n\to\infty}\left(1+\frac{\frac{t^2}{2}+no\left(\frac{t^2}{n\sigma^2}\right)}{n}\right)^n = \lim_{n\to\infty} e^{\frac{t^2}{2}+no\left(\frac{t^2}{n\sigma^2}\right)}$$

$$= e^{\frac{t^2}{2}+\lim_{n\to\infty} no\left(\frac{t^2}{n\sigma^2}\right)} = e^{\frac{t^2}{2}}.$$

This is the moment generating function of standard normal measure. Note that this finding does not reveal any property of the individual random variable itself other than that it is independently distributed with common means and variance and has a moment generating function.

## The delta method

From the preceding section, we have seen that, under commonly occurring conditions, the measure of the sample mean of a random variable (suitable normalized) can be very reasonably approximated by a normal distribution. In the examples that we have developed thus far, the mean and variance of the random variable have been directly available.

However, frequently, we are interested in the distribution of functions of these random variable that do not have an easily calculated mean and variance. A procedure, commonly known as the delta method, provides a very helpful result. It is at its heart, the application of the central limit theorem with the use of a Taylor expansion to approximate the mean and variance of the random variable in question.

## Convergence in probability

Convergence in distribution has provided a helpful approach to the notion of limiting probabilities, not the least of which is of course the Central Limit Theorem. However convergence in distribution is just one of four implementations of limits and probability.

A second approach is what is termed convergence in probability. Perhaps a more evocative name for this type of convergence is weak random variable convergence, for reasons that we will discuss later. First, though, the statement of the definition.

## Definition of convergence in probability

Let $\{X_n\}$ be a sequence of random variables. Then the sequence is said to converge in probability to a random variable *X*, the probability of large deviations of the random variable from its limit is small, i.e., if for each ε > 0 $\lim_{n\to\infty} \mathbf{P}\left[\left|X_n - X\right| \leq \varepsilon\right] = 1.$

There are important differences between convergence in distribution and convergence in probability. Convergence in distribution focuses on the convergence of the cumulative distribution function. There is no statement about the random variable's behavior, only that of $\mathbf{F}_n(x)$.

Here for the first time, we are focusing on the behavior of the random variable itself (note that the requirement $\lim_{n\to\infty} P\left[\left|X_n - X\right| \leq \varepsilon\right] = 1.$ Convergence in probability makes no comment about the cumulative distribution function.

However, convergence in probability does not imply that all but finitely many of the values of the sequence $\{X_n\}$ must be within $\varepsilon$ of the limiting value $X$. This ability to let some random variable realizations "escape" from the requirement of being within $\varepsilon$ of the limiting value $X$ accounts for the "weakness" aspect of convergence in probability.

Consider the random variable follows a $\mathbf{U}\left(0, \frac{1}{n}\right)$ distribution, but only with probability $\frac{n}{n+1}$. With the remaining probability $\frac{1}{n+1}$ it takes on the value $n^2$. Does this random variable converge in probability?

Taking a step back to considering the behavior of the random variable for a moment. It seems to have somewhat of a "split personality". As *n* gets larger, the random variable has a greater and greater probability of falling into ranges that are closer and closer to zero. However, there is always a positive probability that the random variable will take on a value outside this range, and when it does, for large values of *n*, the random variable takes on huge sizes $\left(n^2\right)$. To what extent does this random variable converge?

In order to have convergence in probability we must show $\lim_{n\to\infty} \mathbf{P}\left[\left|X_n - X\right| \leq \varepsilon\right] = 1.$ So, focusing on the argument of the limit. Choose any arbitrary small $\varepsilon > 0$. Then

$$\mathbf{P}\left[\left|X_n - X\right| \le \varepsilon\right] = \mathbf{P}\left[X_n \le \varepsilon\right]$$

and this probability is $\frac{n}{n+1}$.

Now, taking the limit, as *n* increases, we easily see that this probability is one. Thus $X_n \to 0$ in probability.

However, the reverse is not necessarily the case. As a simple thought experiment, consider our previous demonstration that a suitably normalized Poisson random variable converges to a standard normal measure in distribution. This simply means that the cumulative distribution of the Poisson random variable can be approximated by standard normal measure. But does this equilibration of distributions suggest that the standardized Poisson random variable and the normal random variables approximate each other? Certainly not.

Now, consider the following, more formal evaluation. Assume that $X_n$ is a sequence of i.i.d. random variables following a Bernoulli distribution with parameter $\frac{1}{2}$. Let $X$ follow this same distribution. Then clearly $\mathbf{F}_X(x) - \mathbf{F}_{X_n}(x) = 0$ for both values of $x$ and $X_n$ converges to $X$ in distribution.

Now define $Y = 1 - X$. Then $\mathbf{P}\left[X = 0\right] = \mathbf{P}\left[Y = 1\right] = \frac{1}{2}$. This is also true for $\mathbf{P}\left[X = 1\right]$. thus both $X$ and $Y$ have the same distribution function, and $X_n$ must also converge to $Y$ in distribution. However $X$ if never equal to $Y$ nor are the two within $\varepsilon$ of each other. They (cumulative distributions) converge, but the random variables do not.

Note that convergence in probability does not imply that all values of $X_n$ in the sequence must be close to the limit *X***.** We can only say the probability that discrepant values of the random variables occur approaches zero. These discrepant values are unlikely, but we cannot preclude their occurrence. For this reason, convergence in probability is commonly referred to as *weak* convergence. However, we will see later in the chapter that weak convergence can produce very powerful and useful results.

## Examples of convergence in probability

As an example of how weak convergence works, consider a sequence of random variables that are independent and take on the value of either zero or one in accordance with the following rule:

$$\mathbf{P}\left[X_n = 1\right] = 1 - \frac{1}{n}$$
$$\mathbf{P}\left[X_n = 0\right] = \frac{1}{n}$$

Before we try to consider any type of convergent behavior, let's just consider the behavior of these random variables. For $n = 1$, $X_n$ is zero. As *n* increases, however, probability moves away from random variable values of zero to one. Note however, for any value of *n*, $X_n$ can still be either zero or one. It can still bounce back and forth, but the likelihood of a bounce gets smaller and smaller. This is the hallmark of convergence in probability. We would expect then that $X_n$ converges to one in probability.

To prove this, we select our $\varepsilon > 0$ and write

$$\lim_{n\to\infty}\mathbf{P}\left[\left|X_n-1\right|\le\varepsilon\right]=\lim_{n\to\infty}\mathbf{P}\left[X_n=1\right]=\lim_{n\to\infty}\left(1-\frac{1}{n}\right)=1.$$

and we have the desired result. Now suppose we have the following sequence of random variables:

$$\mathbf{P}\left[X_n=1\right]=\frac{1}{2}\left(1-\frac{1}{n}\right)$$
$$\mathbf{P}\left[X_n=0\right]=\frac{1}{n}$$
$$\mathbf{P}\left[X_n=-1\right]=\frac{1}{2}\left(1-\frac{1}{n}\right)$$

Again, before we carry out any formal probability analysis, let's just examine and try to predict the behavior of these random variables. For $n=0$ we have that $X_n$ must be zero. The value of $X_n$ moves away from 0 'in probability" but what does it become? It moves randomly and unpredictably from 1 to -1, never staying at either of these poles for long before it bounces back to the other pole. For this random variable, we know that $X_n$ moves away from zero but it does not settle into a destination that is within $\varepsilon$ of either 1 or -1. This means that for any $\delta>0$ for any value of *N* we can find an $n>N$ such that $\mathbf{P}\left[X_n=-1\right]>\delta.$ and $\mathbf{P}\left[X_n=1\right]>\delta.$ This sequence of random variables does not converge in probability. Of course the sequence of random variables

$$\mathbf{P}\left[X_n=1\right]=\frac{1}{2n}$$
$$\mathbf{P}\left[X_n=0\right]=1-\frac{1}{n}$$
$$\mathbf{P}\left[X_n=-1\right]=\frac{1}{2n}$$

converges to zero in probability.

It is important for you to practice generating sequences like this to first observe their properties, then see if you can prove what you have observed. This types of random variables are not particularly useful in public health applications, but they help you to grasp the underlying concept of convergence in probability.

## Markov's and Chebyshev's inequalities

As it turns out, we can examine the behavior of the probability of events involving a random variable as *n* gets large with some very straightforward mathematical steps.

Let *X* be a non-negative random variable (that is, takes positive measure on the nonnegative reals). Then we can of course write

$$\mathbf{E}\left[X\right]=\int_{\Omega_X}xd\mathbf{P}=\int_0^\infty xd\mathbf{P}.$$

Now, let's choose a positive constant *c*. Then we know

$$\mathbf{E}[X]=\int_0^\infty xd\mathbf{P}=\int_0^c xd\mathbf{P}+\int_c^\infty xd\mathbf{P}\ \geq\ \int_c^\infty xd\mathbf{P}$$

Since $X$ takes positive measure on the nonnegative real number line, then $\int_0^c xd\mathbf{P}\geq 0.$ So we have $\mathbf{E}[X]\geq\int_c^\infty xd\mathbf{P}.$ However, on the set $c\leq x<\infty,\ x\geq c.$ So $\int_c^\infty xd\mathbf{P}\geq\int_c^\infty cd\mathbf{P}=c\int_c^\infty d\mathbf{P}=c\,\mathbf{P}[X\geq c].$ Putting it all together, we write

$$\mathbf{E}[X]=\int_{\Omega_X} xd\mathbf{P}=\int_c^\infty xd\mathbf{P}\geq\int_c^\infty cd\mathbf{P}=c\int_c^\infty d\mathbf{P}=c\,\mathbf{P}[X\geq c].$$

So $\mathbf{P}[X\geq c]\leq\frac{\mathbf{E}[X]}{c}.$ This is known as Markov's inequality.

In order to apply this to asymptotic theory, let's consider a random variable that follows a negative exponential distribution with parameter $\lambda=1.$ We know $\mathbf{E}[X]=1.$ Let $c=n,$ positive integer. Then, using Markov's inequality, we can write $\mathbf{P}[X\geq n]\leq\frac{1}{n}.$

So what does $\lim_{n\to\infty}\mathbf{P}[X\geq n]$ mean? The sequence of probabilities that we must evaluate is straightforward: $\mathbf{P}[X\geq 1]$, $\mathbf{P}[X\geq 2]$ $\mathbf{P}[X\geq 3]$, $\mathbf{P}[X\geq 4]$... . Using Markov's inequality, we write

$$\lim_{n\to\infty}\mathbf{P}[X\geq n]<\lim_{n\to\infty}\frac{1}{n}=0.$$

So this limit of probabilities is zero. The larger the value of $n$, the smaller the value of $\mathbf{P}[X\geq n]$ in such a way that we can get as close to (that is, within $\xi$ of) zero as we want.

Of course, this makes good sense. Since the mean of the random variable is one, the probability that $X$ is very large becomes smaller and smaller. However, note, that while the probability gets smaller and smaller, it is always possible that $X$ can be large. For example, it is possible to have $X\geq 10{,}000.$ It is not very likely, but it remains possible. What is limited here is not the value of $X$, only its probability.

Secondly, the bound on $\mathbf{P}[X\geq n]$ is not very sharp. We can easily compute directly $\mathbf{P}[X\geq n]=\int_n^\infty e^{-x}dx=e^{-n}$. Of course we also know that $\lim_{n\to\infty}\mathbf{P}[X\geq n]=\lim_{n\to\infty}e^{-n}=0.$ If we compare the rates of convergence of the exact value of $\mathbf{P}[X\geq n]=e^{-n}$, versus the Markov limit $\frac{1}{n},$ we see that the exact probability gives a much truer picture of what the rate of convergence is really like. However, Markov's inequality provides a useful approximation.

Suppose we have that

$$\mathbf{P}\left[\left|\overline{x}_n-\mu\right|\geq\varepsilon\right]=\mathbf{P}\left[\left|\overline{x}_n-\mu\right|^2\geq\varepsilon^2\right]=\mathbf{P}\left[\left(\overline{x}_n-\mu\right)^2\geq\varepsilon^2\right].$$

Now we use Markov's inequality to write

$$\mathbf{P}\left[\left(\overline{x}_n-\mu\right)^2\geq\varepsilon^2\right]\leq\frac{\mathbf{E}\left[\left(\overline{x}_n-\mu\right)^2\right]}{\varepsilon^2}.$$

This last statement is a form of Chebyshev's inequality. [*]

## The weak law of large numbers

Application of Markov's inequality gives us one of the most famous laws of probability and statistics - the Law of Large Numbers. All we have to do is apply Markov's inequality to the quantity $|\overline{x}_n-\mu|$ where $\overline{x}_n$ is the mean of independent and identically distributed random variables, and $\mathbf{E}\left[\overline{x}_n\right]=\mu$ and $\mathbf{Var}\left[\overline{x}_n\right]=\frac{\sigma^2}{n}$. If this is the case, then what can we say about $\mathbf{P}\left[|\overline{x}_n-\mu|\geq\varepsilon\right]$?

We set $\varepsilon$ to be some small value greater than zero. The idea is ultimately to show that the probability that $\overline{x}_n$ is very close to its mean $\mu$ (we say that it is "within $\varepsilon$ of $\mu$" goes to one as the sample upon which $\overline{x}_n$ is constructed gets larger and larger. If this is true, then for large $n$, the probability of large differences between the sample mean and its mean becomes small. Begin with

$$\mathbf{P}\left[|\,\overline{x}_n-\mu\,|\geq\varepsilon\right]=\mathbf{P}\left[\left|\,\overline{x}_n-\mu\,\right|^2\geq\varepsilon^2\right]=\mathbf{P}\left[\left(\overline{x}_n-\mu\right)^2\geq\varepsilon^2\right].$$

Now we use Markov's inequality to write

$$\mathbf{P}\left[\left(\overline{x}_n-\mu\right)^2\geq\varepsilon^2\right]\leq\frac{\mathbf{E}\left[\left(\overline{x}_n-\mu\right)^2\right]}{\varepsilon^2}.$$

This last statement is a form of Chebyshev's inequality. [†] And since $\mathbf{E}\left[\left(\overline{x}_n-\mu\right)^2\right]=\mathbf{Var}\left[\overline{x}_n\right]=\frac{\sigma^2}{n}$, we can say

$$\mathbf{P}\left[|\,\overline{x}_n-\mu\,|\geq\varepsilon\right]\leq\frac{\sigma^2}{n\varepsilon^2}.$$

This is all that we need, finishing with $\lim\limits_{n\to\infty}\mathbf{P}\left[|\,\overline{x}_n-\mu\,|\geq\varepsilon\right]\leq\lim\limits_{n\to\infty}\frac{\sigma^2}{n\varepsilon^2}=0.$ This is the proof of the weak law of large numbers, which follows

**Weak Law of Large Numbers**

Let $\{X_i\}$ be a collection of random variables that are independent, each with the same mean $\mu$. Then the sample mean $\overline{X}_n$ converges the population mean $\mu$ in probability.

[*] Chebyshev's inequality is typically stated as $\mathbf{P}\left[|\overline{x}_n-\mu|\geq k\right]=\mathbf{P}\left[\left(\overline{x}_n-\mu\right)^2\geq k^2\right]\leq\frac{\sigma^2}{k^2}.$

[†] Chebyshev's inequality is typically stated as $\mathbf{P}\left[|\overline{x}_n-\mu|\geq k\right]=\mathbf{P}\left[\left(\overline{x}_n-\mu\right)^2\geq k^2\right]\leq\frac{\sigma^2}{k^2}.$

### *Comments on the law of large numbers*

The result that the sample mean converges to the population mean may appear as no surprise to the modern student in statistics. However this finding was by no means clear to 16th century probabilities who were confronted with several competing estimators for the probability of a successful Bernoulli trial. The weak law of large numbers was first demonstrated by Bernoulli in 1713 for independent Bernoulli trials.

However, this original proof did not incorporate Chebyshev's inequality, but instead involved a laborious evaluation of $\mathbf{P}\left[\left|\bar{x}_n - \mu\right| \geq \varepsilon\right]$. However, it was the demonstration of Bernoulli's version of the law of large numbers that supported the natural belief that the proportion of successful Bernoulli trials is an accurate assessment of the probability of a Bernoulli trial.

However, while this intuitive sense that relative frequency computations can be a good measure of the probability of a successful Bernoulli trial is a correct interpretation of this version of the law of large numbers, there remain today some common interpretations of this law that are also intuitive but incorrect.

A false interpretation of the law of large numbers is all too easily injected into the world of gambling. A gambler who is experiencing a run of bad luck commonly believes that continuing to play will assure victory at the gambling table.

If we parameterize his situation by letting $X_n = 1$ if his $n$th gamble earns him money, and $X_n = 0$ if he loses money, then his conviction might be more mathematically stated as a belief that, in the long run, $\overline{X_n}$ must be close to $p$, the probability that the gambler will win his bet.

Therefore, he reasons that the string of $X_n$'s that he has observed for which $X_n = 0$ cannot continue to go on, and will soon be reversed by the occurrence of a compensatory collection of $X_n$'s for which $X_n$ will be one. Thus, he believes that by continuing to play, his "luck will turn around".

However, there are two difficulties with this approach. The first, more obvious one is that, all too commonly, the value of $p$ (the probability of his winning) is lower than he anticipates. This low value would require a far longer appearance at the gambling table than his diminishing fortune will allow.

Secondly, and perhaps less obviously, a change in the gambler's luck occurs far less commonly than our gambler might anticipate. The gambler assumes that the number of times the gambler is winning should increase in proportion to the number of times he gambles. This is not the case. Feller (1) has demonstrated that the number of times the lead changes in a sequence of Bernoulli trials increases not as a function of $n$, but as a function of $\sqrt{n}$. Thus the number of gambles the gambler has to make before he regains his losses is likely to be far longer than he can tolerate.

## Additional results in weak convergence

It seems intuitive that combinations of weakly convergent sequences should themselves be weakly convergent. Here are the demonstrations.

Let $\{X_n\}$ is a sequence of random variables that converges to $X$ in probability and $a$ is a scalar constant. We would expect that the sequence $\{W_n\}$ defined by $W_n = aX_n$ $W_n$ converges in probability to $W = aX$.

We must show that $\lim_{n\to\infty} \mathbf{P}\left[\left|W_n - W\right| \leq \varepsilon\right] = 1$. We write, for $a > 0$

$$\mathbf{P}\left[\left|W_n - W\right| \le \varepsilon\right] = \mathbf{P}\left[\left|aX_n - aX\right| \le \varepsilon\right] = \mathbf{P}\left[a\left|X_n - X\right| \le \varepsilon\right]$$
$$= \mathbf{P}\left[\left|X_n - X\right| \le \frac{\varepsilon}{a}\right]$$

and we can now write that $\lim_{n\to\infty} \mathbf{P}\left[\left|W_n - W\right| \le \varepsilon\right] = \lim_{n\to\infty} \mathbf{P}\left[\left|X_n - X\right| \le \frac{\varepsilon}{a}\right] = 1$ since $\left\{X_n\right\}$ converges in probability to *X*. By the same reasoning, it follows that the convergence of probability of $\left\{X_n\right\}$ to *X* implies that $\{-X_n\}$ converges to $-\boldsymbol{X}$ and therefore the sequence $\{aX_n\}$ converges to *aX* when *a* is negative.

Also, if $\left\{X_n\right\}$ is a sequence of random variables that converges to *X* in probability and $\left\{Y_n\right\}$ is a sequence of random variables that converge to *Y* in probability then the random variable $\left\{W_n\right\}$ where $W_n = X_n + Y_n$ converges in probability to $W = X + Y$.

We must show that, when we are challenged with an $\varepsilon \ge 0$, then we can show that the $\lim_{n\to\infty} \mathbf{P}\left[\left|W_n - W\right| \ge \varepsilon\right] = 0$. Begin by writing

$$\mathbf{P}\left[\left|W_n - W\right| \ge \varepsilon\right] = \mathbf{P}\left[\left|X_n + Y_n - \left(X + Y\right)\right| \ge \varepsilon\right]$$
$$= \mathbf{P}\left[\left|X_n - X + Y_n - Y\right| \ge \varepsilon\right].$$

Now identify $\varepsilon_x$ and $\varepsilon_y$ such that $\varepsilon_x > 0$, $\varepsilon_y > 0$ and $\varepsilon = \varepsilon_x + \varepsilon_y$. Then

$$\mathbf{P}\left[\left|X_n - X + Y_n - Y\right| \ge \varepsilon\right] \le \mathbf{P}\left[\left|X_n - X\right| \ge \varepsilon_{\mathbf{x}} \cup \left|Y_n - Y\right| \ge \varepsilon_{\mathbf{y}}\right]$$
$$\le \mathbf{P}\left[\left|X_n - X\right| \ge \varepsilon_{\mathbf{x}}\right] + \mathbf{P}\left[\left|Y_n - Y\right| \ge \varepsilon_{\mathbf{y}}\right]$$

Thus

$$\lim_{n\to\infty} \mathbf{P}\left[\left|W_n - W\right| \ge \varepsilon\right]$$
$$\le \lim_{n\to\infty} \mathbf{P}\left[\left|X_n - X\right| \ge \varepsilon_{\mathbf{x}}\right] + \lim_{n\to\infty} \mathbf{P}\left[\left|Y_n - Y\right| \ge \varepsilon_{\mathbf{y}}\right] = 0.$$

## Slutsky’s theorem

Other results involving the concept of convergence in probability are available. These include the convergence of sums, differences, products and quotients of random variables which themselves converge in probability. The vehicle through which these results are obtained is Slutsky’s Theorem.

Through its use, we learn that, just as limit functions passed though the continuous function argument for real numbers, we find that the “limit in probability” function passes through continuous functions of random variables that themselves converge in probability.

## Convergence with probability one

Prerequisite

Convergence of sequences of sets

While weak convergence involves the convergence of probabilities, strong convergence involves the direct consideration of whether the random variables themselves converge. This strong converges actually implies weak convergence, i.e., if the random variable converges, then the probabilities of that random variable must also converge.

Recall that when a sequence of random variables $\{X_n\}$ converges in probability, this convergence does not preclude the presence of random variable values that can be far from the limiting value no matter how large *n*.

However, if the probability of this extreme value is small, its occurrence does not block the actual convergence of the relevant probabilities. It is true, though, that this occurrence does attenuate the meaning of convergence in probability.

With convergence with probability 1 (wp1) also known as convergence "almost surely (as)) these extreme random variable values are not just unlikely, they are impossible.*

Suppose we have a sequence of random variables, $\{X_n\}$ such that $X_n = -1^{n+1}$. This is nothing but an alternating series of $1, -1, 1, -1, 1 \ldots$ As we saw in our sequencing discussion, this sequence does not converge because we are never guaranteed to be at the limit for all *n* greater than some *N*. However, consider the sequence, $1, -1, 1, 1, 1, 1, 1, 1, 1, 1, \ldots$ Recall that this sequence converges to 1 since it only takes on the non-limiting value finitely many times. This is the heart of convergence wp1. The random variable value is "guaranteed" after some index N to be within $\varepsilon$ of the limit all the time. For this random variable, infinite subsequences that do not converge to one simply do not occur.

Let's look at another easy example. Let $X_n \sim \mathbf{U}\left(0, \frac{1}{n}\right)$. This sequence of random variables each provides a probability for an interval of real numbers, but the intervals of positive measure get closer and closer to zero for increasing *n*.

However, note that we are guaranteed to never be greater than $\frac{1}{n}$ a boundary that itself is guaranteed to decrease. These are our clues that we may have strong convergence to the value 0.

So let's choose an $\varepsilon > 0$. Will all values of the random variable be less than $\varepsilon$ beyond some point in the sequence? If this is so, then we have established convergence wp1 (convergence almost surely (as)).

The answer is "yes". For any $\varepsilon > 0$, one can find a large enough $N$ such that $\frac{1}{N} \leq \varepsilon$. Since the random variable $X_n$ must follow a follow a $\mathbf{U}\left(0, \frac{1}{N}\right)$ then all values of $X_n$ are trapped between $\left[0, \frac{1}{N}\right] \subseteq [0, \varepsilon]$. For $n > N$, the random variable $X_n$ is even more closely trapped from above to an interval with zero as the lower bound. Thus the sequence of $\{X_n\}$ converges to zero almost surely. As the case in our previous example, sequences of the random variables that do not converge never occur.

Almost surely convergence is a powerful tool because it relies on the guaranteed behavior of the random variable, as opposed to convergence in probability which permits 'renegade" random variable values but ensures that this nonconvergent behavior must become less likely as *n* increases.

* Some writers term this extreme event as occurring with probability (or measure) zero; however, since this language construction is unhelpful at this point in our didactic approach we will avoid it.

The definition of convergence almost surely is as simple as we would have hoped.

**Convergence almost surely (also known as convergence with probability one).**
A sequence of random variables $\{X_n\}$ converges almost surely to $X$, when for every ε > 0, $X_n$ is within ε of $X$ all but finitely many times.

## Strong law of large numbers

While we have demonstrated that the sample mean of a sequence of i.i.d. random variables $\{X_n\}$ converges weakly (i.e., in probability) to the population mean $\mu$, it can be shown that this convergence also has the fundamental feature of strong convergence. The implication of this is that, while weak convergence assures us that, while large deviations of $\overline{X}_n$ from $\mu$ can occur, they happen with vanishingly small probabilities, strong convergence guarantees us that large deviations do not happen, i.e., there is no sequence of random variables that meets the criteria of the law that does not converge.

While this strengthening brings little new force to applications, it is of considerable theoretical interest. However its proof is quite complicated.*

## Convergence in r$^{th}$ mean

Finally, we will define convergence in mean square.

**Definition 7.4. Convergence in mean square**
*Let $\{X_n\}$ be a sequence of random variables. Then if $\lim_{n\to\infty} E\left[|X_n - X|^r\right] = 0,$ then $X_n$ converges to $X$ in r$^{th}$ mean. In particular, if r = 2, we say that $\{X_n\}$ converges to X in mean square.*

Unlike convergence almost surely, convergence in mean square does not focus on the random variables behavior as much as it does on the measure's behavior, assessing this behavior through the expectation.

Convergence in $r^{th}$ mean is quite useful, very powerful, and oftentimes is easy to prove. When $r = 2$, assessing convergence in mean square reduces to assessing the long term behavior of the variance.

For example, in order to show that the sample mean $\overline{X}_n$ of a collection of $n$ independent observations from a normal distribution with known mean $\mu$ and finite variance $\sigma^2$, converges to $\mu$ in mean square, we need merely write $\lim_{n\to\infty} \mathbf{E}\left[\overline{X}_n - \mu\right]^2 = \lim_{n\to\infty} \mathbf{Var}\left[\overline{X}_n\right] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$

## Example: Covid19 positivity rates

One aspect of managing a nationwide or statewide contagion is relationship between the proportion of the community infected and the consequences of the infection. If a link can be drawn between how common the infection is, and sequela, e.g., hospitalizations and deaths that occur due to the agent and subsequent to the agent, then, *ceteris paribus* the epidemiologists can understand what a particular proportion infected implies for hospitalizations and deaths, permitting health care administrators and health providers the opportunity to help ensure adequate resources are available.

* For example, see https://terrytao.wordpress.com/2008/06/18/the-strong-law-of-large-numbers/

Researchers are interested in assessing the proportion of counties in a state that have a COVID19 positivity rate of 20% or more. How can they use the county based data to reliably estimate this essential metric?

The county level assessment is accumulated from each of the testing sites within the county. Each testing site controls the number of tests it carries out, and also through its testing regime, knows which of these tests is positive. It can therefore measure its own positivity rate. This positivity rate for the county is a weighted average of rates reported by each testing center, the weights being related to the number of tests conducted by each center.

If there are *n* counties, then each provides its estimate $X_1. X_2, X_3, \ldots X_n$. Let's now convert the $i^{th}$ county's estimate into a new variable $W_i$ which is 1 if the estimate is greater than 20% positivity, 0 otherwise. The workers wish to know if they can estimate $\mathbf{P}\left[X_i \leq 0.20\right]$ or $\mathbf{F}_X\left(0.20\right)$ by $\overline{W}_n$, the mean of the dichotomous random variable set $\left\{W_i\right\}, i = 1, 2, 3, \ldots n.$

For this we can look to the properties of the estimator. In particular, we are interested in showing if

$$\lim_{n\to\infty} \mathbf{E}\left[\overline{W}_n - \mathbf{F}_X(0.20)\right]^2 = 0.$$

We begin with an examination of the properties of

$\overline{W}_n$? From our work with Bernoulli trials, we know that the expected value of $W_i$ is

$$\begin{aligned}\mathbf{E}\left[W\right] = \mathbf{E}\left[\mathbf{1}_{\left[0\leq X_i \leq 0.20\right]}\right] &= \mathbf{P}\left[X_i \leq 0.20\right] + \mathbf{P}\left[X_i > 0.20\right] \\ &= \mathbf{P}\left[X_i \leq 0.20\right].\end{aligned}$$

Since $\mathbf{P}\left[X_i \leq 0.20\right] = \mathbf{F}_X(0.20),$ we note that $\mathbf{E}\left[\overline{W}_n - \mathbf{F}_X(0.20)\right]^2$ is simply the variance of $\overline{W}_n$, $\mathbf{Var}\left[\overline{W}_n\right]$.

We now compute this variance.

$$\begin{aligned}\mathbf{Var}\left[W_i\right] = \mathbf{E}\left[W_i^2\right] - \mathbf{E}^2\left[W_i\right] &= \mathbf{F}_X\left(0.20\right) - \left[\mathbf{F}_X\left(0.20\right)\right]^2 \\ &= \mathbf{F}_X\left(0.20\right)\left[1 - \mathbf{F}_X\left(0.20\right)\right].\end{aligned}$$

Which uses

$$\begin{aligned}\mathbf{E}\left[W_i^2\right] = \mathbf{E}\left[\mathbf{1}_{X_i \leq 0.20}\right] &= 1\mathbf{P}\left[X_i \leq 0.20\right] + 0\mathbf{P}\left[X_i > 0.20\right] \\ &= \mathbf{P}\left[X_i \leq 0.20\right] = \mathbf{F}_X\left(0.20\right).\end{aligned}$$

Proceeding

$$\begin{aligned}\mathbf{Var}\left[\mathbf{F}_n\left(0.20\right)\right] = \mathbf{Var}\left[\sum_{i=1}^{n} W_i \Big/ n\right] &= \frac{n\mathbf{Var}\left[W_i\right]}{n^2} \\ &= \frac{\mathbf{F}_X\left(0.20\right)\left[1 - \mathbf{F}_X\left(0.20\right)\right]}{n}.\end{aligned}$$

We may now conclude that

$$\lim_{n\to\infty}\mathbf{E}\left[\mathbf{F}_n\left(0.20\right)-\mathbf{F}_X\left(0.20\right)\right]^2=\lim_{n\to\infty}\mathbf{Var}\left[\mathbf{F}_n\left(0.20\right)\right]$$

$$=\lim_{n\to\infty}\frac{\mathbf{F}_X\left(0.20\right)\left[1-\mathbf{F}_X\left(0.20\right)\right]}{n}=0$$

Thus the estimate converges in mean square to the state estimate. The fact that the empirical measure of the distribution function converges to the theoretical cumulative distribution function is sometimes described as the Fundamental Theorem of Statistics. ■

## Convergence modality relationships

As we suspected, the introduction of probability into our understanding of limiting process has been complicated.

We have discussed four different modes of convergence that involve the concept of probability. These are convergence in distribution, convergence in probability, convergence almost surely (or convergence with probability 1), and convergence in mean square. Each of these modes has been the foundation of important theoretical work in probability, but they have different implications.

A sequence of random variables that converges in probability also must converge in distribution. A sequence of random variables that converges almost surely, also converges in probability, and therefore converges in distribution. Similarly, a sequence of random variables that converges in mean square also converges in probability, and must converge in distribution as well.

The relationship between convergence almost surely and convergence in mean square is complex. Each of them implies convergence in probability and convergence in distribution but if a sequence of random variables $\left\{X_n\right\}$ $\left\{X_n\right\}$ converges almost surely to *X*, it need not converge in mean square to *X*. Also, convergence in mean square does not imply convergence with probability one.

Finally, as pointed out above, a sequence of random variables that converges almost surely must also converge in probability. However, if $\left\{X_n\right\}$ is a sequence of random variables that converges in probability to a random variable *X*, although $\left\{X_n\right\}$ does not converge to *X* with probability one, it contains a sub-sequence $\left\{X_{n_k}\right\}$ that does converges to the value *X* with probability one.

The demonstration of convergence in distribution of the binomial distribution and the negative binomial measure to the Poisson distribution, and convergence of the Poisson to normal measure each represent examples of the utility of convergence in distribution. The continued, vibrant applicability of these sturdy results no doubt cause confusion when these results are described as "weak" in some tracts.

However, we must keep in mind that, within mathematics, weak is not synonymous with the terms "fragile", "pathetic" or puny". To the mathematician, describing results as weak merely means that the finding implies less than other types of findings. It does not imply that the implications of weak results are useless. As we have seen, (central limit theorem) the consequences of weak convergence are quite useful and powerful .

Another type of convergence that we saw earlier in this chapter is that of convergence almost surely. Recall that convergence almost surely implied that it was impossible for any possible value of the random variable to be far from the limit point when *n* was large enough (as opposed to “weaker” forms of convergence where it is possible, but unlikely for the random variable to be far from its limit point for large *n*).

In some cases, almost sure convergence produces its own probability law. For example, the strong law of large numbers tells us that not only does the sample mean converge to the population mean in probability (i.e., weakly), but also with probability one (i.e., strongly). These can be very useful results in probability theory.

In addition, probabilists have worked to relax some of the assumptions that underlie the law of large numbers. The most common form of this law is when the sequence of observations is independent and identically distributed. Relaxations of these criteria are available, but they require new assumptions about the existence of variances and assurances that the variances do not grow too fast.

Feller[1] contains an examination of the degree to which the assumptions that support the central limit theorem can be relaxed. There are also multivariable extensions of this very useful theorem (Sterling).

## Conclusions

Asymptotic distribution theory is central to the study of the properties of infinite sequences of random variables. The use of probability and statistics is a daily occurrence, and we must have reliable estimates of statistical estimators on which important scientific and public health decisions reside.

While there are several reasons why an estimate obtained from a sample can be inaccurate, an ultimate explanation lies in the fact that the estimate is based not on the population, but instead on only a sample of the population. Thus, statisticians, when confronted with several different and competing estimators, commonly focus on its asymptotic properties. Our intuition serves us well here. One useful style of intuition informs us that, if the estimate is a good one, then the larger the sample becomes, the closer the estimate based on the sample will be to the true population value.

These asymptotic properties are used in very comprehensible contexts. For example, a researcher may ask whether the estimator under consideration consistently overestimates or underestimates the population parameter that it attempts to measure. If it does, then that estimator is asymptotically biased, suggesting that perhaps another estimator should be sought.

The properties of convergence in distribution, convergence in probability, convergence in law and convergence almost surely help us to establish different useful metrics to assess the utility of these estimators. When there are several estimators that purport to measure the same population parameter, then a comparison of the large sample variances of these estimators might be a useful way to select from among the competitors.

These are each important questions that asymptotic theory addresses. In order to examine these issues, we must understand the use of limits in probability and statistics, and learn some of the classic results of the utilization of this theory.

An Introduction to the Concept of Measure
Elementary Set Theory
Sequences of Sets
Sequences of Functions
Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

---

References

Dudewizc, E.J., Mishra, S.N. (1988) *Modern Mathematical Statistics*. New York. John Wiley and Sons.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications* – vol 1. Third Edition New York. John Wiley & Sons. p 84.

Parzen, E. (1960) *Modern Probability Theory and Its Applications*.. New York. John Wiley & Sons. p 372.

Rao, M.M. (1984) *Probability Theory with Applications.* Orlando. Academic Press. p 45.

Stirling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. New York. John Wiley and Sons.

# Convergence of Binomial to Poisson Distribution

## Prerequisites

The Concept of the Limit
Convergent Series
Cauchy Sequences
Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Advanced Binomial Distribution
General Poisson Process

Our study of the binomial and Poisson distributions demonstrated that the random variables following theses distributions are related. For example work in conditional probability theory using the Poisson distribution revealed binomial random variables.

## Heuristic perspective

However, these events are related in yet a more intricate setting. Consider a Poisson process with on average $\lambda$ arrivals in a unit time. Let's divide that unit time into $n$ time intervals equally spaced where $n$ is quite large (for the sake of example, say greater than ten thousand). Then the probability of an arrival in this tiny time interval is quite small, with the arrival rate of $\frac{\lambda}{n}$. Here the intervals are so small that the only probability of a nonzero event in any interval that is worth worrying about is the probability of one arrival in the interval. Thus $\mathbf{P}_k$, or the probability of $k$ arrivals in unit time is the probability that there is one arrival in each of $k$ of the $n$ time intervals. Since arrivals are independent, this probability follows the binomial law, $\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$, so this serves as an approximation to $\mathbf{P}_k$. In order to remove the approximation, we must demand that the time intervals get exceeding small, i.e., to let $n$ increase to infinity. This approach we used in examination of the contagion model.

However, this is really only a demonstration. A more exact prove that can be accomplished in just a short argument involves the generating functions of the binomial and

Poisson distributions. In order to execute this, we must familiarize ourselves with an important limit equality from calculus.

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x.$$

Let's begin with probability generating function for the binomial distribution, which we showed was $\mathbf{G}_s(t) = (q+ps)^n$. Now let's operate under the discussion above, permitting $p = \frac{\lambda}{n}$. Then we may write

$$\mathbf{G}_s(t) = (q+ps)^n = (1-p+ps)^n = (1+p[s-1])^n$$
$$= (1+p[s-1])^n = \left(1+\frac{\lambda[s-1]}{n}\right)^n.$$

Now taking limits, we have $\lim_{n\to\infty}\mathbf{G}_s(t) = \lim_{n\to\infty}\left(1+\frac{\lambda[s-1]}{n}\right)^n = e^{\lambda[s-1]}$

Which we recognize as the probability generating function of the Poisson distribution. It is the Lévy-Cramér Continuity Theorem that permits us to argue that since the generating function of the binomial distribution function converges to the Poisson distribution, then probabilities that are in fact binomial can be approximated by those of the Poisson distribution. In this case, the convergence (what we will come to know as convergence in distribution) works best when *p* is small and *n* is large.

Note, this does not conclude that Poisson random variables and binomial random variables are the same. Only that binomial probabilities can be approximated by Poisson probabilities.

Hypergeometric Measure
The Geometric and Negative binomial measures
Limits and Continuity
Moment Generating and Probability Generating Functions
Generating Function Inversion

# Alternative Demonstration of Poisson to Normal Convergence

## Prerequisites

The Concept of the Limit
Convergent Series
Cauchy Sequences
Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Advanced Binomial Distribution
Normal Measure

We have provided one derivation of convergence in distribution of the Poisson to normal measure. However, another demonstration of this result illustrates the importance of a finite variance in the limiting process.

## Initial discussion

In this circumstance, we are interested in identifying the moment generating function for the standardized sum of *n* Poisson random variables that we will write as

$$T_n = \frac{\sum_{i=1}^{n} x_i - n\lambda}{\sqrt{n\lambda}}$$

We will first use standard moment generating function techniques to find the moment generating function of $\mathbf{M}_{T_n}(t)$. Begin by writing  expression

$$T_n = \frac{\sum_{i=1}^{n} X_i - n\lambda}{\sqrt{n\lambda}} = \frac{1}{\sqrt{n\lambda}} \sum_{i=1}^{n} \left( X_i - \lambda \right)$$

We may now write

$$\mathbf{M}_{T_n}(t) \;=\; \mathbf{M}_{\sum_{i=1}^{n}(x_i-\lambda)}\left(\frac{t}{\sqrt{n\lambda}}\right) \;=\; \left[\mathbf{M}_{(x_i-\lambda)}\left(\frac{t}{\sqrt{n\lambda}}\right)\right]^n$$

We can now evaluate $\mathbf{M}_{(x_i-\lambda)}\left(\frac{t}{\sqrt{n\lambda}}\right)$.

$$\mathbf{M}_{(x_i-\lambda)}\left(\frac{t}{\sqrt{n\lambda}}\right) = 1 + \frac{\mathbf{E}\left[X_i-\lambda\right]E}{1}\left(\frac{t}{\sqrt{n\lambda}}\right) + \frac{\mathbf{E}\left[X_i-\lambda\right]^2}{2}\left(\frac{t}{\sqrt{n\lambda}}\right)^2 + o\left(\frac{t^2}{n\lambda^2}\right)$$

$$= 1 + \frac{\lambda}{2}\left(\frac{t^2}{n\lambda}\right) + o\left(\frac{t^2}{n\lambda^2}\right) \;=\; 1 + \left(\frac{t^2}{2n}\right) + o\left(\frac{t^2}{n\lambda^2}\right)$$

This reveals that $\mathbf{M}_{T_n}(t) = \left[\mathbf{M}_{(x_i-\lambda)}\left(\frac{t}{\sqrt{n\lambda}}\right)\right]^n$, we may write

$$\mathbf{M}_{T_n}(t) \;= \left[1 + \left(\frac{t^2}{2n}\right) + o\left(\frac{t^2}{n\lambda^2}\right)\right]^n = \left[1 + \frac{\left(\frac{t^2}{2}\right) + no\left(\frac{t^2}{n\lambda^2}\right)}{n}\right]^n$$

and now it only remains for us to take a limit

$$\lim_{n\to\infty}\mathbf{M}_{T_n}(t) \;= \lim_{n\to\infty}\left[1 + \frac{\left(\frac{t^2}{2}\right) + no\left(\frac{t^2}{n\lambda^2}\right)}{n}\right]^n = e^{\lim\limits_{n\to\infty}\left[\left(\frac{t^2}{2}\right) + no\left(\frac{t^2}{n\lambda^2}\right)\right]} = e^{\frac{t^2}{2}}$$

Recognizing that $e^{\frac{t^2}{2}}$ is the moment generating function for the standard normal distribution, we have our desired result.

# Convergence of Negative Binomial to Poisson Measure

## Prerequisites

The Concept of the Limit
Convergent Series
Cauchy Sequences
Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Geometric and Negative binomial measures
General Poisson Process

Let $X$ follow a Negative binomial measure, i.e., $\mathbf{P}\left[X=k\right]=\binom{k+r-1}{r-1}p^r q^k$. The following demonstration reveals without the use of generating functions how a negative binomial measure can be approximated by the Poisson distribution.

## Preliminaries

Begin with

$$\mathbf{P}\left[X=k\right]=\binom{k+r-1}{r-1}p^r q^k=\frac{(k+r-1)!}{k!(r-1)!}(1-q)^r q^k$$
$$=\frac{1}{k!}\frac{(k+r-1)!}{(r-1)!}(1-q)^r q^k=\frac{1}{k!}\frac{(k+r-1)!}{r^k(r-1)!}(1-q)^r (qr)^k.$$

Note the expression $\dfrac{(k+r-1)!}{r^k(r-1)!}$ has exactly $k+r-1$ terms in the numerator and denominator. We can therefore write

$$\begin{aligned}\frac{(k+r-1)!}{r^k(r-1)!} &= \frac{(k+r-1)(k+r-2)(k+r-3)....r}{r^k}\\ &= \left(\frac{k+r-1}{r}\right)\left(\frac{k+r-2}{r}\right)\left(\frac{k+r-3}{r}\right)...\left(\frac{r}{r}\right)\\ &= \left(\frac{r-(1-k)}{r}\right)\left(\frac{r-(2-k)}{r}\right)\left(\frac{r-(3-k)}{r}\right)...\left(\frac{r}{r}\right)\\ &= \left(1-\frac{(1-k)}{r}\right)\left(1-\frac{(2-k)}{r}\right)\left(1-\frac{(3-k)}{r}\right)...\left(\frac{r}{r}\right).\end{aligned}$$

So we have

$$\begin{aligned}\mathbf{P}[X=k] &= \frac{1}{k!}\left(1-\frac{(1-k)}{r}\right)\left(1-\frac{(2-k)}{r}\right)\left(1-\frac{(3-k)}{r}\right)...\left(\frac{r}{r}\right)(1-q)^r(qr)^k\\ &= \frac{1}{k!}\left(1-\frac{(1-k)}{r}\right)\left(1-\frac{(2-k)}{r}\right)\left(1-\frac{(3-k)}{r}\right)...\left(\frac{r}{r}\right)\left(1-\frac{qr}{r}\right)^r(qr)^k\\ &= \frac{1}{k!}\left(1-\frac{(1-k)}{r}\right)\left(1-\frac{(2-k)}{r}\right)\left(1-\frac{(3-k)}{r}\right)...\left(\frac{r}{r}\right)\left(1-\frac{qr}{r}\right)^r(qr)^k\end{aligned}$$

If we now let $r \to \infty,$ and $q \to 0$ such that $rq \to \lambda$ we have

$$\begin{aligned}&\mathbf{limP}[X=k]\\ &= \lim\frac{1}{k!}\left(1-\frac{(1-k)}{r}\right)\left(1-\frac{(2-k)}{r}\right)\left(1-\frac{(3-k)}{r}\right)...\left(\frac{r}{r}\right)\left(1-\frac{\lambda}{r}\right)^r\lambda^k\\ &= \frac{1}{k!}\lim_{r\to\infty}\left[\left(1-\frac{(1-k)}{r}\right)\left(1-\frac{(2-k)}{r}\right)\left(1-\frac{(3-k)}{r}\right)...\left(\frac{r}{r}\right)\right]\lim_{r\to\infty}\left(1-\frac{\lambda}{r}\right)^r\lambda^k\\ &= \frac{\lambda^k}{k!}\lim_{r\to\infty}\left(1-\frac{\lambda}{r}\right)^r = \frac{\lambda^k}{k!}e^{-\lambda}.\end{aligned}$$

Note that the random variable *X* is never Poisson. It always follows a negative binomial measure. However, the negative binomial measure can be approximated by the Poisson distribution.

# The Delta Method

The delta method is a practical application of the central limit theorem and Taylor's expansion. Let $\{X_i\}$ $i = 1$ to $n$ be a collection of independent and identically distributed random variables with mean $\mu$ and variance $\sigma^2$.

Now suppose we have a random variable $Y$ defined as $Y = g(X)$ where $g$ is a differentiable function of $X$. The delta method tells us that

$\sqrt{n}\left(g\left(\overline{X}_n\right) - g(\mu)\right)$ is approximately $N\left(0, \left[g'(\mu)\right]^2 \sigma^2\right)$

The result is almost too good to be true. Finding the exact measure $g(x)$ can be complicated. However, we can invoke the central limit by normalizing $g\left(\overline{X}_n\right)$ [1]. In addition the approximate mean and variance of $g\left(\overline{X}_n\right)$ is each readily computed.

## Derivation

Let $X$ be a random variable with expectation $\mu$ and variance $\sigma^2$. Our goal is to identify the mean and variance of $g(X)$. Our first approach might be to compute

$$\mathbf{Var}[g(X)] = \int_{\Omega_x} \left[g(x)\right]^2 \mathrm{d}\mathbf{P} - \left[\int_{\Omega_x} \left[g(x)\right] \mathrm{d}\mathbf{P}\right]^2$$

however, on many occasions, the integrals in this equation cannot be directly evaluated.

However, an indirect approach is available to us.
In such circumstances, a more indirect approach is available to us.

Consider a simple Taylor series expansion. Writing $Y = g(X)$ as a simple Taylor series approximation around the point $X = \mu,$ we have

$$g(x)$$
$$= g(\mu) + \left[\frac{dg(\mu)}{dx}\right](x-\mu) + \left[\frac{d^2 g(\mu)}{dx^2}\right]\frac{(x-\mu)^2}{2}$$
$$+ \left[\frac{d^3 g(\mu)}{dx^3}\right]\frac{(x-\mu)^3}{3!} + ....$$

Since the higher power terms will be negligible, we will severely truncate this series so that $g(X)$ is a linear function of *X*, or

$$g(x) \approx g(\mu) + \left[\frac{dg(x)}{dx}\right]_{x=u}(x-\mu).$$

Substituting $\overline{X}_n$ for *X* in equation reveals

$$g(\overline{X}_n) = g(\mu) + \left[\frac{dg(x)}{dx}\right]_{x=u}(\overline{X}_n - \mu).$$

Taking expectations of both sides of this equation reveals

$$\mathbf{E}\left[g(\overline{X}_n)\right] = \mathbf{E}\left[g(\mu)\right] + \left[\frac{dg(x)}{dx}\right]_{x=u}\mathbf{E}\left[(\overline{X}_n - \mu)\right] = \mathbf{E}\left[g(\mu)\right].$$

Thus, we have identified an approximation to $\mathbf{E}\left[g(\overline{X}_n)\right]$. The variance of $g\left(\overline{X}_n\right)$ also follows from a similar argument.

$$\mathbf{Var}\left[g(\bar{x}_n)\right] = \mathbf{Var}\left[g(\mu) + \left[\frac{dg(\mu)}{dx}\right](\overline{X}_n - \mu)\right].$$
$$= \mathbf{Var}\left[\left[\frac{dg(x)}{dx}\right]_{x=u}(\overline{X}_n - \mu)\right]$$
$$= \left[\frac{dg(x)}{dx}\right]_{x=u}^2 \mathbf{Var}\left[(\overline{X}_n - \mu)\right]$$
$$= \left[\frac{dg(x)}{dx}\right]_{x=u}^2 \frac{\sigma^2}{n}$$

## Example: Velocity coefficients

Cells commonly undergo three sequential phases of growth. The first is the lag phase during which little growth seen. This is followed by the log phase of growth, where the growth rate is exponential. This period can be very short, and is rapidly followed by a stationary phase where the organisms have reached the maximum tolerated number for the environment's resources.

During the exponential phase of growth the growth rate of the number of cells *y* is determined by the equation $\frac{dy}{dt} = ky,$ leading to the equation $y = Ce^{kt}$, where $C$ is a constant and $k$ is known as the velocity coefficient. Its value determines how accelerant the growth rate is.

## Example: Velocity coefficients

Suppose a collection of cell specimens have their velocity coefficients measured and are seen to be normally distributed with mean $\mu$ and variance $\sigma^2$. What can we say about the measure of the number of organisms at time $t$, $y_t = Ce^{kt}$ ?

While identifying the exact distribution will be difficult, we can invoke the delta method as an approximation. We need only compute $g(\mu) = Ce^{\mu t}$, and

$$\left[\frac{dg(x)}{dx}\right]^2_{x=u} = \left[\frac{dCe^{kt}}{dk}\right]^2_{k=\mu} = \left(Cte^{ut}\right)^2 = (Ct)^2 e^{2\mu t}.$$

The mean number of cells at time $t$ is asymptotically normal with mean $Ce^{\mu t}$ and variance $n^{-1}(Ct)^2 e^{2\mu t}\sigma^2$.

References

1 . Kapada AS, Chan W, Moyé. Mathematical Statistics with Applications. Chapman and Hall.CRC (2005).

# Slutsky's Theorem

There are some very useful properties of random variables that converge in probability that are readily available to us from simple applications of Slutsky's Theorem.

## Slutsky's theorem (general form)

Consider a sequence of random variables $\{X_n\}$ and a collection of functions $h_1, h_2, h_3,..., h_k$ defined on the sequence $\{X_n\}$ so that $h_i(X_n)$ converges to a constant $a_i\ i = 1,2,3...k$. Define $g(a_1, a_2, a_3,..., a_k) < \infty$. Then $g(h_1(X_n), h_2(X_n), h_3(X_n),... h_k(X_n))$ converges to $g(a_1, a_2, a_3,..., a_k)$ in probability. In particular, if $g$ is a rational function (i.e., the ratio of two polynomials), then this result is known as Slutsky's Theorem.

## Applications

This can seem a little overwhelming on first blush, so let's begin with a simple example. Start with a sequence of random variables such that $\{X_n\}$ converges to a constant $\mu$ in probability. Define the family of functions $h_k$ as

$$\begin{aligned}
&h_1(\{x_n\}) = \{x_n\}\\
&h_2(\{x_n\}) = \{2x_n\}\\
&h_3(\{x_n\}) = \{3x_n\}\\
&\vdots\\
&h_k(\{x_n\}) = \{kx_n\}
\end{aligned}$$

Now, we can conclude from our work on convergence in probability that $h_i(X_n)$ converges in probability to the constant $i\mu$.

As an example of a function $g$, we next let the function $g(h_1, h_2, h_3, \ldots, h_k) = \sum_{i=1}^{k} h_i$ . We are now in a position to invoke Slutsky's theorem and obtain

$$\lim_{n\to\infty} g(h_1, h_2, h_3, \ldots, h_k) = \sum_{i=1}^{k} h_i = \frac{k(k+1)}{2}\mu$$

Slutsky's theorem does not have as many applied applications as it does theoretical ones. However, in this latter circumstance, Slutsky's theorem is very powerful theorem for two reasons.

First, it allows functions of random variables to have their convergent properties explored directly.

However, perhaps more importantly, the application of Slutsky's theorem in a very simple form illuminates critical properties of random variables that converge in probability. For example, consider the case where we have a sequence of random variables $\{X_n\}$ that converges in probability to *X*, and a separate sequence of random variables $\{Y_n\}$ that converges in probability to *Y*. Define two functions $h_1$, and $h_2$ as

$$h_1\left(\{x_n\}\right)=\{x_n\};\, h_2\left(\{y_n\}\right)=\{y_n\}$$

Thus the functions $h_1$ and $h_2$ are simply identity functions. Finally, define the function $g(h_1, h_2)$ as $h_1 - h_2$ (i.e., the difference of the two identity functions). Then, from Slutsky's theorem, we find that the function $g(h_1,h_2)$ converges to the value $X-Y$ in probability.

Another way to write this result is if $\{X_n\}$ is a sequence of random variables that converges in probability to *X* and $\{Y_n\}$ is a sequence of random variables that converges in probability to *Y* then the random variable $\{W_n\}$ where $W_n = X_n - Y_\text{n}$ converges in probability to the random variable $W = X–Y$. We have the result directly.

Slutsky's theorem can also be used to prove additional useful lemmas.

If $\{X_n\}$ is a sequence of random variables that converges in probability to *X* and $\{Y_n\}$ is a sequence of random variables that converge in probability to *Y* then the random variable $\{W_n\}$ where $W_n = X_nY_n$ converges in probability to $W = XY$.

If $\{X_n\}$ is a sequence of random variables that converges to *X* in probability and $\{Y_n\}$ is a sequence of random variables that converge to *Y* in probability ($Y \neq 0$) then the random variable $\{W_n\}$ where $W_n=\frac{X_n}{Y_n}$ converges in probability to $W=\frac{X}{Y}$, assuming of course that $Y \neq 0$.

Finally, our first statements introducing the concept of Slutsky's theorem above may appear to be complicated. A much more common and useful form of this theorem follows:

If a random variable $W_\text{n}$ converges in distribution to *W*, and another random variable $U_n$ converges in probability to the constant *u*, then

1) the sequence $\{W_n + U_n\}$ converges in distribution to W + *u*.
2) the sequence $\{W_nU_n\}$ converges in distribution to *uW*.

While the devices suggested by the preceding lemmas are helpful in demonstrating convergence in probability, it is sometimes useful to prove that a sequence of random variables converge in probability from first principles. Consider the following example.

A sequence of random variables $\{X_n\}$ each follow a $U(0,1)$ distribution. Of course, the sequence does not converge in probability. However, what happens to the max $\{X_n\}$ as *n* goes to infinity? In this case, we create a new sequence of random variables $W_1 = \max(X_1)$, $W_2 = \max(X_1, X_2)$, $W_3 = \max(X_1, X_2, X_3)$. Does the sequence $\{W_\text{n}\}$ converge in probability to 1?

We must show that $\lim_{n\to\infty}\mathbf{P}\left[\left|W_n-1\right|>\varepsilon\right]=0.$ We note as before that $\lim_{n\to\infty}\mathbf{P}\left[\left|W_n-1\right|>\varepsilon\right]=\mathbf{P}\left[1-W_n>\varepsilon\right]=\mathbf{P}\left[W_n<1-\varepsilon\right].$ The probability distribution of the $W_n$ is also easily identified. We may write

$$\begin{aligned}\mathbf{P}\left[W_n<1-\varepsilon\right]&=\mathbf{P}\left[Max\left(X_1,X_2,X_3,...X_n\right)<1-\varepsilon\right]\\&=\mathbf{P}\left[X_1<1-\varepsilon,\,X_2<1-\varepsilon,\,X_3<1-\varepsilon,...\;X_n<1-\varepsilon\right]\\&=\prod_{i=1}^{n}\mathbf{P}\left[X_i<1-\varepsilon\right]=\left(1-\varepsilon\right)^n\end{aligned}$$

We can now write

$$\lim_{n\to\infty}\mathbf{P}\left[\left|W_n-1\right|>\varepsilon\right]=\lim_{n\to\infty}\left(1-\varepsilon\right)^n=0.$$

and we have shown that the maximum of a sequence of independent *U(0,1)* random variables converges in probability to one. Finally, we can demonstrate that continuous functions of a sequences of random variables that converge in probability themselves converge in probability, as demonstrated in the following lemma.

Lemma
Let $\{X_n\}$ be a sequence of random variables that converge in probability to *X*. If $f\left(X_n\right)$ is a continuous function at all points $X_n$ in the sequence, then the sequence $\left\{f\left(X_n\right)\right\}$ converges to $f(X)$.

This lemma is yet another implication of Slutsky's theorem. However, a direct demonstration of this lemma is also available. We know that the concept of a continuous function might be expressed in the observation that as two points on the real line get closer to each other, then the value of the continuous function of those two points must also get closer together. The convergence in probability assumption allows us to say that, if it is very likely that those two points *X* and *a* are close to each other, then it is very likely that *f(X)* will be close to *f(a)*.

More formally, we must show that $\lim_{n\to\infty} \mathbf{P}\left[\left|f\left(X_n\right)-f\left(X\right)\right| \leq \varepsilon\right]=1$.

We know that, since $X_n$ converges to *X* in probability, we can go far enough out in the sequence $\{X_n\}$ so that $\mathbf{P}\left[\left|X_n - X\right| \leq \delta\right]$ is large. However, the continuity of the function *f* assures us that if $\left|X_n - X\right| \leq \delta$, then it must be true that $\left|f(X_n) - f(X)\right| \leq \varepsilon$. Thus, $\left|X_n - X\right| \leq \delta$ implies that $\left|f\left(X_n\right)-f(X)\right| \leq \varepsilon$. But if this implication is true that the inequality

$$\mathbf{P}\left[\left|f\left(X_n\right)-f(X)\right| \leq \varepsilon\right] \geq \mathbf{P}\left[\left|X_n - X\right| \leq \delta\right]$$

must also be true. Therefore we know that the quantity $\mathbf{P}\left[\left|f\left(X_n\right)-f(X)\right| \leq \varepsilon\right]$ can be brought arbitrarily close to one by choosing *n* large enough, proving $\lim_{n\to\infty}\mathbf{P}\left[\left|f\left(X_n\right)-f(X)\right| \leq \delta\right]=1$
We may succinctly write this as

$$\lim_{n\to\infty}\mathbf{P}\left[\left|f\left(X_n\right)-f(X)\right| \leq \varepsilon\right] \geq \lim_{n\to\infty}\mathbf{P}\left[\left|X_n - X\right| \leq \delta\right]=1$$

Thus, just as limit function passed though the continuous function argument for real numbers, we find that the limit in probability function passes through continuous functions of random variables.

# Tail Event Measure

Prerequisite
Sequences of Functions
Sequences of Sets

## Elementary discussion and concepts

The occurrence of events in a sequence is a natural phenomenon in applied biostatistics. Questions e.g., “In the next fifty patients, how many will have been exposed to measles?” or “In the next five years, how many patients will have heart failure meeting the indications for transplantation?” are questions that can be parameterized into sequences of random variables and are addressable from familiar measure paradigms.

However, how about “ever” questions? These are questions that do not fix the durations of time. Examples of such questions would be “Will there ever be a cure for cancer?” or “Will there ever be another coronavirus pandemic?” These are questions that deal with an event that plays out over an infinite period of time. Such events are high impact and challenge us with the possibility of their occurrences.[*]

What distinguishes tail events from the consideration of other random variable sequences is the injection of the complication of infinity. The question “Will small pox return in the next 100 years” requires an approach to the solution that is different from the approach that will answer the question “Will smallpox ever return?” The first question considers a finite number of years; the second requires the contemplation of infinity. Tail events revolve around the question of infinity.

While it is natural to conclude that since neither mankind nor public health will endure throughout infinity, consideration of tail events is neither helpful nor germane is incorrect. A negative answer to the question “Will there ever be a thermonuclear holocaust” contains clear public health pertinence.

## Definition

[*] See for example, Barberis N, “The psychololgy of tail events: Progress and Challenges. Yale School of Management, New Haven CT 06540. Nick_bareris@yale.edu

Tail events are events that require the consideration of an infinite sequence of events; we must have a σ-algebra that reflects the incorporation of infinity.

Consider the distinction between the σ-algebras for the two examples 1) there will be a patient with typhus diagnosed in the next 100 days and 2) there will be a patient with typhus diagnosed beyond the first 100 days.

The σ-algebra for event one consists of all possible events in the first 100 days. This is concept with which we are familiar. However, for the second event we have to consider not just an infinite sequence of events (i.e., a patient is diagnosed with typhus on day 101 but no other patient is ever diagnosed with typhus) but all infinite sequences of events. This second σ-algebra is far more complicated. By requiring that the event involve an infinite number of the $X_n$, it falls into the "tail" of the sequence.

Sometimes we can compute these probabilities directly. For example, let *X* be a random variable that follows a Bernoulli distribution where the $\mathbf{P}\left[X=1\right]=p$, and the $\mathbf{P}\left[X=0\right]=1-p$. Let this be an independent and identically distributed random variable indexed by *n*, i.e., $X_n$ is i.i.d. Bernoulli with probability *p*. Can we find the probability that $X_n$=1 infinitely often?

Note that the probability that we seek is indexed over *n*. For any given *n,* the sample space Ω and σ-algebra Σ are clearly specified and the rule of probability is clear, e.g., $p+q=1$. However, once we index over *n*, the sample space and σ-algebra change. The sample space is now the space that considers the joint collection of Bernoulli events over *n*. For example, $\sum_{n=1}^{\infty}\mathbf{P}\left[X_n=1\right]$ is not one and clearly diverges.

With this as background, our intuition tells us that the probability that $X_n=1$ infinitely often should be high. However small the value of *p*, its small value is overcome by the sheer size of $n$.

The key to demonstrating this finding lies in a detailed consideration of the sequence. Whatever the experience has been so far in the sequence, the probability of having at least one more success approaches one in the next infinite subsequence of events, and of course, there are an infinite number of infinite subsequences.[*] We can think of this then by dividing the positive integers into an infinite sequential collection of *N* events. Then, the probability that there is at least one success in one of these epochs is $1-\left(1-p\right)^N$. However, the complete sequence is composed of an infinite collection of these. Let the event that there is an infinite number of these events be $A_\infty$. Then, the probability that there are an infinite number of these events is

$$\mathbf{P}\left[A_\infty\right]=\lim_{n\to\infty}\left(1-\left(1-p\right)^n\right)^n=1.$$

This is true regardless of any positive value of *p* however small.

### *Tail event*

We define a tail event as an event that is based on all σ-algebras after some finite point *m* in the sequence. This tail σ-algebra we can define as $\boldsymbol{T}_m$, where $\boldsymbol{T}_m=\bigcap_{n>m}^{\infty}\Sigma_n$.

For example, let's return to our example of an infinite sequence of i.i.d. Bernoulli trials. We understand what the sample space is for a single trial $X_n$. Call that sample space $\Omega_n$. We also can (in this case, quite simply) denote all members of the σ-algebra for this event

[*] For example consider the family of subsequences for each positive integer *n* $\left\{X_{1_{m \bmod n=0}}\right\}$.

$X_n, \Sigma_n = \{\varnothing, 0, 1\}$. $\boldsymbol{T}_m$ is the intersection of each of these σ-algebras for the rest of the sequence. Since all of these σ-algebras is the same for the remainder of the sequence, then $\boldsymbol{T}_m = \{\varnothing, 0, 1\}$.

Note that the construction of $\boldsymbol{T}_m$ is separate and apart from the measure of the events. For example, assume that our sequence of random variables $\{X_n\}$ is composed of independent Poisson random variables where the $n^{\text{th}}$ member of the sequence has parameter $\lambda_n = n$. Then the probability of small values of $X_n$ decreases for larger $n$. Yet the σ-algebras for $\{X_n\}$ are all the same; $\boldsymbol{T}_m$ is the collection of nonnegative integers.

However, if we set $\Sigma_n = \{0,1\}$ for even $n$ and $\Sigma_n = \{-1, 0\}$ for $n$ odd, then $\boldsymbol{T}_m = \{0\}$ as this is the intersection of these sets.

With this as a definition, we may begin with the Kolmogorov 0-1 law.

## Kolmogorov 0-1 law

*When based on an infinite sequence of independent events* $\{A_n\}$, *the probability of any tail event is either zero or one.*

This is a terse statement with a profound impact. The probability of the occurrence of the event is narrowed down to either 0 or 1 when the event is a tail event. The Borel Cantelli lemma will allow us to compute the probability of several of these events exactly, but first we will prove Kolmogorov's law.

The prove focuses on the truism that if $p^2 = p$, then $p$ is either equal to zero or one. This will be useful if we can show that the probability of a tail event is in fact independent of itself, an observation demonstrating how different these tail events are from our usual considerations.

So, let's begin by denoting an event that is based on the first $n$ events $\{A_1, A_2, A_3, \ldots A_n\}$ in the infinite sequence. This is not a tail σ-algebra, and let's define $\boldsymbol{X}_n$ as an event based on this this σ-algebra.

Define another σ-algebra as $\boldsymbol{T}_n = \Sigma\left(\{A_{n+1}, A_{n+2}, A_{n+3}, \ldots\}\right)$. This meets our definition of a tail σ-algebra. We can conclude at once $\boldsymbol{X}_n$ is an event that is independent of events based on $\boldsymbol{T}_n$ since the $\{A_n\}$ are independent of each other and the σ-algebras that generated them are independent of each other.

Now, let's define a tail event $\Upsilon_n$ as an event that is based on the tail σ-algebra $\boldsymbol{T}_n$. Then, since $\Upsilon_n$ is contained in the tail σ-algebra $\boldsymbol{T}_n$, and $\boldsymbol{T}_n$ is independent of $\boldsymbol{X}_n$ then so too is the event $\Upsilon_n$ independent of $\boldsymbol{X}_n$.

Now, look at a different σ-algebra, $\boldsymbol{X}_\infty$. What does $\boldsymbol{X}_\infty$ look like? It is the collection of all of the σ-algebras of the entire sequence, i.e., $\boldsymbol{X}_\infty = \Sigma\left(\{A_1, A_2, A_3, \ldots A_n, A_{n+1}, A_{n+2}, A_{n+3}, \ldots\}\right)$. In fact $\boldsymbol{X}_\infty$ includes both $\boldsymbol{X}_n$ and $\boldsymbol{T}_n$ on which the tail event $\Upsilon_n$ is based.

We can conclude that $\Upsilon_n$ is independent of $\boldsymbol{X}_\infty$[*]. But $\Upsilon_n$ independent of $\boldsymbol{X}_\infty$, was based on $\boldsymbol{X}_\infty$. The event $\Upsilon_n$ must also be independent of itself. Therefore $\mathbf{P}[\Upsilon_n] = \mathbf{P}[\Upsilon_n \cap \Upsilon_n] = \mathbf{P}[\Upsilon_n]\mathbf{P}[\Upsilon_n] = \mathbf{P}^2[\Upsilon_n]$, and the probability of the tail event is either 0 or 1.

But what types of events are independent of themselves? Events that we think of in the course of our day to day lives are not. With a finite collection of events, possibilities are always exhausted. The event has not occurred, and that is assurance that it will not occur. This is a proper perspective for finite collections of events.

However the situation changes when one deals with the infinite. In that realm, the absence of the event, (say that there are 100 runs of ten successes) in an infinite sequence of Bernoulli trials, may not occur as far out as one goes in the sequence. However, no matter how far one goes out in the sequence, the current location is merely a reflection of starting over again, facing the same (infinite number) of possibilities. In this realm, the non-occurrence of an event says nothing about its occurrence. This is a property of infinity.

## Borel Cantelli

The Borel Cantelli lemma permits a more direct approach to computing the probability of some tail events.

The Borel-Cantelli lemma begins with an infinite sequence of events $A_n$ where $\mathbf{P}[A_n]$ is defined. Then if $\sum_{n=1}^{\infty}\mathbf{P}[A_n]$ converges, then $\mathbf{P}[\limsup A_n] = 0.$ Since $\limsup A_n$ is the subset of the sequence of sets $A_n$ that occurs infinitely often, there are no tail events; all events occur only finitely many times. Thus, we simply need to examine $\sum_{n=1}^{\infty}\mathbf{P}[A_n]$ in order to determine if $A_n$ occurs infinitely often. In our previous example, $\sum_{n=1}^{\infty}\mathbf{P}[A_n] = \sum_{n=1}^{\infty} p = \infty$, so by Borel Cantelli, we know that successes occur infinitely often (as do failures since $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n^c\right] = \sum_{n=1}^{\infty}(1-p) = \infty$ as well).

In fact, since so many events to which we are accustomed to observing are guaranteed by Borel-Cantelli to occur infinitely often, the interesting question is which events do not occur infinitely often. Before we get to that, here is a proof of Borel-Cantelli that using concepts that we have already reviewed.

## Proof of the Borel-Cantelli lemma

Let there be a infinite sequence of events $A_n$, where $\mathbf{P}[A_n]$ is defined. Then if $\sum_{n=1}^{\infty}\mathbf{P}[A_n]$ converges, then $\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] = 0.$

Proof: Given that $\sum_{n=1}^{\infty}\mathbf{P}[A_n]$ converges, we know that $\sum_{n=1}^{\infty}\mathbf{P}[A_n] < \infty.$ In addition, we know that

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] = \mathbf{P}\left[\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_n\right] = \mathbf{P}\left[\lim_{n\to\infty}\bigcup_{m>n}^{\infty} A_n\right].$$

[*] This is an interesting point that deserves some exploration. Essentially the σ-algebra $\boldsymbol{X}_\infty$ is the union of two independent σ-algebras, $\boldsymbol{X}_n$ and $\boldsymbol{T}_n$. If we only know that the union has occurred, we know nothing about whether $\Upsilon_n$ has occurred.

The last term in the above expression has produced $\mathbf{P}\left[\lim_{n\to\infty}\bigcup_{m>n}^{\infty} A_n\right]$. Thus we write

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right]=\mathbf{P}\left[\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty}A_n\right]=\mathbf{P}\left[\lim_{n\to\infty}\bigcup_{m>n}^{\infty}A_n\right]=\lim_{n\to\infty}\mathbf{P}\left[\bigcup_{n=1}^{\infty}A_n\right]$$

$$\le\lim_{n\to\infty}\sum_{n=N}^{\infty}\mathbf{P}\left[A_n\right]=0.$$

the last inequality is justified by the convergence of $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n\right].$

Another approach to this proof is to show that $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n\right]=\infty$ implies $\mathbf{P}\left[\limsup A_n\right]=1.$ Let's find the probability that the infinite sequence of events $\mathbf{P}\left[\limsup A_n\right]=1.$

We have this result if we can show that $\mathbf{P}\left[\left\{\limsup A_n\right\}^c\right]=0.$ Then

$$\mathbf{P}\left[\left\{\limsup_{n\to\infty} A_n\right\}^c\right]=\mathbf{P}\left[\left\{\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty}A_n\right\}^c\right]=\mathbf{P}\left[\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty}A_n^c\right]$$

$$=\mathbf{P}\left[\lim_{n\to\infty}\bigcap_{m>n}^{\infty}A_n^c\right]=\lim_{n\to\infty}\mathbf{P}\left[\bigcap_{m>n}^{\infty}A_n^c\right].$$

Since the $A_n$ events are independent, we write

$$\lim_{n\to\infty}\mathbf{P}\left[\bigcap_{m>n}^{\infty}A_n^c\right]=\lim_{n\to\infty}\prod_{m>n}^{\infty}\mathbf{P}\left[A_n^c\right]=\lim_{n\to\infty}\prod_{m>n}^{\infty}\left(1-\mathbf{P}\left[A_n\right]\right)\le\lim_{n\to\infty}\prod_{m>n}^{\infty}\text{-e}^{\mathbf{P}\left[A_n\right]}.\ ^{*}$$

Continuing,

$$\lim_{n\to\infty}\mathbf{P}\left[\bigcap_{m>n}^{\infty}A_n^c\right]\le\lim_{n\to\infty}\prod_{m>n}^{\infty}\text{-e}^{\mathbf{P}\left[A_n\right]}=\lim_{n\to\infty}e^{-\sum_{m>n}^{\infty}\mathbf{P}\left[A_n\right]}=0$$

Since $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n\right]=\infty.$ Thus $\mathbf{P}\left[\left\{\limsup_{n\to\infty} A_n\right\}^c\right]=0$ and $\mathbf{P}\left[\limsup_{n\to\infty} A_n\right]=1.$

Let's look at two examples of the behavior of a sequence where $\mathbf{P}\left[A_n\right]$ is a function of *n*. Consider two sequences $A_n$ and $B_n$ each of which can take the value of zero. Here $\mathbf{P}\left[A_n=0\right]=\frac{1}{n},$ and $\mathbf{P}\left[B_n=0\right]=\frac{1}{n^2}.$ In this case $\mathbf{P}\left[\limsup_{n\to\infty} A_n\right]=1$ since $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n\right]$ diverges. However, since $\sum_{n=1}^{\infty}\mathbf{P}\left[B_n\right]=\sum_{n=1}^{\infty}\frac{1}{n^2}$ converges, then $\mathbf{P}\left[\limsup_{n\to\infty} B_n\right]=0.$

[*] From $1-x\le e^{-x}$, which can be proved using the Mean Value Theorem

Are there similar findings for the event $\liminf_{n\to\infty} A_n^c$? Consider the following

$$\mathbf{P}\left[\left\{\liminf_{n\to\infty} A_n\right\}^c\right] = \mathbf{P}\left[\left\{\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_m\right\}^c\right] = \mathbf{P}\left[\left\{\bigcap_{n=1}^{\infty}\bigcup_{m>n}^{\infty} A_m^c\right\}\right]$$
$$= \mathbf{P}\left[\left\{\lim_{n\to\infty}\bigcup_{m>n}^{\infty} A_m^c\right\}\right] = \lim_{n\to\infty}\mathbf{P}\left[\bigcup_{m>n}^{\infty} A_m^c\right]$$
$$\le \lim_{n\to\infty}\sum_{m=n+1}^{\infty}\mathbf{P}\left[A_m^c\right] \le \lim_{n\to\infty}\sum_{n=1}^{\infty}\mathbf{P}\left[A_n^c\right] = 0.$$

Thus, if $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n^c\right]$ converges, then $\mathbf{P}\left[\left\{\liminf_{n\to\infty} A_n\right\}^c\right]$and $\mathbf{P}\left[\liminf_{n\to\infty} A_n\right] = 1.$

Similarly, starting with $\mathbf{P}\left[\liminf_{n\to\infty} A_n\right] = \mathbf{P}\left[\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n\right] = \mathbf{P}\left[\lim_{n\to\infty}\bigcap_{m>n}^{\infty} A_n\right] = \lim_{n\to\infty}\mathbf{P}\left[\bigcap_{m>n}^{\infty} A_n\right].$

Now,

$$\mathbf{P}\left[\bigcap_{m>n}^{\infty} A_n\right] = \prod_{m>n}^{\infty}\mathbf{P}\left[A_n\right] = \prod_{m>n}^{\infty}\left(1 - \mathbf{P}\left[A_n^c\right]\right) \le \prod_{m>n}^{\infty} -e^{\mathbf{P}\left[A_n^c\right]}$$
$$= e^{-\sum_{m>n}^{\infty}\mathbf{P}\left[A_n^c\right]}$$

So

$$\mathbf{P}\left[\liminf_{n\to\infty} A_n\right] = \mathbf{P}\left[\bigcup_{n=1}^{\infty}\bigcap_{m>n}^{\infty} A_n\right] = \mathbf{P}\left[\lim_{n\to\infty}\bigcap_{m>n}^{\infty} A_n\right] = \lim_{n\to\infty}\mathbf{P}\left[\bigcap_{m>n}^{\infty} A_n\right]$$
$$\le \lim_{n\to\infty}\prod_{m>n}^{\infty} -e^{\mathbf{P}\left[A_n^c\right]} = \lim_{n\to\infty} e^{-\sum_{m>n}^{\infty}\mathbf{P}\left[A_n^c\right]} = e^{\lim_{n\to\infty} -\sum_{m>n}^{\infty}\mathbf{P}\left[A_n^c\right]} = 0.$$

Thus $\mathbf{P}\left[\liminf_{n\to\infty} A_n\right] = 1$ when $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n^c\right]$ diverges and $\mathbf{P}\left[\liminf_{n\to\infty} A_n\right] = 0$ when $\sum_{n=1}^{\infty}\mathbf{P}\left[A_n^c\right] < \infty.$

We finish our discussion of tail events with the Kolmogorov's three series theorem, which we offer without proof

## Kolmogorov's three series theorem

Let $\{X_n\}$ be a sequence of random variable. Then the sequence $\sum_{n=1}^{\infty} X_n$ converges almost surely if there exists a constant $A > 0$ for which each of the following conditions holds.

i. $\sum_{n=1}^{\infty}\mathbf{P}\left[\left|X_n\right| \ge A\right]$ converges

ii. If $Y_n = X_n 1_{X_n \le A}$ then $\sum_{n=1}^{\infty}\mathbf{E}\left[Y_n\right]$ converges

iii. $\sum_{n=1}^{\infty} \mathbf{Var}\left[X_n\right]$ converges

The theorem becomes iff when the conditions hold for all $A > 0$. Note that the three series theorem provides conditions for demonstrating that $\mathbf{P}\left[\limsup X_n\right] = 0$ through the application of the Borel Cantelli theorem.

# Introduction to Sigma Notation

In order to manage the manipulation of complicated events, we will need to develop increasingly sophisticated ways to summarize events. We begin with the simple concept of summation, introducing here the tool of sigma notation $(\Sigma)$.

Prerequisite ─ none

Our purpose here is to be able to compute and manipulate the simple sums of numbers. Since each summand can be different, we just denote each by a $x_i$, In order to sum $x_i$, $i$ = 1, 2, 3, ..., *n*, we could simply write the sum as

$x_1 + x_2 + x_3 + \ldots x_n$. However, let's define $\sum_{i=1}^{n} x_i$ as

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n.$$

Constants (that is, summands that do not have the index *i* and therefore do not change as *i* changes) are handled differently and simply. For the simple case where $x_i = c$, $i$ = 1, 2, 3, ..., *n*, then we can see that

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n = c + c + c + \ldots + c = nc.$$

Combinations of constants and summands require special attention, but are easily handled. Thus

$$\sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Similarly $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} (x_i + a) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} a = \sum_{i=1}^{n} x_i + na.$

However, if we define $y_{i,}$ $i$ = 1, 2, 3, ..., *n*,

$$\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (x_i + y_i).$$

No further simplification is available to us since both *x* and *y* are indexed by *i*.

We can continue by writing

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2.$$

However, in order to compute $\sum_{i=1}^{n} (x_i + c)^2$, we have to first carry out the expansion of the argument within the sigma sign to find

$$\begin{aligned}\sum_{i=1}^{n} (x_i + c)^2 &= \sum_{i=1}^{n} \left(x_i^2 + 2cx_i + c^2\right) = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} 2cx_i + \sum_{i=1}^{n} c^2 \\ &= \sum_{i=1}^{n} x_i^2 + 2c\sum_{i=1}^{n} x_i + nc^2.\end{aligned}$$

However, if we have $\sum_{i=1}^{n} (x_i + y_i)^2$, we write

$$\sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} \left(x_i^2 + 2x_i y_i + y_i^2\right) = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$$

We can do nothing more to simplify $\sum_{i=1}^{n} x_i y_i$ since both $x_i$ and $y_i$ are indexed. Alternatively $\left(\sum_{i=1}^{n} (x_i + y_i)\right)^2$ can be written as

$$\begin{aligned}\left(\sum_{i=1}^{n} (x_i + y_i)\right)^2 &= \left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\right)^2 \\ &= \left(\sum_{i=1}^{n} x_i\right)^2 + 2\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + \left(\sum_{i=1}^{n} y_i\right)^2.\end{aligned}$$

Background
Why Probability
From Whence it Came – An Early History of Probability
Probability and the Renaissance

Mathematics Review
Factorials Permutations, and Combinations
Binomial Theorem
Vandermond's Inequality
Pascal's Triangle
Properties of Real Numbers
The Concept of the Limit

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Factorials and Combinations

Prerequisite: None

Factorials reflect useful computations involving the whole numbers (0, 1, 2, 3, 4, ...).[*] While they can become large very fast, they are relatively easy to manipulate and are essential in solving important probability problems.

Prerequisite ─ none.

Factorials are conceptually quite easy. We begin by describing $n!$ for any integer $n$ > 0 as

$$n! = n(n-1)(n-2)(n-3)...(2)(1) = \prod_{k=1}^{n} k.$$

Thus, $4! = (4)(3)(2)(1) = 24$.

The number 1! is simply 1, and it is helpful to define 0! = 1.

Many times, we will achieve extensive cancellation when dividing factorials by other factorials.

For example, while we could compute the quantity $\frac{n!}{(n-1)!}$ by 1) first computing the numerator, then 2) computing the denominator, and finally 3) carrying out the division, it is easier to observe

$$\frac{n!}{(n-1)!} = \frac{n(n-1)(n-2)(n-3)...1}{(n-1)(n-2)(n-3)...1} = n$$

A useful computation for large factorials is Sterling's approximation,

$$n! > \sqrt{2\pi n}\left(\frac{n}{e}\right)^{n}$$

[*] While they can be negative as well, we will not be working with that set of computations here.

an approximation that is most useful in the proof of the DeMoive-Laplace Theorem, one of the first proofs of the Central Limit Theorem.

Other concepts including combinatorics will be introduced in context of counting problems that arise in probability.

## Order, permutations, and combinations

A permutation is the number of possible sequences of events. Consider sequencing the order in which patients are seen at an urgent care clinic. We will denote them as simply patients *A*, *B*, and *C*. The following enumeration is the universe of sequences in which all three patients could be seen.

$$ABC\ ACB\ BAC\ BCA\ CAB\ CBA.$$

Let's think for a minute about how these are computed, taking each of the selections one at a time. Note that a patient occurs once and only once in each sequence. There are three choices for the first patient. However, once we have chosen the first patient, we have two choices for the second patient. Once the second patient is chosen, we have one and only once choice for the third patient. The number of possible orders is $(3)(2)(1)=3!=6.$ Thus, the number of possible permutations is simply a factorial.

Now, let's complicate the problem just a little bit. In this new situation we have five patients *A,B,C,D,E* each of whom requires a knee operation. The clinic can only carry out two operations per day. What is the possible number of ways to identify the two patients who will have procedures on the first day.

We can follow the lead of our previous example and just enumerate them. Thinking of this as 5 subjects "competing" for two slots, we have

$$\begin{array}{c} AB\ AC\ AD\ AE \\ BA\ BC\ BD\ BE \\ CA\ CB\ CD\ CE \\ DA\ DB\ DC\ DE \\ EA\ EB\ EC\ ED \end{array}$$

There are 20 possibilities that we find by rotating through the five subjects, ensuring that all five are always considered for each of the two slots, and – importantly – that a single patient cannot be selected for both slots.

We can think of this problem as selecting two subjects from a collection of 5. There were five choices for the first slot, and then four for the second. We can compute this easily (now without enumerating) as $(5)(4)=20.$

This process can be succinctly written using factorial notation as

$$\frac{5!}{(5-2)!}=\frac{5!}{3!}=(5)(4)=20.$$

The denominator $3!=(3)(2)(1)$ is used to divide or remove the possibilities of ordering for the 3 subjects for whom there are no slots. In general, if we are permuting *n* candidates or objects through *k* possible slots without replacement, then the number of sequences is

$$\frac{n!}{(n-k)!}.$$

## Combinations

If we look at our collection of twenty sequences, it turns out that there may be some duplicates, depending on what we want to count.

For example, the two sequences *AB* and *BA* are each considered. Yet, they are the same event for the day (i.e., both patients *A* and *B* have their knee surgery on the same day. Now, if it matters which is first (for example *A* has a morning surgery and *B* has an afternoon surgery is different than subject *B* having an AM surgery and *A* being seen in the afternoon), (i.e., that order counts) then it is appropriate to have *AB* and *BA* listed as separate events and we are done. However, if we are only concerned about the day and not the order of events within the day, then these are duplicates.

Let's assume order does not count. Then we have to reduce the 20 possibilities by the number of duplicates in order to resolve the duplicate problem.

But how do we find these duplicates?

A quick way to see what the adjustment must be follows. For the selection of a sequence of two patients for the day, there are two possible choices for the first slot, and once chosen, there is only one possible selection for the second producing (2)(1) duplicates. Thus to remove the duplicates, we simply divide the number of permutations by the number of duplicates, in this case reducing $\frac{5!}{3!}$ to $\frac{5!}{3!2!} = 10$. This is the correct computation when order does not count.

This final computation is called a *combination*, and we say the number of distinct sequences of *n* objects when taken *k* at a time is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The *k*! in the denominator is the correction necessary to reduce the duplication in the permutation. This section will be quite helpful for computing probabilities by counting events

Background
Why Probability

Mathematics Review
Sigma Notation
Binomial Theorem
Vandermond's Inequality
Pascal's Triangle
Properties of Real Numbers
The Concept of the Limit

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations

Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Binomial Theorem and Pascal's Triangle

Prerequisite:
Introduction to Sigma Notation

The binomial theorem is attributed to Pascal, and it's simple statement has widespread use. It's use is a fine example of how equalities which may seem difficult if not impossible to prove (in fact to even believe) can be easily managed with the simple application of this theorem.

The simple statement of the binomial theorem is that for any quantities *a* and *b* and any non-negative integer *n*, then

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

We will first prove this using an induction argument. In order to check that the theorem is true for *n* = 1, we simply write

$$(a+b)^1 = a+b = \binom{1}{0} b + \binom{1}{1} a = \binom{1}{0} a^0 b^1 + \binom{1}{1} a^1 b^0 = \sum_{k=0}^{1} \binom{1}{k} a^k b^{1-k}.$$

## Pascal's Recursion

Now assuming that the binomial theorem is true for $n=1,$ we proceed by assuming that it is also true for any $n$, $n>1$ and then must prove that it is true for *n* + 1. For this, we will first need an equality known as Pascal's Recursion, which states

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

To prove this, we develop the right hand side of this equation

$$\begin{aligned}\binom{n-1}{k}+\binom{n-1}{k-1}&=\frac{(n-1)!}{k!(n-1-k)!}+\frac{(n-1)!}{(k-1)!(n-k)!}\\&=(n-1)!\left[\frac{1}{k!(n-1-k)!}+\frac{1}{(k-1)!(n-k)!}\right]\\&=(n-1)!\left[\frac{(n-k)}{k!(n-1-k)!(n-k)}+\frac{k}{k(k-1)!(n-k)!}\right]\\&=(n-1)!\left[\frac{(n-k)}{k!(n-k)!}+\frac{k}{k!(n-k)!}\right]\\&=(n-1)!\left[\frac{n}{k!(n-k)!}\right]\\&=\frac{n!}{k!(n-k)!}=\binom{n}{k}.\end{aligned}$$

## Proving the binomial theorem

This equality will speed our proof of the binomial theorem. Begin with the assertion that for an arbitrary $n>1$, then

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}$$

To prove that this is also true for $n + 1$, we multiply both sides by $(a+b)$ to observe

$$(a+b)^{n+1}=(a+b)\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}=\sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k}+\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k+1}$$

We focus on the right hand side of the preceding equation. Examining the first term, $\sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k}$, we let $h = k + 1$. Then, $k=h-1$ and $n-k=n-h+1$. We now have

$\sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k}=\sum_{h=1}^{n}\binom{n}{h-1}a^hb^{n-h+1}$. For the second term we can simply let $h = k$ to write $\sum_{h=0}^{n}\binom{n}{h}a^hb^{n-h+1}$.

Now recall that

$$(a+b)^{n+1}=(a+b)\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}=\sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k}+\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k+1}$$

Its right hand side now becomes

$$\sum_{h=1}^{n}\binom{n}{h-1}a^h b^{n-h+1}+\sum_{h=0}^{n}\binom{n}{h}a^h b^{n-h+1}$$
$$=\sum_{h=1}^{n}\left[\binom{n}{h-1}+\binom{n}{h}\right]a^h b^{n-h+1}+b^{n+1}$$

The last term being the $h=0$ term from $\sum_{h=0}^{n}\binom{n}{h}a^h b^{n-h+1}$. We now define $L = h$ -1 and rewrite

$$\sum_{h=1}^{n}\left[\binom{n}{h-1}+\binom{n}{h}\right]a^h b^{n-h+1}+b^{n+1}$$
$$=\sum_{L=0}^{n-1}\left[\binom{n-1}{L}+\binom{n-1}{L-1}\right]a^L b^{n-L}+b^{n+1}$$
$$=\sum_{L=0}^{n-1}\binom{n}{L}a^L b^{n-L}+b^{n+1}\ \text{(Pascal's recursion here)}$$
$$=\sum_{L=0}^{n}\binom{n}{L}a^L b^{n-L}$$

This may seem like a complicated proof, but it has remarkable dividends.

## Implications of the binomial theorem

There are many useful results from the binomial theorem. Here are just a few examples. Imagine if we have to prove that

$$0=\binom{n}{0}-\binom{n}{1}+\binom{n}{2}-\binom{n}{3}+\binom{n}{4}+\ldots$$

How could we possible begin such a task? The binomial theorem makes it easy. Simply write

$$0^n=1=(-1+1)^n=\sum_{k=0}^{n}\binom{n}{k}(-1)^k 1^{n-k}$$
$$=\binom{n}{0}-\binom{n}{1}+\binom{n}{2}-\binom{n}{3}+\binom{n}{4}+\ldots$$

Other equally intriguing results are

$$2^n=1=(1+1)^n=\sum_{k=0}^{n}\binom{n}{k}(1)^k 1^{n-k}=\binom{n}{0}+\binom{n}{1}+\binom{n}{2}+\ldots$$
$$(-1)^n=(-2+1)^n=\sum_{k=0}^{n}\binom{n}{k}(-2)^k 1^{n-k}=\binom{n}{0}-2\binom{n}{1}+4\binom{n}{2}+\ldots$$
$$(x+1)^n=\sum_{k=0}^{n}\binom{n}{k}(x)^k 1^{n-k}=\binom{n}{0}+\binom{n}{1}x+\binom{n}{2}x^2+\ldots$$

## Vandemonde's equality

As a final example of the binomial theorem in operation, consider Vandemonde's equality

$$\binom{n+m}{k}=\sum_{i=0}^{k}\binom{n}{k}\binom{m}{k-i}.$$

We know that for any constant a, $(a+1)^n(a+1)^m=(a+1)^{n+m}$. Using the binomial theorem we write

$$(a+1)^n=\binom{n}{0}a^0+\binom{n}{1}a+\binom{n}{2}a^2+\binom{n}{3}a^3+\ldots+\binom{n}{n}a^n$$
$$(a+1)^m=\binom{m}{0}a^0+\binom{m}{1}a+\binom{m}{2}a^2+\binom{m}{3}a^3+\ldots+\binom{m}{m}a^n$$
$$(a+1)^{n+m}=\binom{n+m}{0}a^0+\binom{n+m}{1}a+\binom{n+m}{2}a^2+\binom{n+m}{3}a^3+\ldots+\binom{n+m}{n+m}a^{n+m}.$$

We can also multiply the first two equalities term by term to get

$$\begin{aligned}(a+1)^{n+m}&=\left[\binom{n}{0}+\binom{m}{0}\right]a^0+\left[\binom{n}{0}\binom{m}{1}+\binom{n}{1}\binom{m}{0}\right]a+\\&+\left[\binom{n}{0}\binom{m}{2}+\binom{n}{1}\binom{m}{1}+\binom{n}{2}\binom{m}{0}\right]a^2+\sum_{i=0}^{3}\binom{n}{i}\binom{m}{3-i}a^3+\ldots+\end{aligned}$$

Now, given that we have two expressions for $(a+1)^{n+m}$, we simply equate the coefficients of like powers of $(a+1)$ to find the desired result.

## Pascal's triangle

Another innovation of Pascal is Pascal's Triangle. Rather than compute the combinatoric quantities necessary to use the binomial theorem, Pascal devised a way for them to be computed nearly automatically. They are derived from sums of whole numbers (Figure 1).

```
                1
            1       1
          1     2     1
      1     3       3     1
  1       4     6     4       1
1       5    10    10      5      1
```

**Figure 1.** A portion of Pascal's triangle

The third row's elements are the sum of the two numbers above them, continuing in this fashion for each succeeding row. The first row is $\binom{0}{0}=1$. Thus for the third row, we see that $\binom{2}{0}=1$, $\binom{2}{1}=2$, $\binom{2}{2}=1$. This is a quick way to compute binomial coefficients.

Background
Why Probability

Mathematics Review
Sigma Notation
Factorials Permutations, and Combinations
Properties of Real Numbers
The Concept of the Limit

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Properties of Real Numbers

Prerequisite - None

The real numbers are essential for our work. An important property of the real number line is the character and relative positioning of its numbers. Many of the events whose probabilities we will wish to compute involve real numbers, e.g., $\mathbf{P}\left[n=5\right]$ or $\mathbf{P}\left[-3 \leq x \leq 5\right]$. How we compute these probabilities depends on the probabilities (and ultimately) the properties of real numbers.

## Properties of the universe

But, before we begin to think about this, let's take a trip.

To the edge of the universe.

Suppose that you and I have a comfortable perch at the very outermost edge of the universe[*]. This is a an incredible distance from everything that we recognize, and certainly far from what we call "home". It is in fact farther than even our vivid and "unbounded" imaginations can take us.

Yet, here we are, seated at a vantage point that permits us to see everything that the universe contains.

So, from this view, what consumes our vision?

Space.

Vast, empty, still, cold, silent space.

It is everywhere. In fact unless we peer very carefully, space is all that we see. No mass, no planets, no suns, no rocks. Just motionless, stark, void space.

However if we persistently and carefully focus, we will see some distant specs. They are all but devoured by the surrounding space but there they are. So we race towards them, and find collections of galaxies.

Approaching these galactic consortiums, we are struck with how huge they are. They must be teeming with mass. Now, plunging into the mist of this galaxy collective, we find −

More space.

We expected to see (from our first vantage point) huge quantities of matter, but now that we are here, we observe that mass once again appears in the distance. It is closer, but by and large, we have to confess that we are surrounded not by mass but by space. The same space that inhabited the vastness of the universe, also inhabits the open expanses of galaxies. There is far more space than there are nebula, gases and stars.

Now, we focus in on a single galaxy that from the outside, looks like it is bursting with mass. We dive into it, only to find that this galaxy is inhabited principally not by planets or stars, but by space. Again, the change in perspective reveals space where we expected galaxy filling mass.

[*] Whereever that might be because the current consensus is that the universe is ever expanding.

Anxious for mass, we propel ourselves into the heart of the galaxy towards a cluster of stars, but when we get there, what looked like a closely connected group of stars is actually a small number of stars in a vast empty space. We see a collection of planets, and again, upon arrival we find yet more space. We find a planet, get to the planet and see that it is inhabited principally by space. On and on. Down to molecules. Down to atoms where we find that the elements of the atoms – the protons, and electrons − are separated by a vast space. [*]

The totality of mass that we may think of as overwhelming[†] is inconsequential when compared to the vast emptiness that envelopes it.

It is your willingness to change perspective (i.e., to see things not from the point where you started, but from the vantage point of the environment that surrounds you) that will power your understanding of the real number line.

And we will find space is a property of all collections of items, with one great exception.

## Natural numbers

Let's begin our examination of relationships between numbers by considering the non-negative natural numbers, 0,1,2,….*n*… There are infinitely many of them. However, although we can never complete a count of them, we understand how to enumerate them, that is to say we can follow a sequence of counting so that we do not miss or skip any of them. We call this feature denumerability and say that the natural numbers are infinite and denumerable. We know how to count them, although we cannot actually count them all.

However, it only takes a consideration of the natural numbers to demonstrate how complex the concept of infinity can be.

For example, let's begin with the sequence of natural numbers, and then remove every natural number divisible by three, i.e., 3, 6, 9, 12, 15, …

This leaves us the sequence of natural numbers 0, 1, 2, 4, 5, 7, 8, 10…

This final sequence is an infinite sequence. Yet the number sequence that we have removed, 3, 6, 9, 12,… is itself infinite.[‡] Thus we have removed an infinite sequence of natural numbers from an infinite sequence of natural numbers, but wind up with….an infinite sequence of natural numbers.

It is natural for us to think that we have removed a third of the natural numbers in this subtraction operation, but of what is it a third? With infinity there is no total, and therefore fractions of it need not be finite.

In fact, we can repeat the process, removing the infinite sequence of all natural numbers divided by seven, then those divisible by eleven, then thirteen, etc. In the end, we still have an infinite collection of natural numbers remaining.

However, the essential property of infinity (i.e., its "unendingness") is removed from the remaining sequence if we remove all but finitely many of the natural numbers (e.g., all natural numbers beyond the number 100). In this case we are left with a finite sequence of natural numbers (the first 100 of them). Here, removing an infinite sequence of natural numbers leaves a finite sequence, while in previous subtractions, we had an infinite sequence remaining.

At least at first blush, sometimes subtracting infinity from infinity leaves infinity, other times it does not.

[*] In a hydrogen atom, suitably scaled, if the single proton was on the pitcher's mound in the middle of a baseball field, then the electron is out in the bleachers.

[†] For example, the sun, weighing two gigatons is massive. It can lose ten million tons a second in mass, and still exist for billions of years.

[‡] From this perspective $\infty - \infty \neq 0,$ but instead $\infty - \infty = \infty.$ However, if one subtracted the natural numbers from themselves, then $\infty - \infty = 0.$ This only means that infinity is complex topic requiring careful thought.

The concept of infinity requires us to leave the idea of "total" behind us, a concept not easily accomplished by our practical minds. Also, "simple" arithmetic operations need not be so simple. With infinity, we have to develop a new intuition, one that requires time and contemplation.

It is also true that, although we can never finish counting natural numbers, (that is, from one perspective, they are all that we see), there is in fact space between these numbers – space in which no other natural number resides. In fact if one were to try to accumulate the space taken up by the natural numbers on the one hand, and the space between them on the other, one could conclude that there is substantial more space that does not contain natural numbers than there is of space that actually holds them (Figure 1).

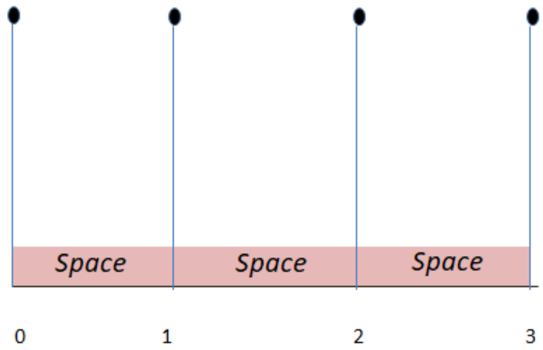


**Figure 1.** Demarcation of space between the whole numbers.

There are an infinite number of natural numbers, but if you are among them, you can see that they are not particularly close together (as in our universe example). There is an infinite number of them, but there is room for more space occupying entities. In mathematics, infinity does not necessarily mean there is no room for nothing else.

## Rational numbers

Next, we consider the rational numbers, that is, those numbers that can be denoted by $\frac{p}{q}$ where $p$ and $q$ are both integers. These are "the fractions", and we are comfortable with them, thinking with and manipulating them in our practical activities of daily living.

The question for us here is, are there more rational numbers on the positive real number line than there are natural numbers?

From one perspective the answer appears to be yes. We begin with the fact that every natural number is itself a rational number. Then, in addition to the natural numbers, look at their reciprocals $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{n}$. With these two sequences (the rational numbers and their reciprocals) we have already "outnumbered" the natural numbers.[*] Plus there are other infinite sequences of rational numbers not contained in these two (e.g., the set of rational numbers between $\frac{1}{2}$ and 1). So there must be more rational than natural numbers.

[*] Here we are acting like $\infty + \infty > \infty$, which is not always the case.

Mustn't there?

No.

When the rational numbers are seen from another perspective, "ordered" such as $1,\frac{1}{2},\frac{1}{3},\frac{2}{3},\frac{1}{4},\frac{3}{4},\frac{1}{5},\frac{2}{5},\frac{3}{5},\frac{4}{5},$ then, they can be placed in one-to-one correspondence with the natural numbers. Removing the natural numbers from the rational numbers is akin to removing the natural numbers divisible by two from the entire set of natural numbers; it is case of $\infty-\infty$ again. Thus from another perspective, there are just as many rational numbers as there are natural numbers.

Thus, we conclude that the rational numbers are countable or denumerable just like the natural numbers even though it seems that there are many more of them. Again, the property of infinity confounds our intuition, intuition that is accustomed to navigating the world of the finite.

Now, how about the space between rational numbers. Are the rational numbers like the natural numbers in that regard as well?

We saw that, from the perspective of a natural number, there is a huge gap between it and the next greater natural number. In fact that gap contains at least the rational numbers, and we now know that there are an infinite number of rational numbers in that gap. From this perspective, the natural numbers are widely spaced indeed.

Clearly the rational numbers are much closer to each other than natural numbers. But are they as close together as they can possibly be? Do they take up all of the "space"?

### *Neighborhoods and limit points.*

To more formally explore this concept, let's create the concept of a neighborhood around a rational number. A neighborhood is all the points within a tiny distance $\varepsilon$ of the original point. On the real number line, it is an interval centered on the original point and extending length $\varepsilon$ both to the left and right of it. (Figure 2)

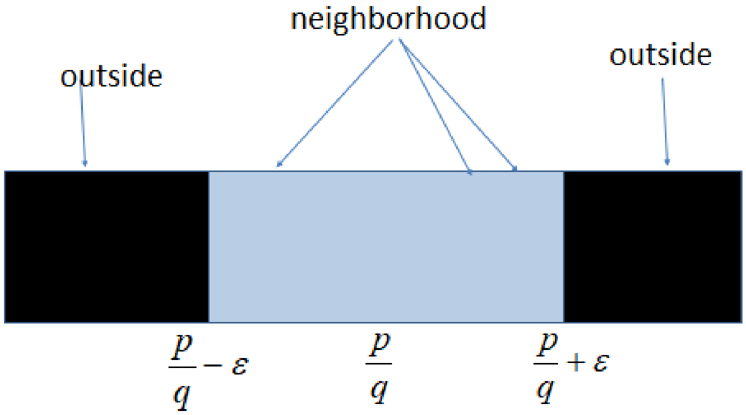


**Figure 2.** A one dimensional epsilon neighborhood around a rational number p/q

Finding a point in the neighborhood of a number regardless of the size of $\varepsilon$ is a demonstration of how close these similar numbers can be. Since, the neighborhood can be as small as we like, then the numbers will be very close to each other. A limit point is defined as a point such that every neighborhood around it – no matter how tiny - contains at least one point with the same property. If every neighborhood contains a point with the property of the original point, then the points may be considered "dense", because the neighborhoods can be as small as we would like.

Certainly, from Figure 1, the natural numbers are not limit points, because for any interval centered on the natural number and a width of less than one, the only natural number in this neighborhood is the original number. Thus, the natural numbers are not dense.

Are the rational numbers on $[0,1]$ limit points? Let's start with a rational number on $[0,1]$ identified as $\frac{p}{q}$. Let's now define a neighborhood around this rational number as $\left[\frac{p}{q}-\varepsilon, \frac{p}{q}+\varepsilon\right]$ where $\varepsilon$ is as small as we like.

We now use a limiting argument. Since we know that $\lim_{n\to\infty}\frac{1}{n}=0,$ we can choose an $n$ large enough so that $\frac{1}{n}$ is arbitrarily close to zero. Is there a rational number between $\frac{p}{q}$ and $\frac{p}{q}+\varepsilon$? Lets choose $N$ large enough such that $\frac{1}{N}\leq\varepsilon.$ If we write $\frac{p}{q}+\frac{1}{N}$ as $\frac{Np+q}{Nq}$, then for any natural number $m$ such that $0\leq m\leq q,$ the number $\frac{Np+m}{Nq}$ permits us to create the following sequence of inequalities:

$$\frac{p}{q} = \frac{Np}{Nq} < \frac{Np+m}{Nq} < \frac{Np+q}{Nq} = \frac{p}{q}+\frac{1}{N} < \frac{p}{q}+\varepsilon.$$

Thus the "new" rational number $\frac{Np+m}{Nq}$ falls in our neighborhood however small $\varepsilon$ is.

Since this is true for all rational numbers on $[0,1]$, then every rational number is a limit point and is in a dense neighborhood of other rational numbers. This is the criteria we need for denseness.

A dense set of numbers is a set of numbers such that every point is a limit point. Thus we know that the rational numbers are infinite, denumerable and dense. So, these, are much closer to each other than the natural numbers. It looks like we are well on our way to squeezing the "space" out of the real number line.

In order for the rationale numbers to do this, they have to fill $[0,1]$? Do they actually do this?

An initial, intuitive answer might be "Yes", because after all the rational numbers are pretty close to each other. But are they so tightly packed that there is no space for anything else?

Now is the time for us to change our perspective.

Recall that the rational numbers are denumerable, and can be sequentially counted $\frac{1}{1},\frac{1}{2},\frac{1}{3},\frac{2}{3},\frac{1}{4},\frac{2}{4},\frac{3}{4}\ldots$ If we can count then, the very process of counting them acknowledges distance between them. The distance may become shorter and shorter as they get closer and closer together, but the fact that we, in our counting, can disconnect from one, and go to the next without skipping any means that we are crossing "space" between them. One way to say this is that, even though they are infinite and dense, there are still gaps between them.

A reasonable depiction of what dense means and does not mean is that of glitter. Suppose one spills glitter on the floor. It appears "everywhere", i.e., the glitter elements are very close to each other. Yet there is "space" between the speckles

The dispersed glitter particles are both dense and separable. They appear "everywhere" from one perspective, but if you change your point of view to that of a single point of glitter, we see that there is space between any two particles. This property of having interleaving space is sometimes denoted as "sparseness". Put another way, just because the rational numbers are dense does not mean that they are jammed in close together with no room between them. It turns out that the rational numbers are both dense and sparse.

Well, then, if there is space, then we must ask, "Is there anything in that space?"

## Properties of irrational numbers

Just because the rational numbers exist in every neighborhood of a rational number does not mean that only rational numbers inhabit that neighborhood. This brings us to the irrational numbers which we will see are far more numerous than the infinite rational numbers.

First, let's show that there are irrational numbers at all. Begin with $y = \log_2 3.$ We will demonstrate its irrationality using a proof by contradiction.

If $y$ is rational, then we can write

$$\log_2 3 = \frac{p}{q}$$

$$2^{\frac{p}{q}} = 3 : \left(2^{\frac{p}{q}} = 3\right)^q$$

$$2^p = 3^q$$

Recall that $p$ and $q$ are integers. Thus $2^p$ must be even. But for any positive integer $q$, $3^q$ must be odd. The finding that an even number equals and odd number is a contradiction, hence, $\log_2 3$ cannot be rational.

The famous finding that $\sqrt{2}$ is irrational is also demonstrated indirectly. Assume that $\sqrt{2}$ is rational, therefore equal to $\frac{p}{q}$ where each of $p$ and $q$ have no common factors. Then $2 = \frac{p^2}{q^2}$, or $p^2 = 2q^2$, which implies that $p^2$ is even. However, since the square of odd integers is never even, $p$ must be even. Thus, there is a $k$ for which $p = 2k.$ then $(2k)^2 = 2q^2$, or $4k^2 = 2q^2$, i.e., $q^2 = 2k^2$ which means that $q$ must also be even. However, if both $p$ and $q$ are even, then they have a common factor, which violated the assumption of our demonstration. Thus $\sqrt{2}$ cannot be rational.

As rigorous as these proofs are, they can be unsatisfying. The steps don't reveal a property of the irrational number that distinguished it from a rational number (other than it is not rational). However, if we dig into the proof a little deeper, we see that what distinguishes an irrational number from a rational one is that the irrational number cannot be written as a simple irreducible fraction.

The irrational numbers are quite numerous. For example, while the product of two irrational numbers can clearly be rational $\left(\sqrt{2}\sqrt{2} = 2\right)$ or irrational as $\sqrt{2}\sqrt{3} = \sqrt{6},$ the product of a rational number and an irrational number must be irrational unless one of them is zero. To see this, let *z* be an irrational number and $\frac{p}{q}$ be rational. If their product is rational, then $z\frac{p}{q}=\frac{a}{b}$, or $z=\frac{aq}{bp}$, which would mean that *z* was rational after all.

Similarly, we can show that sum of a rational and irrational number (for irrational numbers not equal to zero) is irrational. As before, we let *z* be an irrational number and $\frac{p}{q}$ be rational. Then, if the sum is rational, then $z+\frac{p}{q}=\frac{a}{b}$. This implies that $z=\frac{a}{b}+\frac{p}{q}=\frac{aq-pb}{bq}$ which would mean that z is rational, a contradiction.

However, sometimes the sum of two irrational numbers (e.g., $\sqrt{3}+\left(1-\sqrt{3}\right)$ or the difference of two irrational numbers $\left(\sqrt{5}+7\right)-\sqrt{5}$ is rational. However, if $z_1$ and $z_2$ are irrational, then at least one of the $z_1+z_2$ or $z_1-z_2$ is irrational as long as $z_1$ and $z_2$ are not zero.

To see this, let's assume that the sum and difference of $z_1$ and $z_2$ are both rational. Then $z_1+z_2=\frac{p_1}{q_1}$ and $z_1-z_2=\frac{p_2}{q_2}$. Adding these two equations produces $2z_1=\frac{p_1}{q_1}+\frac{p_2}{q_2}$, and $z_1=\frac{p_1}{2q_1}+\frac{p_2}{2q_2}=\frac{p_1q_2+p_2q_1}{2q_1q_2}$ which is clearly rational and a contradiction.

## Some properties of irrational numbers

To demonstrate some of the surprises presented by irrational numbers, consider the following example. It is clear that the sum of two rational numbers must be rational. However, the sum of an infinite number of rational numbers need not be. For example.
$e=1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+...=1+1+\frac{1}{2}+\frac{1}{9}+....$ which is clearly an infinite sum of rational numbers. In fact, every irrational number can be written as an infinite sum of rational numbers
$\sqrt{2}=1.414...=1+\frac{4}{10}+\frac{1}{100}+\frac{4}{1000}+....$ Thus, finite sums of rational numbers are rational.
However, infinite sums of rational numbers may be rational or irrational.

### *Denseness of Irrational Numbers*

It is fairly straightforward to demonstrate that the irrational numbers are dense. Choose $\varepsilon>0$. Then, we know that there exists a positive integer *N* such that for all $n>N$, $\frac{1}{N}<\varepsilon$. Thus, $z+\frac{1}{N}$ is in the $\varepsilon$–neighborhood of *z*. But since $z+\frac{1}{N}$ is the sum of a rational and irrational number, we know that $z+\frac{1}{N}$ must be irrational. Thus, z is a limit point satisfying the definition of density.

It is also easy to show that there is an irrational number between two real numbers *x* and *y*. Since the irrationals are dense we can find an irrational number *r* such that $\frac{x}{\sqrt{2}} < r < \frac{y}{\sqrt{2}}$. This implies that $x < \sqrt{2}r < y.$ We simply let $t = \sqrt{2}r,$ which we know must be irrational*.

## Uncountability of irrational numbers

Thus the set of irrational numbers is dense. But, as we have seen with the rational numbers, this property of density does not mean the neighborhoods consist only of irrational numbers. Number classes can be both dense and sparse. Is there some property that permits us to differentiate rational numbers from irrational numbers.

Yes.

Irrational numbers have a new and distinct property. One cannot count them.

The irrational numbers cannot be counted as can the rational numbers or whole numbers. In fact, there are so many irrational numbers that the very process of counting makes no sense. On the real line, they appear to be indistinguishable from one another. We say that they are nondenumerable.

As an illustration of the difference in denumerability between rational and irrational numbers, consider the following observation. Take a paint brush and, after its bristles are laden with paint, snap the paintbrush into the air repeatedly close to a wall. Now count the droplets on the wall.

One can begin a process of counting the paint droplets and speckles that landed, even though it would (seem to) take forever. This is like the rational numbers –one can lay out a process by which you can count each one without missing any.

Now suppose one takes the brush and simply paints the wall in even strokes until the wall is covered. How does one count droplets in this scenario? It is not just that one cannot begin to count the drops. The concept of a droplet makes no sense. You cannot tell where one ends and the other begins. They are so close together that application of the concept of a “droplet” is lost, because droplets have borders, while confluent paint does not within its margins.

This is what the irrational numbers are like.† Even though we can identify them by name e.g., $\sqrt{11},$ they cannot be disentangled from each other. They are connected to each other.

## Georg Cantor

Irrational numbers were first identified by the Greeks‡. However, it was Georg Cantor who first showed this confluence of irrational numbers in a very revealing way.

He began with the proposition that the rational numbers are countable. Therefore, if the irrational numbers are countable as well, then one should be able to find a one-to-one correspondence with the rational numbers (just as one can find a one-to-one correspondence between the rational numbers and the natural numbers).

Furthermore, he reasoned that if the irrational numbers are countable, then all of the numbers within an interval should be countable as well, and it should be possible to “fill up” the interval with just countable numbers. If the interval cannot be filled, then there must be some other number in the interval, and since it cannot be countable, its cardinality or type of infinity must be different.

* Since it is the product of a rational and irrational number, $t = \sqrt{2}r$ must be irrational.

† Another example of this property is counting individual drops of water on a counter, as opposed to counting the drops in a puddle. The puddle does not permit drops to be separated, and therefore since once does not know where one ends and the other begins, they are impossible to count. So we use another metric e.g., volume.

‡ Hippasius is credited to have discovered that the $\sqrt{2}$ was irrational, and supposedly was drowned at sea by the Gods for upseting the order of things.

So the question becomes, can we ever, using an infinite collection of infinite sequences of rational numbers, fill an interval. For example, lets choose the closed interval $\left[0,1\right]$, and let's try to fill it with the sequence of rational numbers $A_n = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$ .

We can observe that, while this sequence takes up infinitely many available slots on the $\left[0,1\right]$ interval, there are many intervals that remain vacant.

Such an example would be the interval $\left(\frac{1}{2}, 1\right)$ which is not filled by any subset of numbers that is contained in $A_n$. So, in order to correct this deficiency, let's choose another infinite sequence to "fill it". $B_n = \left\{\frac{1}{2} + \frac{1}{n}\right\} = 1,\ \frac{1}{2} + \frac{1}{3},\ \frac{1}{2} + \frac{1}{4},\ \frac{1}{2} + \frac{1}{5}, \ldots$ Since $B_n$ is constructed to focus precisely on the $\left(\frac{1}{2}, 1\right)$ interval that we wish to fill, it fills in many position handsomely.

However, we do notice that many positions in this interval (which in fact is just a subinterval of the original $\left[0,1\right]$ interval) are not filled. In fact, we have created an infinite but countable number of subintervals that continue to retain available spaces waiting to be filled.

To his surprise, Cantor observed that this process continues interminably; using a sequence of tailor-made rational numbers to close up an interval fails to do so, but instead creates a new infinite sequence of subintervals that require filling. The process does not end, precisely because one can always find a new interval between two rational numbers.

In addition, as we have seen before, there is always space between rational numbers, no matter how close the two numbers are to each other. It is this property that makes the rational numbers sparse. It is this space that is filled by the irrational numbers.

Cantor demonstrated that, however one constructs the collections of rational numbers, there will always be irrational numbers that are not members of the sequence. Therefore the irrationals are uncountable or "nondenumerable".

This property connects irrational numbers in a way that rational or natural numbers are not. Irrational numbers are so tightly connected that they cannot be separated and counted. It is as though they blend in to one another on the real line. They are confluent.

Of course we can say that $\sqrt{2}$ is a single, clear and concise number. But you cannot precisely identify its location on the real line, even with (countably infinite) precision. As far down as one goes down with infinite resolving power into the real number line around $\sqrt{2}$, winnowing out numbers close by, you never get to it. It is intermingled and indivisible from other numbers, so much so, that you cannot identify an adjacent one. It is tougher to find the $\sqrt{2}$ on the real number line than it is to find an atom of carbon in a tightly woven carpet. It is both identifiable and also indecipherable and connected.*

However, the irrational numbers have additional surprises. Just as it was Cantor who discovered that the irrational numbers are nondenumerable, he also discovered the Cantor set which is a set of uncountably many irrational numbers that have no length.

## What else is on the real number line?

* This is one of the strangest properties of irrational numbers. Too much thought can challenge sanity, earning these numbers the sobriquet "irrational".

Since the identification of the irrational number was a surprise, one can ask “what other surprises does the real number line hold”. As it turns out, it we stay with the real (and not the complex) number line, there is only one. There is a class of number called transcendental.

At first glance a transcendental number appears to be a special case of an irrational number Many times irrational numbers can be found as the solution to a nonzero polynomial, e.g., $x^2 - 2 = 0$. This is a polynomial whose coefficients are all rational. However, there are irrational numbers which cannot be so derived, e.g., $\pi$ and *e*. The trigonometric and logarithmic functions are other examples. These are known as transcendental numbers.

Research into transcendental numbers continues. However, we can say that irrational, nontranscendental numbers are countable. This follows because they are each derived only from polynomials with rational coefficients. Since there are countably many rational numbers, there are only countably many of these polynomials. In addition, because each polynomial has only a countable number of roots, the set of these nontranscendental irrational numbers is countable. Thus, the uncountable set of real numbers is made up of the transcendental numbers.

In fact, most of the numbers that inhabit the real line are transcendental.

In conclusion, the real number line can be divided into three classes of infinite numbers. The first is the rational numbers, known to be countable. The second class is the nontranscendental irrationals (the algebraic irrationals), also known to be countable. The third is the transcendental numbers, which are uncountable and the largest of the three sets.

Mathematics Review
The Concept of the Limit
Convergent Series
Cauchy Sequences
Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Cantor Sets

Discovered and developed by Georg Cantor, the Cantor set reveals some surprising features about the real line. These sets are worth study, because they demonstrate how intricate the set of numbers in an interval is, and perhaps remind us that although we may believe we know, “all that there is to know” about a concept as simple as the $[0,1]$ interval, this interval always seems to have a secret that lies beyond the horizon of our intuition.[*]

## Development of the set

There are several Cantor sets – the most popular is the Cantor ternary set. The construction is quite simple. Begin by removing the middle third of the line as an open set, i.e., $C_1 = [0,1] \text{ "minus"} \left(\frac{1}{3}, \frac{2}{3}\right)$. This leaves two sets, $\left[0, \frac{1}{3}\right]$ and $\left[\frac{2}{3}, 1\right]$ as we would expect from a set theory construct.

$$\begin{aligned} C_1 &= [0,1] \text{ "minus"} \left(\frac{1}{3}, \frac{2}{3}\right) \\ &= [0,1] / \left(\frac{1}{3}, \frac{2}{3}\right) \\ &= [0,1] \cap \left(\frac{1}{3}, \frac{2}{3}\right)^c \\ &= \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]. \end{aligned}$$

This is what is left after the middle open third of the real line is removed. Removing the middle third of the closed intervals comprising $C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right] = \left[0, \frac{3}{9}\right] \cup \left[\frac{6}{9}, 1\right]$ generates

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{3}{9}\right] \cup \left[\frac{6}{9}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right].$$

[*] Our development here follows the discussion of Christopher Shave, “An exploration of the the Cantor set”.

We continue removing the middle open interval from each of the denoted closed intervals. Each step reveals $C_k$ as the union of $2^k$ intervals each of which is of length $\frac{1}{3^k}$. The Cantor ternary set $C$ is the intersection of these sets $C = \bigcap_{k=1}^{\infty} C_k$, i.e., what is left after each of these removals.

If we continue in this fashion, we can see that we are removing an infinite but countable number of intervals (Figure 1).

The numbers that ultimately remain comprise the Cantor set.

Figure 1 reveals that after 6 such iterations, most of the $[0,1]$ interval has been removed. However, one of the most interesting features about the Cantor set construction is the lengths of 1) the sum of all of the sets that we remove and 2) the sum of those that remain.

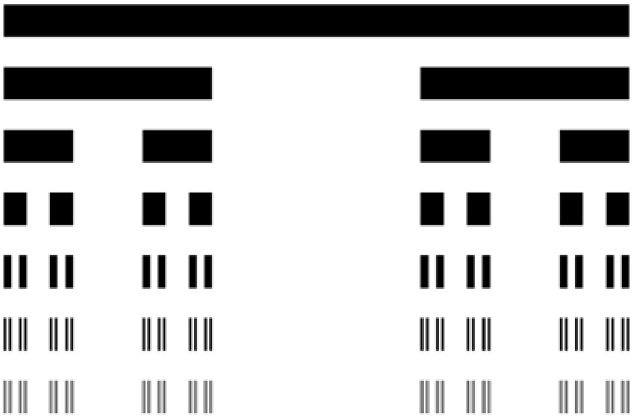


**Figure 1**. Development of the Cantor ternary set by dropping the middle third of each interval

Let's begin with the subintervals of the real number line that are removed. What is their total length $\mathbf{L}(R)$? Counting up, we see that

$\mathbf{L}(R) = \frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \frac{8}{81} + ... + \; = \sum_{k=1}^{\infty} \frac{2^{k-1}}{3^k}$.. But we can write this as

$\mathbf{L}(R) = \sum_{k=1}^{\infty} \frac{2^{k-1}}{3^k} = \frac{1}{3} \sum_{k=1}^{\infty} \frac{2^{k-1}}{3^{k-1}} = \frac{1}{3} \sum_{k=0}^{\infty} \left(\frac{2}{3}\right)^k = \frac{1}{3} \frac{1}{\left(1 - \frac{2}{3}\right)} = 1.$ Thus the length of all of the extractions required to build the Cantor set is in fact length of the $[0,1]$ interval. Yet, we know that the Cantor set is left over. Does this set really have a length of zero?

Yes. While it turns out that the Cantor set is nondenumerable (to be proved in a moment) it must have length of zero.

To see this, we first note that the Cantor set focuses on interval endpoints. Note that in creating the Cantor set, i.e., removing intervals, each interval endpoint can be written in a ternary (base 3) form of only 0's and 2's.* Thus, at every level when an interval is subtracted, what is happening is that we are removing numbers whose ternary expansion contains a one.

* This is an astounishing assertion in and of itself. However, spending some time with a base 10 to base 3 number converter will demonstrate its veracity.

Thus, at the $k^{\text{th}}$ stage of removal, any new interval endpoint either has a 2 in the $3^{-k^{st}}$ ternary place which repeats indefinitely or terminates at the $3^{-k-1^{st}}$ place.

Now consider the number $\frac{1}{4}$. It does not appear as any interval endpoint. However, while we are used to seeing $\frac{1}{4}$ written as 0.25, this is base ten notation. In base 3, $\frac{1}{4} = 0.0202020202020202...$ Thus it is not in any interval that is removed and therefore is in the Cantor set. Yet it is not an interval endpoint. And there are uncountably many such points as 0.25.

To show that the Cantor set is uncountable, we turn to a proof by contradiction. We know that the hallmark of the Cantor set is that it has 1) a 2 in the $3^{-k}$ ternary place which repeats indefinitely or terminates at the $3^{-k-1}$ place.

Let's let the set *w* be the collection of all members of the Cantor set. Now if we assume what we don't wish to prove, i.e., that *w* is countable, then we can index each member of *w* using natural number indices, $w_1, w_2, w_3, ...$ Furthermore, we can write

$$
\begin{aligned}
&w_1 = 0.c_{1_1}, 0.c_{1_2}, 0.c_{1_3}...\\
&w_2 = 0.c_{2_1}, 0.c_{2_2}, 0.c_{2_3}...\\
&w_3 = 0.c_{3_1}, 0.c_{3_2}, 0.c_{3_3}...\\
&...\\
&w_n = 0.c_{n_1}, 0.c_{n_2}, 0.c_{n_3}...0.c_{n_m}
\end{aligned}
$$

where the digit $c_{n_m}$ is either 0 or 2.

Now define the number $z = 0.c_1c_2c_3...c_n...$ where $c_1$ reverses[*] the first digit of $w_1$, $c_2$ reverses the second digit of $w_2$, $c_3$ reverses the third digit of $w_3$ and so on. Since z contains only zeros and twos, it is a member of the Cantor set. But is it a member of *W*?

If *z* were a member of W, say $z = w*$ then every digit of *z* must match every digit of *w** However, we know that it not the case, since z is not equal to *w** in the $n^{\text{th}}$ place. Therefore *z* cannot be member of *W*.[†]

Since we have already shown that the length of the complement of the Cantor set is one, i.e., the length of the entire interval, we know that the Cantor set can contain no intervals.

These properties alone are quite revealing. We could be forgiven for thinking that since 1) the real numbers are nondenumerable and also comprise intervals, and 2) the set of rational numbers is both denumerable and do not solely make up intervals, then it must be the property of nondenumerability that confers the interval making property on the real number line.

The Cantor set demonstrates that this is clearly not the case. Sets of nondenumerable numbers can both be uncountable yet separated and by themselves not form intervals. These such sets have a cumulative length of zero on the $[0,1]$ line, the entire length having been consumed by the extractions to form the Cantor set.

[*] "reverses" in this context means to replace a "2" with a "0" or a "0" with a "2".

[†] This demonstration is the heart of the Cantor diagonalization process that he used to demonstrate that although the rational numbers are infinite, there are more numbers than just the rational numbers, namely the irrationals.

# Limits and Continuity

Prerequisite
Properties of real numbers

The difference between calculus and other areas of finite mathematical structures in mathematics is the concept of the limit, an idea that is principally due to Cauchy. It is through an understanding of the limiting process that we can produce powerful results in mathematics with implications in public health (e.g., the contagion process). At first glance, the limiting process can seem counterintuitive; however, it relies only on the properties of real numbers.

## What does a limit mean?

Let's examine the behavior of the thoroughly understood function $f(k)=\left(\frac{1}{2}\right)^k$ for all non-negative natural numbers. The values of this function comprise a sequence of numbers $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots.$ indexed by *k*.

We begin with the observation that the quantity $\left(\frac{1}{2}\right)^k$ decreases as *k* increases. However, the decrease in this function follows a pattern, allowing us to put bounds on the function's value, regardless of the value of *k*. For example, $f(k)$ cannot be negative. Also, its maximum value is 1. Finally it is always decreasing.

However, given that it can always be decreasing, can $f(k)$ ever be zero? An examination of examples of *k* reveals that whatever value of *k* that we chose, $\left(\frac{1}{2}\right)^k$ is never zero. This interesting property of a function that changes in such a way that, while it approaches a number, is never equal to that number is the heart of the concept of a limit (Figure 1).

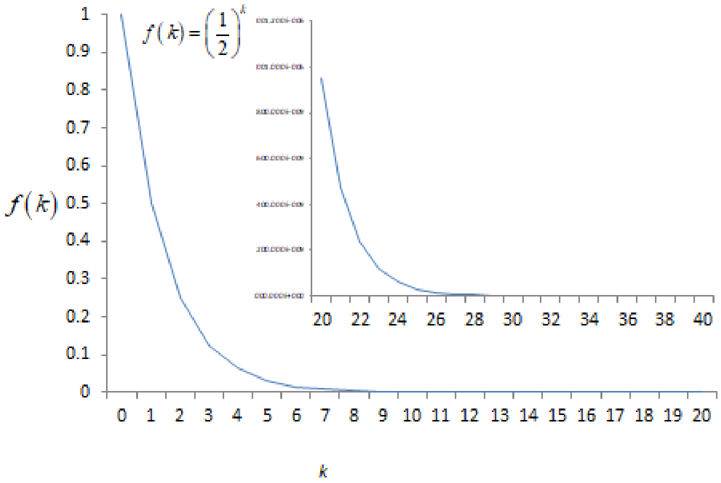


**Figure 1.** Decreasing value of the function $f_k = 2^{-k}$ as $k$ increases. Note from the insert the pattern of the functions shape. A slower rate of decline follows a faster rate.

Figure 1 demonstrates the relationship between $k$ and $f(k)$. From the larger graph in the figure, $f(k)$ first decreases precipitously as $k$ increases. Then as $k$ continues to decrease, the rate at which $f(k)$ decreases is reduced. It appears to "level off".

However, the insert reveals that for larger values of $k$, $f(k)$ continues to demonstrate first a rapid fall followed by a more gradual one. No matter how large the value of $k$, 1) $f(k)$ gets ever closer to zero, and 2) there is always a gap between $f(k)$ and zero.* This is the concept of the limit.

We express this as

$$\lim_{k\to\infty}\left(\frac{1}{2}\right)^k = 0.$$

Other ways to express this are that the function $\left(\frac{1}{2}\right)^k$ converges to 0 or $\left(\frac{1}{2}\right)^k \to 0.$

This does not say that $\left(\frac{1}{2}\right)^k$ is ever equal to zero; we know that it is not. However the concept of "approaching" that is encapsulated in the limit statement is the notion that we can get as close as we want to zero; we simply have to let $k$ be large enough.

For example, if we want $\left(\frac{1}{2}\right)^k \leq 0.00001,$ we simply have to choose $k > 17$. We can get as arbitrarily close to zero as we like. It is this property of the function that will allow us to say $\lim_{k\to\infty}\left(\frac{1}{2}\right)^k = 0.$

* We might think of this as a plane approaching a runway, getting closer and closer, but its wheels never hit the tarmac.

It is fair to ask why the limit could not be some infinitesimal positive value $b > 0$? However, after reflection, we see that, for any positive value *b*, no matter how small, there is some *k* such that $f_k < b$.* Thus $b > 0$ cannot be the limit.

We now will formalize this concept. We say that an infinite sequence of numbers $a_0$, $a_1$, $a_2$, $a_3$, …$a_n$ … approaches the limit *a,* i.e., $\lim_{n\to\infty} a_n = a$, if for any $\varepsilon > 0$, we can find an $n^*$ such that for all $n > n^*$, $|a_n - a| \le \varepsilon$.

This is simply a mathematical way of saying that if $a_n$ truly approaches *a* in the limit, then we can bring $a_n$ in as close as we like to *a* (and staying within that close distance of *a*) by simply going far enough out in the sequence. Therefore, if we are challenged with an ε, (i.e., we are told to have $a_n$ to always be within ε of *a*), we simply have to find the value of *n* ($n^*$) that ensures all subsequent values of $a_n$ are close enough (within ε) of *a*.

With this as a definition, let's now develop a formal argument to show that $\lim_{n\to\infty}\left(\frac{1}{2}\right)^n = 0$. We are challenged with any $\varepsilon > 0$, and must find an $n^*$ such that for any $n > n^*$, $\left|\left(\frac{1}{2}\right)^n - 0\right| = \left(\frac{1}{2}\right)^n \le \varepsilon$.

Our previous work with this function shows how this may be done. Working with limits requires having detailed and intimate knowledge of the functions involved. In this case, the exponential suggests that taking logs is in order. Proceeding,

$$\left(\frac{1}{2}\right)^n \le \varepsilon$$

$$n\ln\left(\frac{1}{2}\right) \le \ln(\varepsilon)$$

$$-n\ln\left(\frac{1}{2}\right) \ge -\ln(\varepsilon)$$

$$n\ln(2) \ge -\ln(\varepsilon)$$

$$n = n^* \ge \frac{-\ln(\varepsilon)}{\ln(2)}.$$

Since we know that $\left(\frac{1}{2}\right)^n$ is always decreasing, all subsequent values $\left(\frac{1}{2}\right)^n$ for $n > n^*$ will be even closer to 0, and we are finished. Note that proving this limit required us to have 1) intimate knowledge of the function $\left(\frac{1}{2}\right)^n$ and also the actual value of the limit.

## All but finitely many times

Looking at the previous example, note that we said that all subsequent values $\left(\frac{1}{2}\right)^n$ for $n > n^*$ will be even closer to 0. Another way to think of this is that there are only finitely many value of $n$ $(n < n^*)$ for which $\left(\frac{1}{2}\right)^n > \varepsilon$, and therefore, reversing it, $\left(\frac{1}{2}\right)^n \le \varepsilon$ for all but these finitely

* To see this, note that $\left(\frac{1}{2}\right)^k \le b$ implies $\ln\left(\frac{1}{2}\right)^k \le \ln b$ and $k\ln\left(\frac{1}{2}\right) \le \ln b$ or $k \le \frac{\ln b}{\ln\left(\frac{1}{2}\right)}$.

many times. If a function $f_k$ is within any $\varepsilon > 0$ of $f$ all but finitely many times, then we can say $f_k \to f$. We will develop this concept of "all but finitely many times" for not just functions but, for sets of objects.

## Examples of convergent series

A sequence that has a limit, converges. There are many useful series that converge. Examples without proof are

$$\sum_{k=0}^{\infty} s^k = \frac{1}{1-s} \text{ for } |s|<1.$$

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

$$\sum_{k=1}^{\infty} \frac{1}{2n} = 2$$

Among the most interesting to us will be

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^{x.}$$

that will be the basis of one of the most useful probability distributions for us, the Poisson distribution.

## Cauchy sequences

For the previous demonstration of the proof that $f(k) = \left(\frac{1}{2}\right)^k$ had the limit of (or converged to) zero, we had to know what the limit actually was. However it is possible to demonstrate that a sequence converges without actually knowing its limit.

We define a sequence $a_0$, $a_1$, $a_2$, $a_3$, …$a_n$ …to be a Cauchy sequence, it for any small value of $\varepsilon > 0$, we can find an $n$* such that for all $n$ and all , $m > n$*, then $|a_n - a_m| \le \varepsilon.$

Note that this definition does not require that we know the limit $a$ of our sequence. However, this definition focuses on a property that makes good sense. If we have a convergent sequence, not only must the elements of the sequence get closer and closer to the limit, but they also must get closer and closer to each other.

We can easily demonstrate that a convergent sequence is a Cauchy sequence. Assume the sequence $\{a_n\}$ is convergent. Then we know that, for a given $\varepsilon > 0$ we can find an $n$* such that

$\forall_{n>n^*}$ then $|a_n - a| \le \frac{\varepsilon}{2}.$ Then choose an integer $m > n.$ Then

$$\begin{aligned} |a_n - a_m| &= |a_n - a + a - a_m| \le |a_n - a| + |a_m - a| \\ &\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

thereby satisfying the Cauchy requirement.*

* Here, we have invoked the rule that any side of a triangle is less than the sum of lengths of the other two sides. This is known as the triangle inequality.

Proof of the reverse direction begins with the assertion that given $\varepsilon > 0$, we can find a value of $n$ and a subsequence $\{a_m\}$ such that $|a_n - a_m|$ is within $\varepsilon$ of this upper bound. Another way to say this is that since $a_m$ does not get to "escape" from $a_n$, $a_m$ is within $\varepsilon$ of $a$.

Assuming $a_n$ and $a_m$ are positive, and that there must be an upper bound $a$ for $a_m$. Thus as $n$ increases, $a_m$ is trapped between $a_n$ and this bound $a$. This implies that for $n$ large enough, $a_n$ is within $\varepsilon$ of $a$. Since $a_n \leq a_m \leq a$, then $a_m$ must also be within $\varepsilon$ of $a$. Thus the sequence converges to $a$. The "successful pursuit" of $a_m$ by $a_n$ implies that $a_m$'$s$ "escape" from $a_n$ is blocked by the bound $a$ which is in fact the limit of the sequence.

## Continuity

Continuity is a property of many functions. It is essential to understand this feature not just for the practice that it offers for working with the limit concept, but in addition, to appreciate the "limit pass through" feature it provides.

The property of continuity can be confusing to students new to the concept of the limit because it can be difficult to differentiate continuity from convergence.

Essentially, for a function to be continuous, two levels of convergence must take place. First, the function's argument must converge. Second, the function itself must converge.

Fortunately, like so many elementary concepts in calculus, a little thought provides a natural sense of the underlying idea. An intuitive understanding of continuous functions over a region is based on the sense that such functions contain no breaks as one moves across the $x$ axis. This is colloquially expressed as "The pencil need not be removed from the page" when drawing the function over a continuous region. While this is easy to grasp, we must invoke the limit concept to state this mathematically.

Let's begin with the assumption that, at the point $a$, $f(x) = f(a)$. How does the function actually get to $f(a)$? In drawing the function across the page, moving from left to right uninterruptedly along the $x$ axis, we anticipate that as we allow $x$ to get closer to $a$, $f(x)$ gets closer to $f(a)$. Thus, we have two limit processes. One is the (trivial) process of $x$ approaching $a$, and the other is that as $x$ approaches $a$, $f(x)$ approaches $f(a)$.

We can now state the formal criteria for continuity of a function. A function $f(x)$ is continuous at $f(a)$ if the $\lim_{x\to a} x = a$ implies $\lim_{x\to a} f(x) = f(a)$. This is simplified to the condition that a function $f(x)$ is continuous at $f(a)$ if $\lim_{x\to a} f(x) = f(a)$.

This dual limit phenomenon requires that we have two small arbitrary constants, $\varepsilon$, and $\delta$. With these constants in hand, demonstrating continuity of $f(x)$ at $f(a)$ translates the concept of "if $x$ is close to $a$, then $f(x)$ is close to $f(a)$" to if $|x - a| \leq \delta$ then $|f(x) - f(a)| \leq \varepsilon$.

In general, the practical demonstration of continuity requires us to first examine the inequality $|f(x) - f(a)| < \varepsilon$, trying to get this into a function of $|x - a|$, then use the fact that you can squeeze $|x - a|$ to be as small as possible to get $|f(x) - f(a)| < \varepsilon$.

For example, in order to demonstrate the continuity of the function $f(x) = kx$ for $k$ a positive, known constant, we start by assuming that $x$ is within $\delta$ of $a$. Then $|x - a| \leq \delta$. This implies that $|f(x) - f(a)| = |kx - ka| = k|x - a| \leq k\delta$, or $|f(x) - f(a)| \leq k\delta$. Regardless of the value of $k$, choosing $\delta$ small enough permits $\left|f(x)-f(a)\right|$ to be small. Setting $\varepsilon = k\delta$ finishes the demonstration. If we, for example would like $f(x)$ to be within 0.0001 of $f(a)$, we set $\delta = \frac{0.0001}{k}$.

This is an easy example; not all such continuity demonstrations are so simple. Sometimes demonstrating continuity can be a technical challenge, requiring intimate knowledge about $f(x)$.

However, one relieving factor with tremendous and helpful consequences is the notion of passing a functions through limits, i.e., the "limit pass through" feature.

When can we pass a limit through a function, i.e., when is $\lim_{x\to a} f(x) = f\left(\lim_{x\to a} x\right) = f(a)$? The answer is that this is true for continuous functions, as this is their very definition.

The property of reversing the function and limit sign adds tremendous power to our ability to manipulate limits. From our previous example, since $f(x)$ is continuous at $f(a)$, then $\lim_{x\to a} f(x) = \lim_{x\to a}(kx) = k\lim_{x\to a} x = ka = f(a)$, and we have the continuity of $f(x) = kx$ at $f(a)$ directly.

The demonstration that the sums, differences, products and quotients (with the monitory that the denominator must stay away from zero) of continuous functions are also continuous allows us to build families of continuous functions. For example, since sine and cosine are continuous functions, We know that the tangent is also continuous (as long as we stay away from $\frac{\pi}{2}$, where the cosine is zero).

## Pointwise versus uniform convergence

So far, our concern has concentrated on only the convergence of the function $f(x)$ with no thought at all about the rate at which the function changes as $x$ converges to $a$. However, "rates" do differ. Consider the elementary function $f(x) = x$ and the slightly more complicated $g(x) = x^2$.Do they both converge and if so, do they converge at the same rate.

Remember that if $f(x)$ is to converge to $f(a)$ as $x \to a$, then we must be assured that for $|x-a| \le \delta$, then $\left|f(x)-f(a)\right| \le \varepsilon$.

This is trivial for $f(x)$ since we only have to let our $\varepsilon = \delta$ to find that for $|x-a| \le \delta$, then $\left|f(x)-f(a)\right| = |x-a| = \varepsilon = \delta$.

Taking the same approach for $g(x)$, again, let $|x-a| \le \delta$. What does this imply about the distance between $g(x)$ and $g(a)$? Write $\left|g(x)-g(a)\right| = \left|x^2-a^2\right| = |x-a||x+a|$.

Now, focus on $|x+a|$. We know that as $x$ approaches $a$, the quantity $|x+a|$ approaches $2a$. Thus $\left|g(x)-g(a)\right| = \left|x^2-a^2\right| = |x-a||x+a| \le 2a\delta$. So by choosing $\varepsilon = 2a\delta$, we have $\left|f(x)-f(a)\right| = \left|x^2-a^2\right| \le \varepsilon$ and $f(x) = x^2$ is continuous at the point $x = a$.

However, is $g(x)$ as close to $g(a)$ as $f(x)$ is to $f(a)$ when $|x-a| \leq \delta$? For $a=1$, $2a\delta = 2\delta$, which means that $|g(x)-g(a)|$ is twice the distance as $|f(x)-f(a)|$. When $a=100$, then $|g(x)-g(a)|$ is two hundred times the distance $|f(x)-f(a)|$. These rates are clearly different; moreover, $f(x)$ has a constant rate of convergence, while $g(x)$'s rate is a function of $a$. We say that $f(x)$ doesn't just converge, but it converges uniformly. Alternatively, the function $g(x)$ converges in a pointwise manner.

As another example of pointwise convergence, consider the function sequence $\{f_n(x)\}$ where $f_n(x) = x^n$. Note that this is a function with two determining values. The first is *n*, which indexes the sequence of functions. The second is *x*, which is the argument of $f_n$. One could think of this as a sequence of functions indexed by *n* for each value of *x*.

What does convergence in this case mean?
For each *x* we have a sequence of functions, opening the door to the possibility that each sequence of functions may have different limiting behavior depending on the value of *x*. For fixed $n$, $\lim_{x \to a} x^n = a^n$. [*]

But what about $\lim_{n\to\infty} x^n$? For example, for any *x* > 1, $\lim_{n\to\infty} x^n$ does not exist, since $x^n$ is unbounded, growing to infinity.

For *x* = 1, our function $x^n = 1$ for all $n$. Similarly, for $x = 0$, $x^n = 0$ for $n \geq 1$.

Now, what is the behavior of $x^n$ for $0 < x < 1$? Based on the argument we made from the previous section, our sense suggests that for $0 < x < 1$, $\lim_{n\to\infty} x^n = 0$. How can we show this?

Challenged with an *ε* > 0, we need to find an *n** such that for all *n* > *n** $f_n(x) = x^n < \varepsilon$. We proceed as follows.

$$\begin{gathered} x^n < \varepsilon \\ n\ln(x) < \ln(\varepsilon) \\ -n\ln(x) > -\ln(\varepsilon) \\ n > n^* = \frac{-\ln(\varepsilon)}{-\ln(x)} \end{gathered}$$

And we have our result. However, notice that our *n** is a function of *x*. We might even write this as *n**(*x*). The fact that *n**(*x*) is a function of *x* is a hint to us that if convergence takes place, it takes place at different rates (Figure 2).

[*] To see this note that, to demonstrate the pointwise convergence of $f_3(x) = x^3$,
$|x^3 - a^3| = |x-a||x^2 + ax + a^2| < |x-a||3a^2| \leq 3a^2\delta$

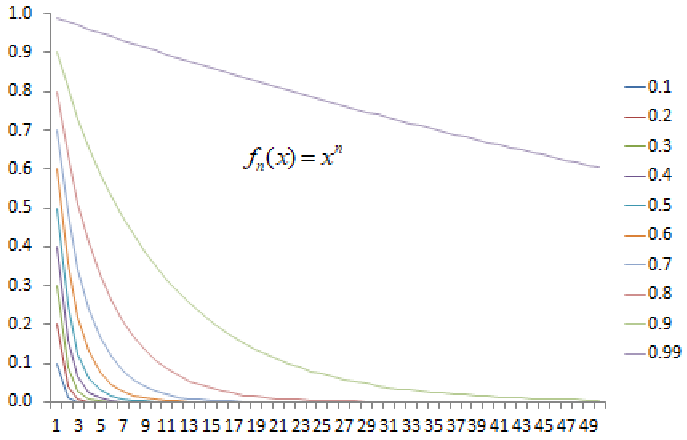


**Figure 2**. Different rates of convergence to zero for a pointwise convergent function.

Certainly each of the functions in Figure 2 converges to zero. However, they converge at difference rates. This is another example of pointwise convergence. since the rate of convergence is conditional on each "point" $x$.

Pointwise convergent functions can be a challenge to work with because we cannot be assured to have the function within $\varepsilon$ of the limit for all values of $x$.

This type of assurance is provided by functions which are uniformly convergent. A function that is uniformly convergent has the rate of convergence independent of $x$, i.e., $n*$ can be written so that it is not a function of $x$.

Commonly we can convert a pointwise convergent function to a uniformly convergent function on an interval of $x$. Consider the same function above, $f_n(x) = x^n$, except we now restrict ourselves to the region $0 < x < 0.9$. Then we modify our proof of convergence as follows;

$$\begin{aligned} x^n &< \varepsilon \\ n\ln(x) &< \ln(\varepsilon) \\ -n\ln(x) &> -\ln(\varepsilon) \\ n > n^* &= \frac{-\ln(\varepsilon)}{-\ln(0.90)} > \frac{-\ln(\varepsilon)}{-\ln(x)} \; . \end{aligned}$$

Which is true for all $0 \le x \le 0.90$.

However, since our proof holds only for $0 < x \le 0.90$, we say that the function is locally uniform convergent.

Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes
Exponential Functions
Differential Equations

The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini's Theorem

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Convergence and Limit Interchanges

Convergence is a critical feature of many functions. However, the different types of convergence and continuity with their implications can sometimes confuse the reader. Here is a brief summary.

**Passing the Limit through a Function**

$\lim\limits_{a_n \to a} f(a_n) = f\left(\lim\limits_{a_n \to a} a_n\right) = f(a)$. The limit and the function are interchangeable. This is a consequence of pointwise convergence and is the basis of the continuity concept.

**Passing a Limit through an Integral or Differential**

This is a property of the uniform convergence of the function $f(x)$ to $f(a)$. Note that the limit passes through both the integral (or derivative) and then again through the function

$$\lim_{a_n \to a} \int f(a_n) = \int \lim_{a_n \to a} f(a_n) = \int f\left(\lim_{a_n \to a} a_n\right) = \int f(a).$$

$$\lim_{a_n \to a} \frac{df(a_n)}{dx} = \frac{d \lim\limits_{a_n \to a} f(a_n)}{dx} = \frac{df\left(\lim\limits_{a_n \to a} a_n\right)}{dx} = \frac{df(a)}{dx}.$$

## Passing Limits Through Functions

One of the consequences of our definition of a limit is that in some circumstances, if a sequence converges, then a function of that sequence can also converge. Such a property would aid our understanding of functions profoundly. However, exactly what functions are these?

We began an examination of this concept in the section on continuity, and summarize here, phrasing the pass through argument somewhat differently. For a function $f$, we say that if $a_0, a_1, a_2, ..., a_n ...$ approaches the limit *a*, then another sequence $f(a_1), f(a_2), f(a_3)...., f(a_n)...$ approaches the limit $f(a)$.

A succinct way of putting this is that $\lim\limits_{n \to \infty} f(a_n) = f\left(\lim\limits_{n \to \infty} a_n\right)$, and we say that the limit "passes through" the function.

## Uniform convergence and continuity

With the notion of uniform convergence as background, we are now in a position to show two important features of continuous functions. The first is that if $f_n(x)$ is continuous and converges uniformly to $f(x)$, then $f$ must be continuous.

In order for $f(x)$ to be continuous at a point, say $x_0$ then for $|x-x_0|<\delta$, then $|f(x)-f(x_0)|<\varepsilon$. Given that we know $f_n(x)$ converges uniformly to $f(x)$ we know that there exists a single $N$ / for all $n>N$, $f_n(x)$ gets as close as we need to $f(x)$ and $f_n(x_0)$ to $f(x_0)$ for which $x$ is uniformly convergent (this is why we need uniform convergence. Thus, we can write

$$\begin{aligned}
&|f(x)-f(x_0)| = |\, f(x)-f_n(x)+f_n(x)-f_n(x_0)+f_n(x_0)-f(x_0)\,| \\
&= \left|\, (f(x)-f_n(x)) + (f_n(x)-f_n(x_0)) + (f_n(x_0)-f(x_0))\,\right| \\
&\le |f(x)-f_n(x)| + |f_n(x)-f_n(x_0)| + |f_n(x_0)-f(x_0)|.
\end{aligned}$$

Now we can write $|f(x)-f_n(x)|<\frac{\varepsilon}{3}$ and $|f_n(x_0)-f(x_0)|<\frac{\varepsilon}{3}$ by the uniform convergence of $f_n(x)$. We can also write $|f_n(x)-f_n(x_0)|<\frac{\varepsilon}{3}$ by the continuity of $f_n(x)$. Thus, we finish by writing, for $|x-x_0|<\delta$,

$$\begin{aligned}
&|f(x)-f(x_0)| = |\, f(x)-f_n(x)+f_n(x)-f_n(x_0)+f_n(x_0)-f(x_0)\,| \\
&= \left|\, (f(x)-f_n(x)) + (f_n(x)-f_n(x_0)) + (f_n(x_0)-f(x_0))\,\right| \\
&\le |f(x)-f_n(x)| + |f_n(x)-f_n(x_0)| + |f_n(x_0)-f(x_0)| \\
&\le \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3}=\varepsilon.
\end{aligned}$$

For an example of the impact of the absence of uniform convergence on the continuity of the limit, consider $f_n(x)=x^n$ on $[0,1]$. Note than on the interval $[0,1)$ we have $\lim_{n\to\infty} f_n(x)=0$. Also at the upper bound we find that $\lim_{n\to\infty} f_n(x)=1$. So the limit is $f(x)=0\,1_{0\le x<1}+1_{x=1}$. This function is clearly not continuous at $x=1$.

Now, if we redefine the function as $f_n(x)=x^n 1_{0\le x\le 0.99}$, we know from our earlier development that $f_n(x)$ is uniformly convergence (the upper bound confers the rate of convergence) and $\lim_{n\to\infty} f_n(x)=0$ for all $x$ on the defined semi-closed interval, is also continuous and therefore we have continuity of the limit function.

Thus, the property of continuity permits us to pass the limit through the function, i.e., $\lim_{x\to a} f(x) = f\left(\lim_{x\to a} x\right) = f(a)$. The continuity of $f(a)$ is assured if $f(x)$ is uniformly convergent

## Uniform convergence, integrals and derivatives

This will be a very useful property for us, since we will commonly take derivatives and integrals term by term in a series. If the series is uniformly convergent, then we can interchange the derivative and the infinite summation, and in the case of integration, interchange the integral and the summation.

Specifically, if $f_n(x)$ converges uniformly to a function $f(x)$, then

$$\lim_{n\to\infty}\int f(x_n)=\int\lim_{n\to\infty} f(x_n)=\int f\left(\lim_{n\to\infty} x_n\right)=\int f(x).$$

We can see a similar, measure-theoretic result in the development of each of the monotone convergence theorem and the Lebesgue Dominated Convergence Theorem.

A similar finding is available for differentiation. Again, if , $f_n(x)$ converges uniformly to a function $f(x)$, then

$$\lim_{n\to\infty}\frac{df(x_n)}{dx}=\frac{d\lim_{n\to\infty} f(x_n)}{dx}=\frac{df\left(\lim_{n\to\infty} x_n\right)}{dx}=\frac{df(x)}{dx}.$$

Interesting series that have these properties are probability generating functions and the exponential function,

$$\sum_{k=0}^{\infty}\frac{x^k}{k!}=e^x.$$

Thus, while continuity permits us to pass the limit through a function, uniform continuity permits us to pass the limit through an integral or through a differential.

Curve Slopes
Exponential Functions
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini’s Theorem

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions

Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability

Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

# Principles of Differential Calculus

The fundamental feature of calculus is the notion of the limit. Here we use that concept in the development of the idea of the derivative. This will be important in our discussions concerning several developments, most notably the Poisson process.

Prerequisite
Properties of Real Numbers
The Concept of the Limit
Convergent Series
Cauchy Sequences
Pointwise vs. Uniform Convergence

## Curve slopes

We have good intuition about the slope of a straight line. In general, if the line is described as $y = mx + b,$ them $m$ is the slope of the line. Given any two points on the line $(x_1, y_1)$ and $(x_2, y_2)$, we can compute the slope $m$ as

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x},$$

motivating the common descriptive phrase of $m$ as the rate of change of $y$ compared to the rate of change of $x$. Note that the use of this definition doesn't mandate which two points $(x_1, y_1)$ and $(x_2, y_2)$ are used. Any two points on the straight line can be used, because they each give the same answer.

The slope of a straight line is constant over the entire real line. It doesn't matter whether $\Delta y$ is large or small; as long as it is connected through the formula $\frac{\Delta y}{\Delta x}$ to the matched $\Delta x,$ the solution for $m$ will be the same and correct.

However, many curves do not have this property. The evaluation of a curve such as $y = x^2$ demonstrates a non-constant slope (Figure 1).

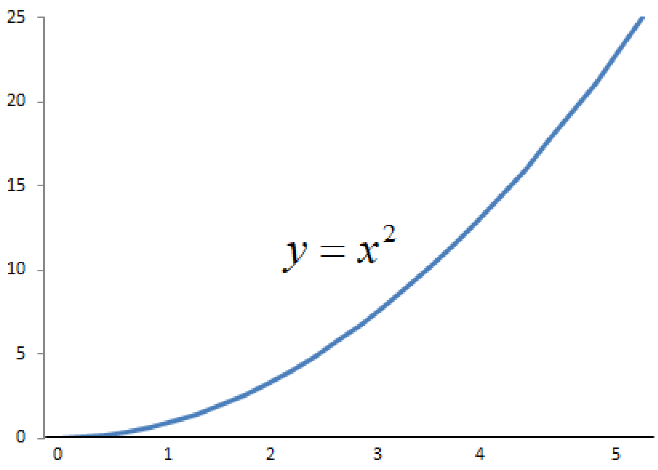


**Figure 1.** Demonstration of a nonconstant slope for the curve $y = x^2$

Note that for $x$ very close to zero, small increases in $x$ produce very small change in $y$. However, as $x$ increases, the increase in $y$ for the same change in $x$ increases. We can visualize this by overlaying the slope computed at each point (Figure 2).

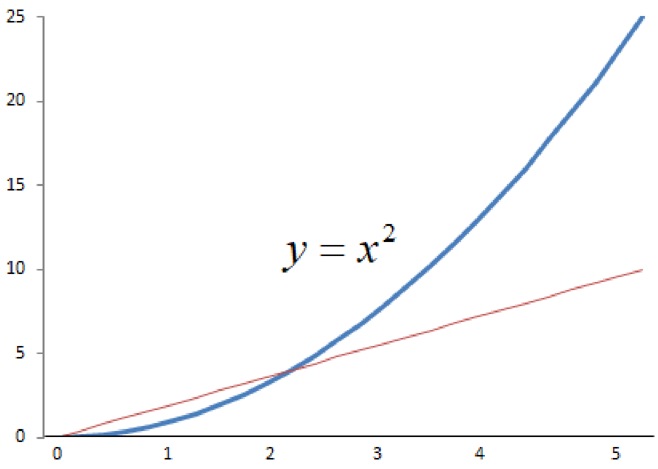


**Figure 2.** The curve $y = x^2$ and its slope.

For the straight line that we first considered, the slope is always the same and, since this slope is a constant, its rate of change is zero. However, for the curve $y = x^2$, the slope is now a function of $x$, and is always changing. In fact, each point $x$ generates a different slope for the curve $y = x^2$.

How can we compute this $x$-value specific slope, for example at the point $x_0$?

Since the slope changes as a function of $x$, we want to get as close to the point $x_0$ as possible before we compute our $\Delta y$. This suggest that the concept of the limit may serve us here. Let $h$ serve as our $\Delta x$. Then, we would like to let $h$ get as small as possible, approaching zero, and look at the quantity $\frac{(x_0+h)^2}{h}$. We define this instantaneous slope as the derivative *of y with respect to x* and for our function write

$$\left.\frac{dy}{dx}\right|_{x=x_0} = \lim_{h\to 0}\frac{(x_0+h)^2 - x_0^2}{h}$$

This is exactly the concept that we sought, i.e., to identify the change in $x^2$ at $x = x_0$ for the smallest possible changes in *x.* In fact, using what we know about limits we can compute this quantity. We begin by noting that $(x_0+h)^2 - x_0^2 = x_0^2 + 2x_0h + h^2 - x_0^2 = 2x_0h + h^2$, permitting us to write the original expression as

$$\left.\frac{dy}{dx}\right|_{x=x_0} = \lim_{h\to 0}\frac{(x_0+h)^2 - x_0^2}{h} = \lim_{h\to 0}\frac{2x_0h + h^2}{h} = \lim_{h\to 0}(2x_0+h) = 2x_0.$$

Thus, the slope is a function of $x_0$, as our intuition suggested. In general to find the derivative of a function $f(x)$ at the point $x = x_0$ we compute

$$\left.\frac{df(x)}{dx}\right|_{x=x_0} = \lim_{h\to 0}\frac{f(x_0+h) - f(x_0)}{h}.$$

There are two caveats to finding derivatives, each involving the properties of our function $f(x)$. The first is that $f(x)$ must be continuous. The second is that taking a derivative by allowing *x* to increases to $x_0$ must give the same value as taking a derivative by allowing *x* to decrease to $x_0$. The function provided in Figure 3 shows how a function can have derivatives in some regions and be missing its derivatives in others.

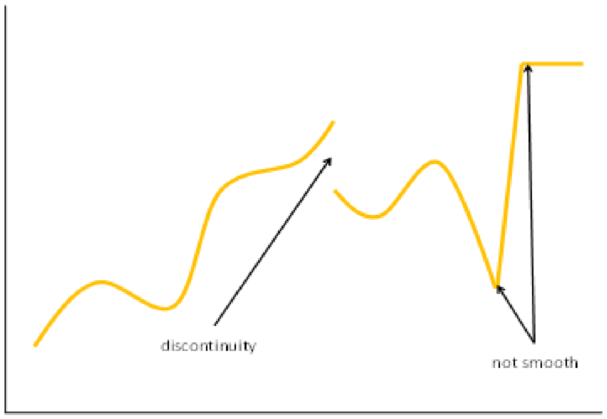


Figure 3. A function must be continuous and "smooth" to have derivatives at each of its points.

The lack of continuity means that one cannot take a limit of the function at the point of discontinuity, principally because $f\left(x_0+h\right)-f\left(x_0\right)$ has no limit at the points of discontinuity. The notion of smoothness is conveyed by the ability to compute the same value for the derivative approaching from the left as for approaching from the right. The simplest and most infamous of these functions with discontinuity is the function $y=|x|$ at $x=0$.

There are many formulas for computing derivatives, and after a time, like integration, "differentiating a function" at a particular point can become very mechanical for students. However, it is important to realize that one is taking advantage of important properties of the function in taking its derivative. Simply "taking the derivative" of a function when the derivative does not exist is a common pitfall.

## Exponential functions

Chief among the functions we will rely on is the exponential function. In the section on limits we defined this function $e^x$ as $e^x=\sum_{k=0}^{\infty}\frac{x^k}{k!}$.* Its derivative is simply $e^x dx$, which means that its antiderivative or integral is also $e^x$. We can compute more complicated derivatives of this function, relying on the adaptation of the chain rule $\frac{de^{g(x)}}{dx}=e^{g(x)}\frac{dg(x)}{dx}$. Thus we have

$$\frac{de^x}{dx}=e^x : \quad \frac{de^{-x}}{dx}=-e^{-x} : \quad \frac{de^{ax}}{dx}=ae^x : \quad \frac{de^{\frac{-x^2}{2}}}{dx}=-xe^{\frac{-x^2}{2}}.$$

Mathematics Review
Exponential Functions
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini's Theorem

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

* Here is an example of a irrational number that is the infinite sum of rational numbers.

Basic Probability Distributions

Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability

Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# The Mean Value Theorem

Prerequisites
Pointwise vs. Uniform Convergence
Convergence and Limit Interchanges
Passing Limits Through Functions
Uniform Convergence and Continuity
Uniform Convergence, Integrals and Derivatives
Curve Slopes

One of the most useful theorems from calculus provides the ability to compare the relative magnitude of a function at one point by its value at another point and its rate of change.

For example, if two functions have the same value at a particular point $x_0$, and, for every point in some interval around $x_0$, the rate of rise of one function is greater than the rate of rise of the other, then the function with the greater slope has greater values than the other. This seems self-evident, but its line of reasoning is quite valuable in calculus and analysis. It is this thought that is encapsulated in the mean value theorem. It is remarkable how such a natural conclusion is so useful in mathematics.

## Statement of the theorem

If a function $f(x)$ is continuous on an interval $\left[a,b\right]$ and differentiable on the open interval $\left(a,b\right)$ then at some point $x_0$, $a < x_0 < b$ the following equality holds:

$$f'(x_0) = \frac{f(b) - f(a)}{b-a}.$$

The important corollaries of the mean value theorem are

1) If $f'(x_0) > 0$ then $f(b) > f(a)$
2) If $f'(x_0) < 0$ then $f(b) < f(a)$
3) If $f'(x_0) = 0$ then $f(b) = f(a)$ *

These corollaries (which follow because the denominator $b-a$ is positive) are sometimes all that we use from the Mean Value Theorem.

## Implementation

* This is sometimes known as Rolle's Theorem.

For example, suppose we wish to show that $e^x > x+1$ for $x>0$. This is the same as showing that $e^x - x - 1 > 0$. We know that $f(0) = e^0 - 0 - 1 = 0$. The derivative $f'(x) = e^x - 1$ is positive on this interval. Since the derivative is positive on the interval, the function is increasing on the interval and for any $x$ in the interval $f(x) > f(0) = 0$. Thus, the function must be positive and $e^x > x+1$.

Similarly, we can show $1 - x \le e^{-x}$. This is equivalent to $e^{-x} + x - 1 = f(x) \ge 0$. $f(0) = e^{-0} - 1 + 0 = 0$, and $f'(x) = -e^{-x} + 1 = 1 - e^{-x} > 0$, satisfying the corollary of the mean value theorem.

Mathematics Review
Exponential Limit
The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini's Theorem

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential

Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Polar Coordinates

Prerequisite
Curve Slopes

In calculus we learn that there are many integrals (we will soon see that this is synonymous with measures) of the real line that may not be straightforward to evaluate. Using the tools of partial fractions and integration by parts, among others, we build up our repertoire of implements that help us with the actual integration.

Once useful tool in double integration is the use of polar coordinates. They are particularly useful when the integrand is a function of an exponent whose power is along the lines of $x^2 + y^2$.[*]

## Polar coordinate system

Polar coordinates are nothing more than an alternative way to map the Cartesian coordinate plane. We customarily think of this as the $(x, y)$ plane, recognizing that every point in this system can be identified by a unique pair of real numbers which we call the *x*-coordinate and *y*-coordinate. However, this is not the only assignment system that provides unique specification of the point. In polar coordinates, we assign to each point a pointer that begins at the origin of a specific length, *r*, and an angle measured from the *x*-axis, $\theta$. Not only can every point be specified uniquely in this matter, but the relationship between $(x, y)$, and $(r, \theta)$ is quite valuable (Figure 1).

[*] Grégoire de Saint-Vincent and Bonaventura Cavalieri independently introduced the concepts in the mid-seventeenth century. Saint-Vincent wrote about polar coordinates privately in 1625 and published his work in 1647, while Cavalieri published his in 1635 with a corrected version appearing in 1653. Cavalieri first used polar coordinates to solve a problem relating to the area within an Archimedean spiral. Blaise Pascal subsequently used polar coordinates to calculate the length of parabolic arcs (from Wikipedia).

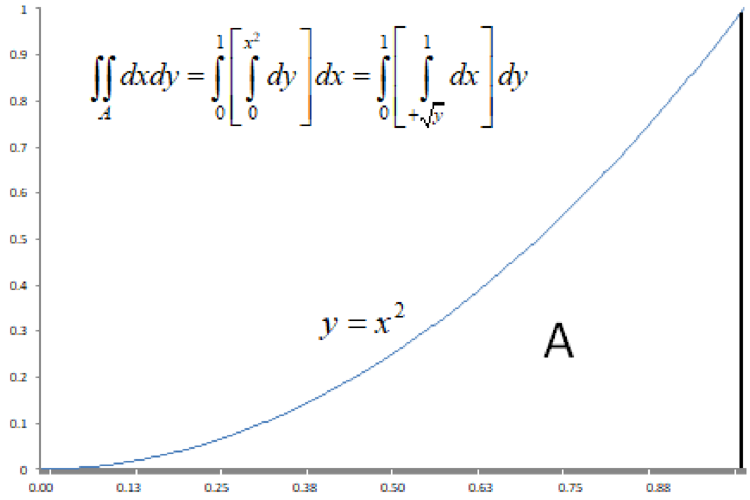


Figure 1. Example of the use of Fubini's theorem.

From this diagram we can see that $\cos(\theta)=\frac{x}{r}; \sin(\theta)=\frac{y}{r}$. This leads to the familiar parameterization commonly used in mapping the Cartesian plane to the polar plane, namely

$$x = r\cos(\theta);\ y = r\sin(\theta).$$

This tells us how to map $(r,\theta)$ to $(x,y)$. If we want to map the reverse, we see from figure 1 then $\theta=\tan^{-1}\left(\frac{y}{x}\right)$, and $x^2+y^2=r^2\cos^2(\theta)+r^2\sin^2(\theta)=r^2\left(\cos^2(\theta)+\sin^2(\theta)\right)=r^2$. Also, a simple use of Pythagorean's theorem reveals the same result. Thus we can map readily from Cartesian to the polar coordinate planes.

Mathematics Review
Exponential Limit
The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini's Theorem

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability

Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Exponential Limit

Prerequisites
Binomial Theorem
The Concept of the Limit
Convergent Series
Pointwise versus Uniform Convergence
Curve Slopes

We need to show

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$$

To proof this assertion, we first invoke the binomial formula

$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$ to see that $\left(1+\frac{x}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{x}{n}\right)^k$.

We can now write

$$\begin{aligned}\sum_{k=0}^{n}\binom{n}{k}\left(\frac{x}{n}\right)^k &= \sum_{k=0}^{n}\frac{n(n-1)(n-2)(n-3)...(n-k+1)}{n^k}\frac{x^k}{k!}\\ &= \sum_{k=0}^{n}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\left(\frac{n-3}{n}\right)...\left(\frac{n-k+1}{n}\right)\frac{x^k}{k!}\\ &= \sum_{k=0}^{n}(1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{3}{n}\right)...\left(1-\frac{k-1}{n}\right)\frac{x^k}{k!}\end{aligned}$$

Taking the limit we have

$$\begin{aligned}\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n &= \lim_{n\to\infty}\sum_{k=0}^{n}(1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{3}{n}\right)...\left(1-\frac{k-1}{n}\right)\frac{x^k}{k!}\\ &= \lim_{n\to\infty}\sum_{k=0}^{n}\frac{x^k}{k!}\\ &= e^x.\end{aligned}$$

Mathematics Review
Differential Equations
The Mean Value Theorem
Polar Coordinates
Exponential Limit

The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini’s Theorem

# Differential Equations

Prerequisite
Curve Slopes
Exponential Functions

Differential equations focus on solutions to problems that involve rates of change. Since these change rates are managed by calculus, it is no surprise that calculus plays an important role in solving these equations.

## Simple rate change equation

Begin with the equation

$$\frac{dy}{dt} = k$$

Here the rate of change is a constant. Without an attempt at an analytic solution, we would expect that the values of *y* are a constant multiple of *t*.

The analytic solution is straightforward

$$\begin{aligned}
&\frac{dy}{dt} = k \\
&dy = kdt \\
&\int dy = k\int dt \\
&y = kt + C
\end{aligned}$$

*C* is an arbitrary constant, which can be identified based on the boundary values of y when $t = 0$. If for example $y = 0$ when $t = 0$, then our final solution is $y = kt$. If on the other hand, $y = y_0 \neq 0$ when $t = 0$, then $C = y_0$ and the final solution is $y = kt + y_0$. The use of boundary values helps us to determine the specific solution to the problem after the general solution has been identified. This will be a recurring motif in our approach to difference-differential equations.

## Separability

We will now add a complication in the equation.

$$\frac{dy}{dt}=ky$$

Here the rate of change is not simply a constant, but a function of $y$. This tells us that as $y$ increases, the rate of change of $y$ increases. Let's try the same approach.

$$\begin{aligned}&\frac{dy}{dt}=ky\\&\frac{dy}{y}=kdt\\&\int\frac{dy}{y}=\int kdt\\&\ln y=kt+C\\&y=e^{kt+C}=e^{kt}e^{C}=C_1e^{kt}\end{aligned}$$

So the same approach, with the exponential being introduced through the integral $\int\frac{dy}{y}$ with resultant exponentiation. Note how we managed the constant. Since it is a multiplicative constant, we just defined a new constant $C_1=e^C$. We have considerable freedom in manipulating constants.

For the boundary condition $y=y_0$ when $t=0,$ we find $y_0=C_1e^{k(0)}=C_1$. Therefore our final solution is $y=y_0e^{kt}$.

What made this differential equation so easy to solve was that the terms involving $y$ and those involving $t$ could be separated. This concept of separability is a recurring theme.

Consider the equation $\frac{dy}{dx}=-\frac{x}{ye^{x^2}}$. We can proceed the same way.

$$\begin{aligned}&\frac{dy}{dx}=-\frac{x}{ye^{x^2}}\\&ydy=-\frac{x}{e^{x^2}}dx=-xe^{-x^2}dx.\end{aligned}$$

Now integrating,

$$\begin{aligned}&\int ydy=\int -xe^{-x^2}dx\\&\frac{y^2}{2}=\frac{1}{2}e^{-x^2}+C\\&y^2=e^{-x^2}+C_1.\end{aligned}$$

Using the boundary condition of $y=y_0$ when $x=0$ we find $y_0^2=1+C_1$ or $C_1=y_0^2-1$. Thus, our solution is $y^2=e^{-x^2}+y_0^2-1$ or $y=\pm\sqrt{e^{-x^2}+y_0^2-1}.$

## Partial differential equations

These require the same type of approach. Since we have rates of change that affect more than one variable, we try to isolate these rates of changes, carry out the simple integration involving these changes, and then use the conditions provided by the original conditions.

For example, a very useful tool for the solution of a partial differential equation that can be written in the form

$$P\frac{\partial F(x,y)}{\partial x}+Q\frac{\partial F(x,y)}{\partial y}=R,$$

Is the collection of equalities

$$\frac{dx}{P}=\frac{dy}{Q}=\frac{dF}{R}.$$

These are called auxillary equations. They are used to find a general solution for $F\left(x,y\right)$. What this general solution is found, we use the boundary condions of $F\left(x,y\right)$ to find its specific solution.

In order to examine this concept more specifically, we might use the auxillaruy equations $\frac{dx}{P}=\frac{dy}{Q}$ to find a constant $c_1\left(x,y\right)$. We can also use $\frac{dy}{Q}=\frac{dF}{R}$ to find another constant $c_2\left(x,y\right)$. We then write $c_2\left(x,y\right)=\Phi\left[c_1\left(x,y\right)\right]$.

For example, lets consider the partial differential equation $y^2\frac{\partial F(x,y)}{\partial x}+x^3\frac{\partial F(x,y)}{\partial y}=x^2y^3$.

Then $\frac{dx}{P}=\frac{dy}{Q}=\frac{dF}{R}$ reveals $\frac{dx}{y^2}=\frac{dy}{x^2}=\frac{dF}{x^2y^3}$. Using the first and second terms let's us calculate $\frac{dx}{y^2}=\frac{dy}{x^2}$. Proceeding

$$\begin{aligned}
&\frac{dx}{y^2}=\frac{dy}{x^2}\\
&x^2dx=y^2dy\\
&\int x^2dx=\int y^2dy\\
&\frac{x^3}{3}=\frac{y^3}{3}+c_a\\
&x^3-y^3=3c_a=c_1\left(x,y\right).
\end{aligned}$$

This is our first constant. Continuing with the second and third equaliy, we see

$$\frac{dy}{x^2} = \frac{dF}{x^2 y^3}$$
$$y^3 dy = dF$$
$$\int y^3 dy = \int dF$$
$$\frac{y^4}{4} + c = F(x,y)$$
$$c_2(x,y) = F - \frac{y^4}{4}.$$

So we can write

$$c_2(x,y) = F(x,y) - \frac{y^4}{4}$$
$$\Phi\left[c_1(x,y)\right] = \Phi\left[x^3 - y^3\right] = F(x,y) - \frac{y^4}{4}.$$

Now we use the boundary conditions for $F(x,y)$. For example if $F(0,y) = y$, then

For example if one has the equation

$$\frac{\partial \mathbf{G}_s(t)}{\partial t} = \upsilon s(s-1)\frac{\partial \mathbf{G}_s(t)}{\partial s}.$$

Then we can write this as

$$\frac{\partial \mathbf{G}_s(t)}{\partial t} - \upsilon s(s-1)\frac{\partial \mathbf{G}_s(t)}{\partial s} = 0,$$

and therefore write

We will use these equalities in combination with the state of the system at time $t$ =0 to identify the generating function $\mathbf{G}_s(t)$. Begin with $\frac{dt}{1} = \frac{d\mathbf{G}_s(t)}{0}$, or $d\mathbf{G}_s(t) = 0$, which implies $\mathbf{G}_s(t)$ is a constant. We write this constant $\Phi(c_1)$.

Next we work with $\frac{dt}{1} = \frac{ds}{-\upsilon s(s-1)}$;

$$\frac{dt}{1}=\frac{ds}{-\upsilon s(s-1)}$$

$$\frac{-\upsilon dt}{1}=\frac{ds}{s(s-1)}=\frac{ds}{s-1}-\frac{ds}{s}$$

$$\int -\upsilon dt=\int\frac{ds}{s-1}-\int\frac{ds}{s}$$

$$-\upsilon t=\ln\left(s-1\right)-\ln\left(s\right)=\ln\left(\frac{s-1}{s}\right)$$

Now we just manipulate the constant

$$-\upsilon t=\ln\left(\frac{s-1}{s}\right)+C_1$$

$$C_1=-\upsilon t+\ln\left(\frac{s}{s-1}\right)$$

$$C_2=\frac{s}{s-1}e^{-\upsilon t}$$

We can now write $\mathbf{G}_s(t)=\Phi\left(c_1\right)=\Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right)$, and invoke the boundary condition for clarification of the form of the function $\Phi$. It we know that, for example $\mathbf{G}_s\left(0\right)=a$ we can therefore write

$$\mathbf{G}_s(0)=s^a=\Phi\left(e^{-\upsilon(0)}\frac{s}{s-1}\right)=\Phi\left(\frac{s}{s-1}\right)$$

Let $z=\frac{s}{s-1}$. Then $s=\frac{z}{z-1}$, and we find $\Phi\left(z\right)=\left(\frac{z}{z-1}\right)^a$ for $t$ = 0. For any other value of $t$ we have $\mathbf{G}_s(t)=\Phi\left(e^{-\upsilon t}\frac{s}{s-1}\right)=\left[\frac{e^{-\upsilon t}\frac{s}{s-1}}{e^{-\upsilon t}\frac{s}{s-1}-1}\right]^a$

Simplifying, we find.

$$\left[\frac{e^{-\upsilon t}\frac{s}{s-1}}{e^{-\upsilon t}\frac{s}{s-1}-1}\right]^a=\left[\frac{se^{-\upsilon t}}{se^{-\upsilon t}-(s-1)}\right]^a=\left[\frac{se^{-\upsilon t}}{1-s\left(1-e^{-\upsilon t}\right)}\right]^a$$

Thus

$$\mathbf{G}_s(t)=\left[\frac{se^{-\upsilon t}}{1-s\left(1-e^{-\upsilon t}\right)}\right]^a.$$

Mathematics Review
The Mean Value Theorem
Polar Coordinates
Exponential Limit
The Exponential and Gamma Functions
Integration of Exponential Families
Integration by Parts
Gamma Function
Fubini's Theorem

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion
Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure

Ordering Random Variables
Asymptotics
Tail Event Measure

# The Exponential and Gamma Functions

Prerequisites
Curve Slopes
Exponential Functions

The exponential and gamma functions are operations that play a necessary role in probability. The exponential function is critical in the Poisson process, and the processes related to it (emigration, contagion, death and the more advanced models) as well as for the Weibul, exponential, gamma, and normal distributions, just to name a few. The gamma function appears commonly and is used explicitly in the beta distribution, and the gamma distribution. Both the distributions have their own probability functions.

## Integration of exponential functions

The exponential function we have seen before in our brief introduction to in the discussion of derivatives. Using this as a background, we can begin some straightforward work with integration. For example, from this discussion, we know $\int e^x dx = e^x$, and $\int_a^b e^x dx = e^b - e^a$. Paying attention to signs produces $\int e^{-x} dx = -e^{-x}$, and therefore $\int_a^b e^{-x} dx = e^{-a} - e^{-b}$. An adaptation of this and a useful result for us is $\int_a^\infty e^{-x} dx = e^{-a}$, and of course $\int_0^b e^{-x} dx = 1 - e^{-b}$.

It is just a short and simple step to find the $\int ae^{ax} dx$. If we think of $u = ax$, then $du = adx$, and now rewrite $\int ae^{ax} dx = \int e^{ax} adx$. We see that we can now write this as $\int e^u du = e^u = e^{ax}$.

This gives us a clue for how to manage $\int e^{ax} dx$. We simply say $\int e^{ax} dx = \frac{1}{a}\int ae^{ax} dx = \frac{1}{a}\int e^{ax} adx = \frac{1}{a}\int e^u du = \frac{1}{a}e^u = \frac{1}{a}e^{ax}$. This ability to multiply by a constant to obtain a simpler integrand will be essential for us.

In the same vein we can find $\int xe^{\frac{-x^2}{2}}dx$ by recognizing that if $u=-\frac{x^2}{2}$, then $du=-x.$ We can now write

$$\int xe^{\frac{-x^2}{2}}dx=-\int -xe^{\frac{-x^2}{2}}dx=-\int e^{\frac{-x^2}{2}}\left(-xdx\right)=-\int e^u du=-e^u=-e^{\frac{-x^2}{2}}.$$

In order to carry out integrations such as $\int xe^{-x}dx$ we need to develop practice with the concept of integration by parts.

## Integration by parts

A useful tool in integration is to deconstruct the integrand and evaluate it piecemeal, using the finding that $\int udv=\int uv-\int vdu,$ where we are using the shorthand $u=u(x)$ and $v=v(x).$

The art to using this tool is seeing an integrand $f(x)=u(x)dv(x)$ in such a way that integration by parts is helpful. Sometimes it is so easy that it is difficult to see it at first blush for example, consider $\int \ln(x)dx.$ We write

$$u=\ln(x):du=\frac{dx}{x}:\ dv=dx:v=x$$

and therefore

$$\int \ln(x)dx=x\ln(x)-\int\frac{dx}{x}x=x\ln(x)-x.$$

For another example, consider $\int xe^{-x}dx.$ Consider

$$u=x:du=dx:\ dv=e^{-x}dx:v=-e^{-x}.$$

Now write

$$\int xe^{-x}dx=-xe^{-x}-\int -e^{-x}dx=-xe^{-x}-e^{-x}=-e^{-x}\left(x+1\right).$$

To compute $\int x^2e^{-x}dx,$ we take advantage of the previous problem, choosing

$$u=x:du=dx:\ dv=xe^{-x}dx:v=-e^{-x}(x+1).$$

We can now compute

$$\begin{aligned}\int x^2e^{-x}dx&=-xe^{-x}(x+1)-\int -e^{-x}(x+1)\\&=-xe^{-x}(x+1)+\int xe^{-x}dx+\int e^{-x}dx\\&=-xe^{-x}(x+1)-e^{-x}(x+1)-e^{-x}\\&=-e^{-x}\left[(x+1)x+(x+1)+1\right]\\&=-\left(x^2+2x+2\right)e^{-x}.\end{aligned}$$

## Gamma functions

The gamma function is an essential function for us in probability. It is defined as

$$\Gamma(r)=\int_0^\infty x^{r-1}e^{-x}dx.$$

Our experience in this section provides some insight into how we can solve this. For example

$$\Gamma(1)=\int_0^\infty x^0e^{-x}dx=\int_0^\infty e^{-x}dx=-e^{-x}\Big|_0^\infty=-\left[e^{-\infty}-e^0\right]=-[0-1]=1.$$

$$\Gamma(2)=\int_0^\infty xe^{-x}dx=-e^{-x}\left(x+1\right)\Big|_0^\infty=-[0-1]=1.$$

Further development makes use of the observation that $\Gamma\left(r+1\right)=r\Gamma(r)$, a result we now prove.

Let's integrate $\Gamma(r+1)=\int_0^\infty x^re^{-x}dx$ by parts. We let

$$u=x^r:\ du=rx^{r-1}:\quad dv=e^{-x}dx:v=\int e^{-x}dx=-e^{-x}.$$

Thus

$$\Gamma(r+1)=\int_0^\infty x^re^{-x}dx=uv\Big|_0^\infty-\int_0^\infty vdu\ =-x^re^{-x}\Big|_0^\infty+r\int_0^\infty x^{r-1}e^{-x}dx\ =r\Gamma(r).$$

For integer $k=r$ this reduces to $\Gamma(k+1)=k!$.

As another example, $\Gamma\left(\frac{1}{2}\right)=\int_0^\infty x^{-\frac{1}{2}}e^{-x}dx.$ Let $x=y^2$ and carry out a change of variables. The region of integration is 1 to 1. $x^{-\frac{1}{2}}=\left(y^2\right)^{-\frac{1}{2}}=y^{-1}.\ dx=2ydy.$ Thus $\int_0^\infty x^{-\frac{1}{2}}e^{-x}dx=\int_0^\infty y^{-1}e^{-y^2}2ydy=2\int_0^\infty e^{-y^2}dy.$ If we define $A=\int_0^\infty e^{-y^2}dy,$ then, using Fubini's theorem named after the Italian mathematician Guido Fubini, we find $A^2=\int_0^\infty e^{-y^2}dy\int_0^\infty e^{-x^2}dx=\int_0^\infty\int_0^\infty e^{-\left(x^2+y^2\right)}dx\,dy.$ Implementing polar coordinates, we see

$$\int_0^\infty\int_0^\infty e^{-\left(x^2+y^2\right)}dx\,dy=\int_0^{\frac{\pi}{2}}\int_0^\infty re^{-r^2}dr\,d\theta=\frac{\pi}{4}.$$ and $A=\sqrt{\frac{\pi}{4}}.$ Thus $\Gamma\left(\frac{1}{2}\right)=2A=2\sqrt{\frac{\pi}{4}}=\sqrt{\pi}.$

# Fubini’s Theorem

One of the most useful procedures in advanced integration is the process of obtaining the measure over a region or volume of space. Attributed to Guido Fubini, this process provides a pathway to compute multiple integrals. It allows one to transform a complicated integral such as the normal measuring tool into a tractable iterated integral.

Prerequisite
Curve Slopes
Exponential Limit

Fubini’s Theorem says that if a function is measurable or integrable over a closed region, then the double integral $\iint_A f(x,y)$ can be evaluated by a succession of single integrals. If for example the region A can be defined as $a \le x \le b; c \le y \le d$, then Fubini’s theorem states that $\iint_A f(x,y) = \int_c^c \left[ \int_a^b f(x,y)dx \right] dy,$ where the inner integral is carried out first. An implication of this is that the double integral can also be written as $\iint_A f(x,y) = \int_a^b \left[ \int_c^d f(x,y)dy \right] dx.$ Thus the user has a choice of the sequence of integration that they wish to carry out.

For example, suppose we wish to find the area A (Figure 1) bounded by the curve $y = x^2$ and the lines $x = 1,$ and $y = 1$.

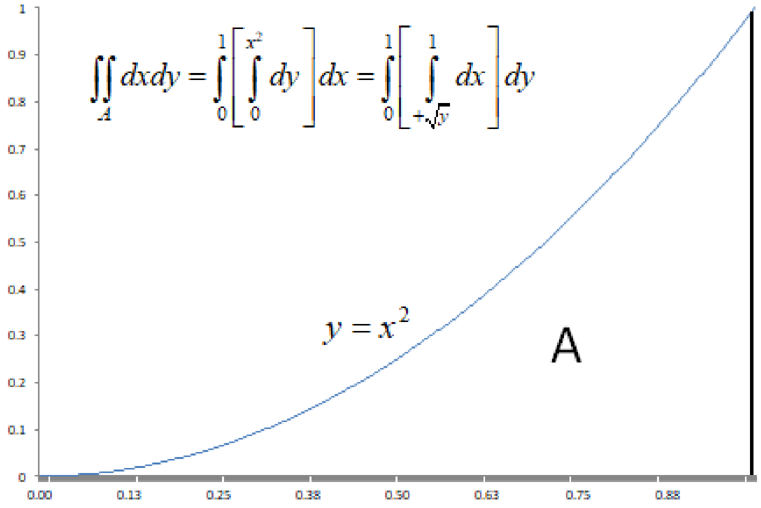


Figure 1. Example of the use of Fubini's theorem.

We can compute $\iint_A dxdy = \int_0^1\left[\int_0^{x^2} dy\right]dx = \int_0^1 x^2 dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}$. However, we can also write $\iint_A dxdy = \int_0^1\left[\int_{+\sqrt{y}}^1 dx\right]dy = \int_0^1\left(1-y^{\frac{1}{2}}\right)dy$. Continuing

$\int_0^1\left(1-y^{\frac{1}{2}}\right)dy = \left[y-\frac{2}{3}y^{\frac{3}{2}}\right]_0^1 = \frac{1}{3}$. Thus, the order is really a matter of convenience.

Measure
An Introduction to the Concept of Measure
Set Functions in Measure Theory
Simple Functions in Public Health
Measure and its Properties
Working with Measure

Probability Foundations
Elementary Set Theory
Basic Properties of Probability
Counting Events
Properties of Real Numbers
An Introduction to the Concept of Measure

Basic Probability Distributions
Basics of Bernoulli Trials.
Basics of the Binomial Distribution
Basics of the Poisson Distribution
Basics of Normal Measure

Advanced Probability
Bernoulli Distribution – In Depth Discussion

Advanced Binomial Distribution
Multinomial Distribution
Hypergeometric Measure
Geometric and Negative binomial measures
General Poisson Process
Survival Measure: Exponential, Gamma, and Related
Cauchy, Laplace, and Double Exponential
Continuous Probability Measure
Moment and Probability Generating Functions
Variable Transformations
Uniform and Beta Measure
Normal Measure
Compounding
F and T Measure
Ordering Random Variables
Asymptotics
Tail Event Measure

# Blaise Pascal

One of the founding fathers of probability is Blaise Pascal

He was born in Clermont on June 19, 1623 into a home where the father, Ėtieene Pascal was an established mathematician in his own right and focused on the intellectual development of all of his children.[1] In fact he was so devoted to Blaise that he, recognizing the abilities of his son, left his position at the Président à la Cour de Aides for Paris in 1631 to focus on Blaise.

After Ėtieene began to remain at home, the home life of the family changed. Pascal's mother died when he was a toddler, and Pascal became very close to his sisters Gilberte and Jaqueline [2].

Pascal was kept at home in order to ensure his not being overworked, and following his father's directive that his education first focus on languages.

Notably young Pascal was prohibited by his father from doing any work in mathematics. This was because his father, recognized his son's quantitative capabilities and avid interest in mathematics, feared that Pascal would be so absorbed in the field that he would ignore other fields.

However, this unusual restriction served only to spark Pascal's curiosity, and at twelve, he asked his tutor about geometry. When he learned what geometry, with its emphasis on figure construction and figure relationship entailed, he gave up his leisure time to focus on these new concepts.

A few weeks later, he had discovered for himself many properties of figures, and in particular the proposition that the sum of the angles of a triangle is equal to two right angles. His father, learning of Pascal's new initiative, abandoned his mathematics constraint, and give him a copy of Euclid's Elements, a book which Pascal absorbed [3]

Pascal's efforts produced some immediate accomplishments Before he turned thirteen, he had proven the 32nd proposition of Euclid. He discovered an error in Rene Descartes geometry. At the age of fourteen he was admitted to the weekly meetings of Roberval, Mersenne, Mydorge, and other French geometricians . At sixteen Pascal wrote an essay on conic sections and began to prepare a treatise on the entire field of mathematics.

However his father now required him to instead compute the sums of long columns of numbers by hand. Pascal responded eagerly by designing a calculating machine, which he perfected when he was thirty. Named the "Pascaline", it was the first accurate mechanical calculator created.*

In a strange turn, in 1650, when Pascal was 27 and in the midst of these researches, he suddenly abandoned his favorite pursuits to study religion, or, as he says in his Pensées, "contemplate the greatness and the misery of man"; and about the same time he persuaded the younger of his two sisters to enter the Port Royal society. In 1646, he and his sister Jacqueline identified with the religious movement within Catholicism known by its detractors as Jansenism. Following a religious experience in late 1654, he began writing influential works on philosophy and theology. His two most famous works include the Lettres provinciales and the Pensées, the former set in the conflict between Jansenists and Jesuits. [4]

In that year, he also wrote an important treatise on the arithmetical triangle.

It was during this time, that he began the many letters that he and Pierre de Fermat exchanged, a correspondence that began in 1654. It was during this correspondence that Pascal came to the observation that there was a fixed frequency at which the sides of die appeared. This finding served as a basis for modern probability.

Pascal conducted this while working on the concept that barometric pressure was in reality the weight of the atmosphere. He also confirmed his theory of the cause of barometrical variations by obtaining at the same instant readings at different altitudes on the hill of Puy-de-Dôme.

In the 1650's, Pascal focused on developing a perpetual motion machine. After stumbling through an accidental invention, he developed and demonstrated the roulette (little wheel) machine.[2]

Three years later, he began administer his father's estate, and began considering marriage when an accident turned his attention back to religion. While driving a cart the horses kicked free and ran away, Pascal saved only by the traces breaking. Always somewhat of a mystic, full of dismay and disgust by society's reactions to his inventions, he considered the accident a mandate from God that he leave the material word. He wrote an account of the accident on a small piece of parchment, which for the rest of his life he wore next to his heart, to perpetually remind him of his covenant.

Now completely renouncing his interests in science and mathematics, he planned to devote the rest of his life to God. During this time, he produced a collection of spiritual essays, Les Pensées. and shortly moved to Port Royal, where he retired.

In a time that should have been consumed by rest and reflection, he was plagued by sleeplessness and a toothache. Suddenly he was consumed with the cycloid. This was a new figure to geometry, its shaped determined by tracing a point on the circumference of a moving circle. To his astonishment, his teeth ceased hurting at once. Regarding this as a divine intervention, he threw himself into the problem, working ceaselessly for eight days, ending with a defensible account of the cycloid's geometry. His work developed new questions, and he held a contest challenging fellow mathematicians to enter and compete with their answers. [5]

However, he had inflicted irreparable damage to his health by his incessant study, and, while no one is certain, he is believed to have died of a brain hemorrhage when he was 39 years old, on August 19, 1662 .His writings on probability were published posthumously.

Why Probability

* The Pacaline was not a commercial success in Pascal's lifetime because by doing the work of six accountants, contemporaries felt it would create unemployment.

From Whence it Came – An Early History of Probability
Probability and the Renaissance
Blaise Pascal
Pierre Fermat
Abraham de Moivre
Famous Correspondence between Pascal and Fermat
Simon Laplace
Bayes and Price
The Inversion Problem

References

[1] . D. E. Smith. History of Mathematics. Volume 1. New York. Dover Publishing Company
[2] https://www.biography.com/scholar/blaise-pascal last accessed 3/17/2020
[3] . http://www.thocp.net/biographies/pascal_blaise.html last accesed 3/17/2020

[4] https://en.wikipedia.org/wiki/Blaise_Pascal last accessed 3/17/2020.
[5] http://www.ms.uky.edu/~carl/ma330/project2/whitt21.html last accessed 3/17/2020

# Pierre Fermat

.

One of the most eminent mathematicians of his time, Pierre de Fermat was born near Montauban in 1601, and died at Castres on January 12, 1665. The son of a leather-merchant, he was educated at home.

At thirty, he obtained the post of counselor for the local parliament at Toulouse, discharging those duties with careful attention to detail. It was a job that provided leisure time, and he devoted that time to mathematics.

Lacking formal training in mathematics, Fermat did not follow established norms for promulgating his work. Except for a few isolated papers, Fermat published nothing in his lifetime. He gave no systematic exposition of his methods. In fact, some of the most striking of his results were found after his death on loose sheets of paper or written in the margins of works which he had read and annotated, and are unaccompanied by any proof. He was by constitution modest and retiring, and does not seem to have intended his papers to be published. It is probable that he revised his notes as occasion required, and that his published works represent the final form of his researches, and therefore cannot be dated much earlier than 1660.

The theory of numbers appears to have been his favorite field of study. Fermat prepared an edition of Diophantus, and the notes and comments contain many elegant theorems. Unfortunately, only a handful of the proofs survived, a testament to the fact that they simply were not rigorous, and that Fermat saw their veracity in his head, perhaps by analogy or induction.

One of the most enigmatic of this mathematician's work is Fermat's Last Theorem, which states that for any integer $n > 2$, there are no integers *a*, *b*, and *c*, such that $a^n + b^n = c^n$. [1] Fermat famously stated that he had a proof, but that it would not fit in the margin of the copy of *Arithmetica*. A rigorous proof of this surprising assertion by Andrew Wiles, published in 1995. [2]

Fermat did not seek controversy. When he found himself in a vitriolic dispute with Descartes, he resolved it with tact and courtesy, bringing the matter to a friendly conclusion. This is seen in his famous letters with Pascal are gentle and encouraging.

Why Probability<br>
From Whence it Came – An Early History of Probability<br>
Probability and the Renaissance<br>
Also…<br>
Blaise Pascal<br>
Pierre Fermat<br>
Abraham de Moivre<br>
Famous Correspondence between Pascal and Fermat<br>
Simon Laplace<br>
Bayes and Price

The Inversion Problem

[1] https://en.wikipedia.org/wiki/Fermat%27s_Last_Theorem
[2] https://www.math.wisc.edu/~boston/869.pdf

# Abraham de Moivre

A friend of Isaac Newton, Edmund Halley, and James Sterling, Abraham de Moivre was a French mathematician famous for de Moiré's formula, which links complex numbers and trigonometry, and for his work on normal measure and probability theory.[*] His book on probability, the Doctrine of Chances, is said to have been prized by gamblers. Perhaps most importantly for the time, de Moivre's ceaseless agitation for acceptance of probability as an elevated view of the workings of the contemporary world and not as a mere gambler's crutch, brought the field much needed respectability.

Abraham de Moivre was born in Vitry in Champagne on May 26, 1667. His father, Daniel de Moivre, a surgeon, believed in the value of education, and though Abraham de Moivre's parents were Protestant, he first attended Christian Brothers' Catholic school in Vitry, which was unusually tolerant of his appearance given religious tensions in France at the time. Although mathematics was not part of his course work, de Moivre read several mathematical works on his own including Elements de mathematiques by Father Prestet and a short treatise on games of chance, De Ratiociniis in Ludo Aleae, by Christiaan Huygens. In 1684 he moved to Paris to study physics and for the first time had formal mathematics training with private lessons from Jacques Ozanam.

However, the world around him was in turmoil. Religious persecution in France became severe when King Louis XIV issued the Edict of Fontainebleau in 1685, revoking the Edict of Nantes and the substantial rights afforded to French Protestants. It forbade Protestant worship and required that all children be baptized by Catholic priests. De Moivre was sent to the Prieure de Saint-Martin, a school the authorities sent Protestant children to for indoctrination into Catholicism. It is unclear when de Moivre left the Prieure de Saint-Martin and moved to England, as the records of the Prieure de Saint-Martin indicate that he left the school in 1688. However, while still in his native France, he was imprisoned for being a Protestant. When released three years later, he and his brother emigrated to England to escape religious prosecution, presenting themselves as Huguenots admitted to the Savoy Church in London on August 28, 1687. De Moivre never returned to France and never again published anything in French.

By the time he arrived in London, de Moivre was a competent mathematician with a good knowledge of many of the standard texts. To make a living, de Moivre became a private tutor of mathematics, visiting his pupils or teaching in the coffee houses of London. De Moivre continued his studies of mathematics after visiting the Earl of Devonshire, discovering Newton's recent book, Principia there. Recognizing the text as a major work, he resolved to read and understand it. However, as he was required to take extended, time consuming walks around London to see his students, reducing his time for study, de Moivre would tear pages from the

[*] Adapted from An Introduction to Mathematical Statistics and its Applications by Richard J. Larsen and Morris L. Marx and
http://www.swlearning.com/quant/kohler/stat/biographical_sketches/bio8.2.html

famous book and carry them around in his pocket to read between lessons. Eventually de Moivre become so knowledgeable about the material that Newton referred questions to him, saying, "Go to Mr. de Moivre; he knows these things better than I do."

By 1692, de Moivre became friends with Edmond Halley and soon after with Isaac Newton himself. In 1695, Halley communicated de Moivre's first mathematics paper, which arose from his study of fluxions in the Principia, to the Royal Society. This paper was published in the Philosophical Transactions that same year. Shortly after publishing this paper, de Moivre also generalized Newton's famous binomial theorem into the multinomial theorem. The Royal Society became apprised of this method in 1697 and made de Moivre a member two months later.

In 1712, De Moivre was appointed to a commission set up by the society, alongside MM. Arbuthnot, Hill, Halley, Jones, Machin, Burnet, Robarts, Bonet, Aston and Taylor to review the claims of Newton and Leibniz as to who discovered calculus.

After de Moivre had been accepted, Halley encouraged him to turn his attention to astronomy. In 1705, de Moivre discovered, intuitively, that "the centripetal force of any planet is directly related to its distance from the center of the forces and reciprocally related to the product of the diameter of the evolute and the cube of the perpendicular on the tangent". Johann Bernoulli proved De Moivre's formula in 1710.

Although his only source of income was from tutoring students and advising gamblers and speculators, de Moivre's successful solutions to the problems he met in his consulting practice led to his writing a major work. His text on probability, The Doctrine of Chances, emanated from an article first published in Latin in 1711 and was published posthumously in its final and third edition in 1756, it expanded the work of his predecessors, particularly Christiaan Huygens and several members of the Bernoulli family.

It is notable (among many other contributions) for the origin of the general laws of addition and multiplication of probabilities, for the origin of the binomial distribution law, and for the origin of the formula for the normal curve, which de Moivre discovered in 1733. This book came out in four editions, 1711 in Latin, and 1718, 1738 and 1756 in English. In the later editions of his book, de Moivre gives the first statement of the formula for normal measure curve, the first method of finding the probability of the occurrence of an error of a given size when that error is expressed in terms of the variability of the distribution as a unit, and the first identification of the probable error calculation. Additionally, he applied these theories to gambling problems and actuarial tables.

De Moivre's second major publication, A Treatise of Annuities on Lives, was published in 1752. In it, he revealed normal measure of the mortality rate over a person's age. From this he produced a simple formula for approximating the revenue produced by annual payments based on a person's age. It was highly original and laid foundations for the mathematics of life insurance. Also, his effort reflects De Moivre's desire to free the science of probability from its connection with gambling, and also to establish a connection between probability and theology, necessary to get the support of the thought leaders of the day for probability as a respectable theory.

Despite these successes, and his development of analytic geometry, de Moivre was unable to obtain an appointment to a Chair of Mathematics at a university. At least a part of the reason was a bias against his French origins which would have released him from his dependence on time-consuming tutoring that burdened him more than it did most other mathematicians of the time.

Throughout his life de Moivre remained poor. It is reported that he was a regular customer of Slaughter's Coffee House, St. Martin's Lane at Cranbourn Street, where he earned a little money from playing chess. Yet he was ever ebullient, always anxious for conversation.

De Moivre continued studying the fields of probability and mathematics. However, as he grew older, he became increasingly lethargic and needed longer time asleep. Observing that he was sleeping an extra fifteen minutes each night, he calculated the date of his death on the day when the additional sleep time accumulated to 24 hours.

November 27, 1754.

This was the day he died. He was initially buried at St Martin-in-the-Fields, although his body was later moved.

Why Probability
From Whence it Came – An Early History of Probability
Probability and the Renaissance

Also…

Blaise Pascal
Pierre Fermat
Abraham de Moivre
Famous Correspondence between Pascal and Fermat
Simon Laplace
Bayes and Price
The Inversion Problem

# Pascal Fermat Correspondence

FERMAT AND PASCAL ON PROBABILITY[*] Italian writers of the fifteenth and sixteenth centuries, notably Pacioli (1494), Tartaglia (1556), and Cardan (1545), had discussed the problem of the division of a stake between two players whose game was interrupted before its close. The problem was proposed to Pascal and Fermat, probably in 1654, by the Chevalier de M´er´e, a gambler who is said to have had unusual ability "even for the mathematics." The correspondence which ensued between Fermat and Pascal, was fundamental in the development of modern concepts of probability, and it is unfortunate that the introductory letter from Pascal to Fermat is no longer extant. The one here translated, written in 1654, appears in the OEuvres de Fermat (ed. Tannery and Henry, Vol. II, pp.288–314, Paris 1894) and serves to show the nature of the problem.

[*] From http://www.socsci.uci.edu/~bskyrms/bio/readings/pascal_fermat.pdf. All but the last two letters were translated from the French by Professor Vera Sanford, Western Reserve University, Cleveland, Ohio, and appear in A Source Book in Mathematics (ed. D E Smith). The last two were translated by by Maxine Merrington and appear in Games, Gods and Gambling by F N David.

Fermat to Pascal 1654 [undated]

Monsieur

If I undertake to make a point with a single die in eight throws, and if we agree after the money is put at stake, that I shall not cast the first throw, it is necessary by my theory that I take 1/6 of the total sum to be impartial because of the aforesaid first throw.

And if we agree after that that I shall not play the second throw, I should, for my share, take the sixth of the remainder that is 5/36 of the total. If, after that, we agree that I shall not play the third throw, I should to recoup myself, take 1/6 of the remainder which is 25/216 of the total.

And if subsequently, we agree again that I shall not cast the fourth throw, I should take 1/6 of the remainder or 125/1296 of the total, and I agree with you that that is the value of the fourth throw supposing that one has already made the preceding plays.

But you proposed in the last example in your letter (I quote your very terms) that if I undertake to find the six in eight throws and if I have thrown three times without getting it, and if my opponent proposes that I should not play the fourth time, and if he wishes me to be justly treated, it is proper that I have 125/1296 of the entire sum of our wagers.

This, however, is not true by my theory. For in this case, the three first throws having gained nothing for the player who holds the die, the total sum thus remaining at stake, he who holds the die and who agrees to not play his fourth throw should take 1/6 as his reward.

And if he has played four throws without finding the desired point and if they agree that he shall not play the fifth time, he will, nevertheless, have 1/6 of the total for his share.

Since the whole sum stays in play it not only follows from the theory, but it is indeed common sense that each throw should be of equal value. I urge you therefore (to write me) that I may know whether we agree in the theory, as I believe (we do), or whether we differ only in its application. I am, most heartily, etc., Fermat.

Pascal to Fermat Wednesday, July 29, 1654

Monsieur,—

Impatience has seized me as well as it has you, and although I am still abed, I cannot refrain from telling you that I received your letter in regard to the problem of the points 1 yesterday evening from the hands of M. Carcavi, and that I admire it more than I can tell you. I do not have the leisure to write at length, but, in a word, you have found the two divisions of the points and of the dice with perfect justice. I am thoroughly satisfied as I can no longer doubt that I was wrong, seeing the admirable accord in which I find myself with you.

I admire your method for the problem of the points even more than that of the dice. I have seen solutions of the problem of the dice by several persons, as M. le chevalier de M´er´e, who proposed the question to me, and by M. Roberval also M. de M´er´e has never been able to find the just value of the problem of the points nor has he been able to find a method of deriving it, so that I found myself the only one who knew this proportion.

Your method is very sound and it is the first one that came to my mind in these researches, but because the trouble of these combinations was excessive, I found an abridgment and indeed another method that is much shorter and more neat, which I should like to tell you here in a few words; for I should like to open my heart to you henceforth if I may, so great is the pleasure I have had in our agreement. I plainly see that the truth is the same at Toulouse and at Paris.

This is the way I go about it to know the value of each of the shares when two gamblers play, for example, in three throws, and when each has put 32 pistoles at stake:

Let us suppose that the first of them has two (points) and the other one. They now play one throw of which the chances are such that if the first wins, he will win the entire wager that is at stake, that is to say 64 pistoles. If the other wins, they will be two to two and in consequence, if they wish to separate, it follows that each will take back his wager that is to say 32 pistoles.

Consider then, Monsieur, that if the first wins, 64 will belong to him. If he loses, 32 will belong to him. Then if they do not wish to play this point, and separate without doing it, the first should say "I am sure of 32 pistoles, for even a loss gives them to me. As for the 32 others, perhaps I will have them and perhaps you will have them, the risk is equal. Therefore let us divide the 32 pistoles in half, and give me the 32 of which I am certain besides." He will then have 48 pistoles and the other will have 16.

Now let us suppose that the first has two points and the other none, and that they are beginning to play for a point. The chances are such that if the first wins, he will win all of the wager, 64 pistoles. If the other wins, behold they have come back to the preceding case in which the first has two points and the other one.

But we have already shown that in this case 48 pistoles will belong to the one who has two points. Therefore if they do not wish to play this point, he should say, "If I win, I shall gain all, that is 64. If I lose, 48 will legitimately belong to me. Therefore give me the 48 that are certain to be mine, even if I lose, and let us divide the other 16 in half because there is as much chance that you will gain them as that I will." Thus he will have 48 and 8, which is 56 pistoles.

Let us now suppose that the first has but one point and the other none. You see, Monsieur, that if they begin a new throw, the chances are such that if the first wins, he will have two points to none, and dividing by the preceding case, 56 will belong to him. If he loses, they will he point for point, and 32 pistoles will belong to him. He should therefore say, "If you do not wish to play, give me the 32 pistoles of which I am certain, and let us divide the rest of the 56 in half. From 56 take 32, and 24 remains. Then divide 24 in half, you take 12 and I take 12 which with 32 will make 44.

By these means, you see, by simple subtractions that for the first throw, he will have 12 pistoles from the other; for the second, 12 more; and for the last 8.

But not to make this more mysterious, inasmuch as you wish to see everything in the open, and as I have no other object than to see whether I am wrong, the value (I mean the value of the stake of the other player only) of the last play of two is double that of the last play of three and four times that of the last play of four and eight times that of the last play of five, etc.

But the ratio of the first plays is not so simple to find. This therefore is the method, for I wish to disguise nothing, and here is the problem of which I have considered so many cases, as indeed I was pleased to do: Being given any number of throws that one wishes, to find the value of the first.

For example, let the given number of throws he 8. Take the first eight even numbers and the first eight uneven numbers as:

2, 4, 6, 8, 10, 12, 14, 16

and

1, 3, 5, 7, 9, 11, 13, 15.

Multiply the even numbers in this way. the first by the second, their product by the third, their product by the fourth, their product by the fifth, etc.; multiply the odd numbers in the same way: the first by the second, their product by the third, etc.

The last product of the even numbers is the denominator and the last product of the odd numbers is the numerator of the fraction that expresses the value of the first throw of eight. That is to say that if each one plays the number of pistoles expressed by the product of the even numbers, there will belong to him [who forfeits the throw] the amount of the other's wager expressed by the product of the odd numbers. This may he proved, but with much difficulty by combinations such as you have imagined, and I have not been able to prove it by this other method which I am about to tell you. but only by that of combinations. Here are the theorems which lead up to this which are properly arithmetic propositions regarding combinations, of which I have found so many beautiful properties:

If from any number of letters, as 8 for example,

A, B, C, D, E, F, G, H,

you take all the possible combinations of 4 letters and then all possible combinations of 5 letters, and then of 6, and then of 7, of 8, etc., and thus you would take all possible combinations, I say that if you add together half the combinations of 4 with each of the higher combinations, the sum will be the number equal to the number of the quaternary progression beginning with 2 which is half of the entire number.

For example, and I shall say it in Latin for the French is good for nothing.

If any number whatever of letters, for example 8,

A, B, C, D, E, F, G, H,

be summed in all possible combinations, by fours, fives, sixes, up to eights, I say, if you add half of the combinations by fours, that is 35 (half of 70) to all the combinations by fives, that is 56, and all the combinations by sixes, namely 28, and all the combinations by sevens, namely 8, and all the combinations by eights namely 1, the sum is the fourth number of the quaternary progression whose first term is 2. I say the fourth number for 4 is half of 8.

The numbers of the quaternary progressions whose first term is 2 are

2, 8, 32, 128, 512, etc.,

of which 2 is the first, 8 the second, 32 the third, and 128 the fourth. Of these, the 128 equals:

+ 35 half of the combinations of 4 letters
+ 56 the combinations of 5 letters
+ 28 the combinations of 6 letters
+ 8 the combinations of 7 letters
+ 1 the combinations of 8 letters.

That is the first theorem, which is purely arithmetic. The other concerns the theory of the points and is as follows:

It is necessary to say first: if one (player) has one point out of 5 for example, and if he thus lacks 4, the game will infallibly be decided in 8 throws, which is double 4.

The value of the first throw of 5 in the wager of the other is the fraction which has for its numerator the half of the combinations of 4 things out of 8 (I take 4 because it is equal to the number of points that he lacks, and 8 because it is double the 4) and for the denominator this same numerator plus all the higher combinations.

Thus if I have one point out of 5, 35/128 of the wager of my opponent belongs to me. That is to say, if he had wagered 128 pistoles, I would take 35 of them and leave him the rest, 93.

But this fraction 35/128 is the same as 105/384, which is made by the multiplication of the even numbers for the denominator and the multiplication of the odd numbers for the numerator.

You will see all of this without a doubt, if you will give yourself a little trouble, and for that reason I have found it unnecessary to discuss it further with you.

I shall send you, nevertheless, one of my old Tables; I have not the leisure to copy it, and I shall refer to it.

You will see here as always, that the value of the first throw is equal to that of the second, a thing which may easily be proved by combinations.

You will see likewise that the numbers of the first line are always increasing; those of the second do the same; those of the third the same.

But after that, those of the fourth line diminish; those of the fifth etc. This is odd.

i have no time to send you the proof of a difficult point which astonished M. (de M´er´e) so greatly, for he has ability but he is not a geometer (which is, as you know, a great defect) and he does not even comprehend that a mathematical line is infinitely divisible and he is firmly convinced that it is composed of a finite number of points. I have never been able to get him out of it. If you could do so, it would make him perfect.

He tells me then that he has found an error in the numbers for this reason . If one undertakes to throw a six with a die, the advantage of undertaking to do it in 4 is as 671 is to 625.

If one undertakes to throw double sixes with two dice the disadvantage of the undertaking is 24.

But nonetheless, 24 is to 36 (which is the number of faces of two dice)2 as 4 is to 6 (which is the number of faces of one die).

This is what was his great scandal which made him say haughtily that the theorems were not consistent and that arithmetic was demented. But you will easily see the reason by the principles which you have.

[Clearly, the number of possible ways in which two dice can fall.] 5

I shall put all that I have done with this in order when I shall have finished the treatise on geometry on which I have already been working for some time.

I have also done something with arithmetic on which subject, I beg you to give me your advice. I proposed the lemma which everyone accepts, that the sum of as many numbers as one wishes of the continuous progression from unity as 1, 2, 3, 4, being taken by twos is equal to the last term 4 multiplied into the next greater, 5.

That is to say that the sum of the integers in A being taken by twos is equal to the product A × (A + 1).

I now come to my theorem: If one he subtracted from the difference of the cubes of any two consecutive numbers, the result is six times all the numbers contained in the root of the lesser number.

Let the two roots R and S differ by unity. I say that R3 − S3 − 1 is equal to six times the sum of the numbers contained in S. Let S be called A, then R is A + 1. Therefore the cube of the root R or A + 1 is A3 + 3A2 + 3A + 13. The cube of S, or A, is A3, and the difference of these is R3 − S3; therefore, if unity he subtracted, 3A2+3A is equal to R3 −S3 −1. But by the lemma, double the sum of the numbers contained in A or S is equal to A×(A+1); that is, to A2+A. Therefore, six times the sum of the numbers in A is equal to 3A3 + 3A. But 3A3 + 3A is equal to R3 − S3 − 1. Therefore R3 − S3 − 1 is equal to six times the sum of the numbers contained in A or S. Quod erat demonstrandum.

No one has caused me any difficulty in regard to the above, but they have told me that they did not do so for the reason that everyone is accustomed to this method today. As for myself, I mean that without doing me a favor, people should admit this to be an excellent type of proof. I await your comment, however, with all deference. All that I have proved in arithmetic is of this nature.

Here are two further difficulties. i have proved a plane theorem making use of the cube of one line compared with the cube of another. I mean that this is purely geometric and in the greatest rigor. By these means I solved the problem: “Any four planes, any four points, and any four spheres being given, to find a sphere which, touching the given spheres, passes through the given points, and leaves on the planes segments in which given angles may be inscribed;” and this one: “Any three circles, any three points, and any three lines being given, to find a circle which touches the circles and the points and leaves on the lines and are in which a given angle may be inscribed.”

I solved these problems in a plane, using nothing in the construction but circles and straight lines, but in the proof I made use of solid loci,[6]—of parabolas, or hyperbolas. Nevertheless, inasmuch as the construction is in a plane, I maintain that my solution is plane, and that it should pass as such. This is a poor recognition of the honor which you have done me in putting up with my discourse which has been plaguing you so long. I never thought I should say two words to you and if I were to tell you what I have uppermost in my heart,—which is that the better I know you the more I honor and admire you,—and if you were to see to what degree that is, you would allot a place in your friendship for him who is,

Monsieur, your etc.

Fermat to Carcavi Sunday, August 9, 1654
Monsieur,

I was overjoyed to have had the same thoughts as those of M. Pascal, for I greatly admire his genius and I believe him to be capable of solving any problem he attempts. The friendship he offers is so dear to me and so precious that I shall not scruple to take advantage of it in publishing an edition of my Treatises.

If it does not shock you, you could both help in bringing out this edition, and I suggest that you should be the editors: you could clarify or augment what seems too brief and thus relieve me of a care which my work prevents me from taking.

I would like this volume to appear without my name even, leaving to you the choice of designation which would indicate the author, whom you could qualify simply as a friend.

Here is the course which I have thought out for the second Part which will contain my researches on numbers. It is a work which is still only an idea, and for which I may not have the leisure to put fully on paper ; but I will send a summary to M. Pascal of all my principles and first theorems, in which, I can promise you in advance, he will find everything not only novel and hitherto unknown but also astounding.

If you combine your work with his, everything will succeed and soon be completed, and we will thus be able to publish the first Part which you have in your care. If M. Pascal approves of my overtures which are based on my great esteem for his genius and his intellect, I will first begin to inform you of my numerical results.

Farewell. I am, Monsieur, your very humble and obedient servant. Fermat.

Pascal to Fermat Monday, August 24, 1654

Monsieur,

I was not able to tell you my entire thoughts regarding the problem of the points by the last post,7 and at the same time, I have a certain reluctance at doing it for fear lest this admirable harmony which obtains between us and which is so dear to me should begin to flag, for I am afraid that we may have different opinions on this subject.

I wish to lay my whole reasoning before you, and to have you do me the favor to set me straight if I am in error or to indorse me if I am correct. I ask you this in all faith and sincerity for I am not certain even that you will be on my side. When there are but two players, your theory which proceeds by combinations is very just. But when there are three, I believe I have a proof that it is unjust that you should proceed in any other manner than the one I have.

But the method which I have disclosed to you and which I have used universally is common to all imaginable conditions of all distributions of points, in the place of that of combinations (which I do not use except in particular cases when it is shorter than the general method), a method Which is good only in isolated cases and not good for others.

I am sure that I can make it understood, but it requires a few words from me and a little patience from you. 2. This is the method of procedure when there are two players. If two players, playing in several throws, find themselves in such a state that the first lacks two points and the second three of gaining the stake, you say it is necessary to see in how many points the game will be absolutely decided.

It is convenient to suppose that this will be in four points, from which you conclude that it is necessary to see how many ways the four points may be distributed between the two players and to see how many combinations there are to make the first win and how many to make the second win, and to divide the stake according to that proportion. I could scarcely understand this reasoning if I had not known it myself before; but you also have written it in your discussion.

Then to see how many ways four points may be distributed between two players, it is necessary to imagine that they play with dice with two faces (since there are but two players), as heads and tails, and that they throw four of these dice (because they play in four throws).

Now it is necessary to see how many ways these dice may fall. That is easy to calculate. There can be sixteen, which is the second power of four; that is to say, the square. Now imagine that one of the faces is marked a, favorable to the first player. And suppose the other is marked b, favorable to the second. Then these four dice can fall according to one of these sixteen arrangements.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | a | a | a | a | a | a | a | b | b | b | b | b | b | b | b |
| a | a | a | a | b | b | b | b | a | a | a | a | b | b | b | b |
| a | a | b | b | a | a | b | b | a | a | b | b | a | a | b | b |
| a | b | a | b | a | b | a | b | a | b | a | b | a | b | a | b |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 |

and, because the first player lacks two points, all the arrangements that have two a's make him win. There are therefore 11 of these for him. And because the second lacks three points, all the arrangements that have three b's make him win. There are 5 of these. Therefore it is necessary that they divide the wager as 11 is to 5.

There is your method, when there are two players, whereupon you say that if there are more players. it will not be difficult to make the division by this method. 3. On this point, Monsieur, I tell you that this division for the two players founded on combinations is very equitable and good, but that if there are more than two players, 8 it is not always just and I shall tell you the reason for this difference. I communicated your method to [some of] our gentlemen, on which M. de Roberval made me this objection: That it is wrong to base the method of division on the supposition that they are playing in four throws seeing that when one lacks two points and the other three, there is no necessity that they play four throws since it may happen that they play but two or three, or in truth perhaps four.

Since he does not see why one should pretend to make a just division on the assumed condition that one plays four throws, in view of the fact that the natural terms of the game are that they do not throw the dice after one of the players has won; and that at least if this is not false, it should he proved.

Consequently he suspects that we have committed a paralogism. I replied to him that I did not found my reasoning so much on this method of combinations, which in truth is not in place on this occasion, as on my universal method from which nothing escapes and which carries its proof with itself. This finds precisely the same division as does the method of combinations.

Furthermore, I showed him the truth of the divisions between two players by combinations in this way. Is it not true that if two gamblers finding according to the conditions of the hypothesis that one lacks two points and the other three, mutually agree that they shall play four complete plays, that is to say, that they shall throw four two-faced dice all at once,—is it not true, I say, that if they are prevented from playing the four throws, the division should be as we have said according to the combinations favorable to each?

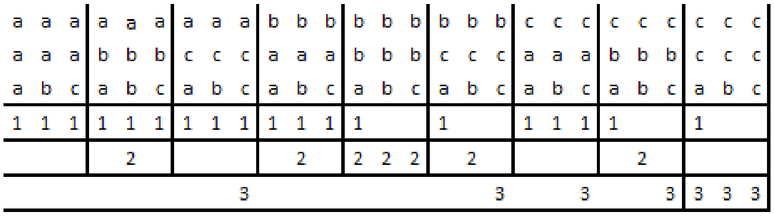
He agreed with this and this is indeed proved. But he denied that the same thing follows when they are not obliged to play the four throws. I therefore replied as follows: It is not clear that the same gamblers, not being constrained to play the four throws, but wishing to quit the game before one of them has attained his score, can without loss or gain be obliged to play the whole four plays, and that this agreement in no way changes their condition? For if the first gains the two first points of four. will he who has won refuse to play two throws more, seeing that if he wins he will not win more and if he loses he will not win less?

For the two points which the other wins are not sufficient for him since he lacks three, and there are not enough [points] in four throws for each to make the number which he lacks. It certainly is convenient to consider that it is absolutely equal and indifferent to each whether they play in the natural way of the game, which is to finish as soon as one has his score, or whether they play the entire four throws.

Therefore, since these two conditions are equal and indifferent, the division should he alike for each. But since it is just when they are obliged to play the four throws as I have shown, it is therefore just also in the other case. That is the way I prove it, and, as you recollect, this proof is based on the equality of the two conditions true and assumed in regard to the two gamblers, the division is the same in each of the methods, and if one gains or loses by one method, he will gain or lose by the other, and the two will always have the same accounting.

Let us follow the same argument for three players and let us assume that the first lacks one point, the second two, and the third two. To make the division, following the same method of combinations, it is necessary to first discover in how many points the game may he decided as we did when there woe two players. This will be in three 9 points for they cannot play three throws without necessarily arriving at a decision. It is now necessary to see how many ways three throws may he combined among three players and how many are favorable to the first, how many to the second, and how many to the third, and to follow this proportion in distributing the wager as we did in the hypothesis of the two gamblers.

It is easy to see how many combinations there are in all. This is the third power of 3; that is to say, its cube, or 27. For if one throws three dice at a time (for it is necessary to throw three times), these dice having three faces each (since there are three players), one marked a favorable to the first, one marked b favorable to the second, and one marked c favorable to the third,—it is evident that these three dice thrown together can fall in 27 different ways as:



| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | a | a | a | a | a | a | a | a | b | b | b | b | b | b | b | b | b | c | c | c | c | c | c | c | c | c |
| a | a | a | b | b | b | c | c | c | a | a | a | b | b | b | c | c | c | a | a | a | b | b | b | c | c | c |
| a | b | c | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c | a | b | c |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | | 1 | | | 1 | 1 | 1 | 1 | | | 1 | | |
| | | | | 2 | | | | | | 2 | | 2 | 2 | 2 | | 2 | | | | | | 2 | | | | |
| | | | | | | | | 3 | | | | | | | | | 3 | | | 3 | | | 3 | 3 | 3 | 3 |

Since the first lacks but one point, then all the ways in which there is one a are favorable to him. There are 19 of these. The second lacks two points. Thus all the arrangements in which there are two b's are in his favor. There are 7 of them. The third lacks two points. Thus all the arrangements in which there are two c's are favorable to him. There are 7 of these. If we conclude from this that it is necessary to give each according to the proportion 19, 7, 7, we are making a serious mistake and I would hesitate to believe that you would do this.

There are several cases favorable to both the first and the second, as aab has the a which the first needs, and the two b's which the second needs. So too, the ace is favorable to the first and third. It therefore is not desirable to count the arrangements which are common to the two as being worth the whole wager to each, but only as being half a point. For if the arrangement occurs, the first and third will have the same right to the wager, each making their score.

They should therefore divide the wager in half. If the arrangement aab occurs, the first alone wins. It is necessary to make this assumption. There are 13 arrangements which give the entire wager to the first, and 6 which give him half and 8 which are worth nothing to him. Therefore if the entire sum is one pistole, there are 13 arrangements which are each worth one pistole to him, there are 6 that are each worth Y2 a pistole, and 8 that are worth nothing.

Then in this case of division, it is necessary to multiply 13 by one pistole which makes 13 6 by one half which makes 3 8 by zero which makes 0 Total 16 and to divide the sum of the values 16 by the sum of the arrangements 27, which makes the fraction 16/27 and it is this amount which belongs to the first gambler in the event of a division; that is to say, 16 pistoles out of 27. 10 The shares of the second and the third gamblers will be the same: There are 4 arrangements which are worth 1 pistole; multiplying, 4 There are 3 arrangements which are worth 3/2 pistole; multiplying, 11 2 And 20 arrangements which are worth nothing 0 Total 27 Total 51 2

Therefore 5 1 2 pistoles belong to the second player out of 27, and the same to the third. The sum of the 5 1 2 , 5 1 2 , and 16 makes 27. 5.

It seems to me that this is the way in which it is necessary to make the division by combinations according to your method, unless you have something else on the subject which I do not know.

But if I am not mistaken, this division is unjust. The reason is that we are making a false supposition,—that is, that they are playing three throws without exception, instead of the natural condition of this game which is that they shall not play except up to the time when one of the players has attained the number of points which he lacks, in which case the game ceases.

It is not that it may not happen that they will play three times, but it may happen that they will play once or twice and not need to play again. But, you will say, why is it possible to make the same assumption in this case as was made in the case of the two players?

Here is the reason: In the true condition [of the game] between three players, only one can win, for by the terms of the game it will terminate when one [of the players] has won. But under the assumed conditions, two may attain the number of their points, since the first may gain the one point he lacks and one of the others may gain the two points which he lacks, since they will have played only three throws.

When there are only two players, the assumed conditions and the true conditions concur to the advantage of both. It is this that makes the greatest difference between the assumed conditions and the true ones.

If the players, finding themselves in the state given in the hypothesis,—that is to say, if the first lacks one point, the second two, and the third two; and if they now mutually agree and concur in the stipulation that they will play three complete throws; and if he who makes the points which he lacks will take the entire sum if he is the only one who attains the points; or if two should attain them that they shall share equally,- in this case, the division should be made as I give it here. the first shall have 16, the second 5 1 2 , and the third 5 1 2 out of 27 pistoles, and this carries with it its own proof on the assumption of the above condition.

But if they play simply on the condition that they will not necessarily play three throws, but that they will only play until one of them shall have attained his points, and that then the play shall cease without giving another the opportunity of reaching his score, then 17 pistoles should belong to the first, 5 to the second, and 5 to the third, out of 27. And this is found by my general method which also determines that, under the proceeding condition, the first should have 16, the second 5 1 2 , and the third without making use of combinations,—for this works in all cases and without any obstacle.

These, Monsieur, are my reflections on this topic on which I have no advantage over you except that of having meditated on it longer, but this is of little [advantage to me] from your point of view since your first glance is more penetrating than are my prolonged endeavors.

I shall not allow myself to disclose to you my reasons for looking forward to your opinions. I believe you have recognized from this that the theory of combinations is good for the case of two players by accident, as it is also sometimes good in the case of three gamblers, as when one lacks one point, another one, and the other two,[8] because, in this case, the number of points in which the game is finished is not enough to allow two to win, but it is not a general method and it is good only in the case where it is necessary to play exactly a certain number of times.

Consequently, as you did not have my method when you sent me the division among several gamblers, but [since you had] only that of combinations, I fear that we hold different views on the subject. I beg you to inform me how you would proceed in your research on this problem. I shall receive your reply with respect and joy, even if your opinions should be contrary to mine.

I am etc.

Fermat to Pascal Saturday, August 29, 1654

Monsieur,

Our interchange of blows still continues, and I am well pleased that our thoughts are in such complete adjustment as it seems since they have taken the same direction and followed the same road.

Your recent Trait´e du triangle aritbm´etique and its applications are an authentic proof and if my computations do me no wrong, your eleventh consequence went by post from Paris to Toulouse while my theorem, on figurate numbers, which is virtually the same, was going from Toulouse to Paris. I have not been on watch for failure while I have been at work on the problem and I am persuaded that the true way to escape failure is by concurring with you.

But if I should say more, it would he of the nature of a Compliment and we have banished that enemy of sweet and easy conversation. It is now my turn to give you some of my numerical discoveries, but the end of the parliament augments my duties and I hope that out of your goodness you will allow me due and almost necessary respite.

I will reply however to your question of the three players who play in two throws. When the first has one [point] and the others none, your first solution is the true one and the division of the wager should he 17, 5, and 5. The reason for this is self-evident and it always takes the same principle, the combinations making it clear that the first has 17 changes while each of the others has but five.

For the rest, there is nothing that I will not write you in the future with all frankness. Meditate however, if you find it convenient, on this theorem: The squared powers of 2 augmented by unity11 are always prime numbers. [That is,] The square of 2 augmented by unity makes 5 which is a prime number;

The square of the square makes 16 which, when unity is added makes 17, a prime number; The square of 16 makes 256 which, when unity is added, makes 257, a prime number; The square of 256 makes 65536 which, when unity is added, makes 65537, a prime number; and so to infinity.

This is a property whose truth I will answer to you. The proof of it is very difficult and I assure you that I have not yet been able to find it fully. I shall not set it for you to find unless I come to the end of it. This theorem serves in the discovery of numbers which are in a given ratio to their aliquot parts, concerning which I have made many discoveries.

We will talk of that another time. I am Monsieur, yours etc.

Fermat.

At Toulouse, the twenty ninth of August, 1654.

Fermat to Pascal Friday, September 25, 1654

Monsieur,

Do not be apprehensive that our argument is coming to an end. You have strengthened it yourself in thinking to destroy it and it seems to me that in replying to M. de Roberval for yourself you have also replied for me.

In taking the example of the three gamblers of whom the first lacks one point, and each of the others lack two, which is the case in which you oppose, I find here only 17 combinations for the first and 5 for each of the others; for when you say that the combination ace is good for the first, recollect that everything that is done after one of the players has won is worth nothing.

But this combination having Made the first win on the first die, what does it matter to the third gains two afterwards, since even when he gains thirty all this is superfluous? The consequence, as you have well called it “this fiction,” of extending the game to a certain number of plays serves only to make the rule easy and (according to my opinion) to make all the chances equal; or better, more intelligibly to reduce all the fractions to the same denomination.

So that you may have no doubt, if instead of three parties you extend the assumption to four, there will not be 27 combinations only, but 81; and it will be necessary to see how many combinations make the first gain his point later than each of the others gains two, and how many combinations make each of the others win two later than the first wins one.

You will find that the combinations that make the first win are 51 and those for each of the other two are 15, which reduces to the same proportion. So that if you take five throws or any other number you please, you will always find three numbers in the proportion of 17, 5, 5.

And accordingly I am right in saying that the combination ace is [favorable] for the first only and not for the third, and that cca is only for the 13 third and not for the first, and consequently my law of combinations is the same for three players as for two, and in general for all numbers. 2.

You have already seen from my previous letter that I did not demur at the true solution of the question of the three gamblers for which I sent you the three definite numbers, 17, 5, 5. But because M. de Roberval will perhaps be better satisfied to see a solution without any dissimulation and because it may perhaps yield to abbreviations in many cases, here is an example:

The first may win in a single play, or in two or in three. If he wins in a single throw, it is necessary that he makes the favorable throw with a three-faced die at the first trial. A single die will yield three chances. The gambler then has 1/3 of the wager because he plays only one third. If he plays twice, he can gain in two ways,-either when the second gambler wins the first and he the second, or when the third wins the throw and when he wins the second. But two dice produce 9 chances. The player than has 2/9 of the wager when they play twice. But if he plays three times, he can win only in two ways, either the second wins on the first throw and the third wins the second, and he the third; or when the third wins the first throw, the second the second, and he the third; for if the second or the third player wins the two first, he will win the wager and the first player will not. But three dice give 27 chances of which the first player has 2/27 of the chances when they play three rounds. The sum of the chances which makes the first gambler win is consequently 1/3, 2/9, and 2/27, which makes 17/27.

This rule is good and general in all cases of the type where, without recurring to assumed conditions, the true combinations of each number of throws give the solution and make plain what I said at the outset that the extension to a certain number of points is nothing else than the reduction of divers fractions to the same denomination.

Here in a few words is the whole of the mystery, which reconciles us without doubt although each of us sought only reason and truth. 3. I hope to send you at Martinmas an abridgment of all that I have discovered of note regarding numbers. You allow me to be concise [since this suffices] to make myself understood to a man [like yourself who comprehends the whole from half a word. What you will find most important is in regard to the theorem that every number is composed of one, two, or three triangles;12 of one, two, three, or four squares; of one, two, three, four, or five pentagons; of one, two, three, four, five, or six hexagons, and thus to infinity.

To derive this, it is necessary to show that every prime number which is greater by unity than a multiple of 4 is composed of two squares, as 5, 13, 17, 29, 37, etc. Having given a prime number of this type, as 53, to find by a general rule the two squares which compose it. Every prime number which is greater by unity than a multiple of 3, is composed of a square and of the triple of another square, as 7, 13, 19. 31, 37, etc. Every prime number which is greater by 1 or by

3 than a multiple of 8, is composed of a square and of the double of another square, as 11, 17, 19, 41, 43, etc. 12[I.e., triangular numbers.]

There is no triangle of numbers whose area is equal to a square number. This follows from the invention of many theorems of which Bachet vows himself ignorant and which are lacking in Diophantus. I am persuaded that as soon as you will have known my way of proof in this type of theorem, it will seem good to you and that it will give you the opportunity for a multitude of new discoveries, for it follows as you know that multi pertranseant ut augeatur scientia. When I have time, we will talk further of magic numbers and I will summarize my former work on this subject.

I am, Monsieur, most heartily your etc.

Fermat.

The twenty-fifth of September.

I am writing this from the country, and this may perhaps delay my replies during the holidays.

Pascal to Fermat Tuesday, October 27, 1654

Monsieur,

Your last letter satisfied me perfectly.

I admire your method for the problem of the points, all the more because I understand it well. It is entirely yours, it has nothing in common with mine, and it reaches the same end easily. Now our harmony has begun again. But, Monsieur, I agree with you in this, find someone elsewhere to follow you in your discoveries concerning numbers, the statements Of which you were so good as to send me. For my own part, I confess that this passes me at a great distance; I am competent only to admire it and I beg you most humbly to use your earliest leisure to bring it to a conclusion.

All of our gentlemen saw it on Saturday last and appreciate it most heartily. One cannot often hope for things that are so fine and so desirable. Think about it if you will, and rest assured that I am etc.

Pascal.

Paris, October 27, 1654.

Fermat to Pascal Sunday, July 25, 1660

Monsieur,

As soon as I discovered that we were nearer to one another than we had ever been before, I could not resist making plans for renewing our friendship and I asked M. de Carcavi to be mediator: in a word I would like to embrace you and to talk to you for a few days ; but as my health is not any better than yours, I very much hope that you will do me the favour of coming half way to meet me and that you will oblige me by suggesting a place between Clermont and Toulouse, where I would go without fail towards the end of September or the beginning of October.

If you do not agree to this arrangement, you will run the risk of seeing me at your house and of thus having two ill people there at once. I await your news with impatience and am, with all my heart,

Yours ever,

Fermat.

Pascal to Fermat
Tuesday, August 10, 1660
Monsieur,

You are the most gallant man in the world and assuredly I am the one who can best recognize your qualities and very much admire them, especially when they are combined with your own singular abilities. Because of this I feel I must show my appreciation of the offer you have made me, whatever difficulty I still have in reading and writing, but the honour you do me is so dear to me that I cannot hasten too much in answering your letter. 1 will tell you then, Monsieur, that if I were in good health, I would have flown to Toulouse and I would not allow a man such as you to take one step for a man such as myself.

I will tell you also that, even if you were the best Geometrician in the whole of Europe, it would not be that quality which would attract me to you, but it is your great liveliness and integrity in conversation that would bring me to see you. For, to talk frankly with you about Geometry, is to me the very best intellectual exercise: but at the same time I recognize it to be so useless that I can find little difference between a man who is nothing else but a geometrician and a clever craftsman. Although I call it the best craft in the world, it is after all only a craft, and I have often said it is fine to try one's hand at it but not to devote all one's powers to it.

In other words, I would not take two steps for Geometry and I feel certain you are very much of the same mind. But as well as all this, my studies have taken me so far from this way of thinking, that I can scarcely remember that there is such a thing as geometry. I began it, a year or two ago, for a particular reason; having satisfied this, it is quite possible that I shall never think about it again. Besides, my health is not yet very good, for I am so weak that I cannot walk without a stick nor ride a horse, I can only manage three or four leagues in a carriage.

It was in this way that I took twenty-two days in coming here from Paris. The doctors recommended me to take the waters at Bourbon during the month of September, and two months ago I promised, if I can manage it, to go from there through Poitou by river to Saumur to stay until Christmas with M. le duc de Roannes, governor of Poitou, who has feelings for me that I do not deserve. But, since I go through Orleans on my way to Saumur by river and if my health prevents me from going further. I shall go from there to Paris.

There, Monsieur, is the present state of my life, which I felt obliged to describe to you so as to convince you of the impossibility of my being able to receive the honour you have so kindly offered me. I hope, with all my heart, that one day I shall be able to acknowledge it to you or to your children, to whom I am always devoted, having a special regard for those who bear the name of the foremost man in the world. I am, etc.

Pascal.
De Bienassis,
10th August, 1660

From Whence it Came – An Early History of Probability
Probability and the Renaissance
Blaise Pascal
Pierre Fermat
Abraham de Moivre
Simon Laplace
The Notion of Random Events

# Simon Laplace

According to some, the 1760s the teenaged Simon Laplace had a problem. Born into a family that was not poor, but committed to agriculture, his parents consigned him to a safe, secure, but inconspicuous existence.

However, incessantly pulled by the restless drive to discover, he determined that he would go another way, even if that new way required money. With no money of his own, and his family unable to contribute, he turned to an unusual and untapped source ─ his neighbors!

Through his unique combination of force of personality, unshakeable will, and an uncanny knack to discern what people needed to hear, young Simon went from house to house, alternately asking, cajoling, and demanding that the different households, themselves poor, provide money for his training. The astonished neighbors acceded to his request, committing their hard earned money to his education.

There has been no community that has earned a greater return on investment, then Beaumont-en-Auge, Laplace's home town.

Whether it is true or not, it speaks to the remarkable drive of the young man who earned the sobriquet, "The Newton of France" [1].

Yet, when he arrived at the university, according to some, young Laplace was treated badly [2] by D'Alembert, a mathematician, philosopher, and musical theorist was particularly unimpressed with the enthusiastic Simon Laplace, d'Alembert dispatched the young man to a collection of thick mathematics tomes, telling Laplace to return when he had mastered them.

Several days later, Laplace returned.

D'Alembert upon questioning Laplace, realized that he had in fact mastered the material. The teacher dropped all reservations, accepted him as a mathematical prodigy.

After completing some university training, Laplace presented a well-received paper before the French Academy of Science in 1773 in which he demonstrated the stability of planetary motion. He was twenty-four years old, yet he followed with the first of four works that rocketed the field of probability forward into a perilous but promising future.

His manuscript, entitled *Mémoire sur la Probabilité des Causes par les Évènements* provided further justification for the use of what would become known as the prior distribution, establishing the legitimacy of the inversion approach in probability and setting the stage for the first Bayesian computation.

In doing so, Laplace opened a door into either a lush valley of intellectual and philosophical fruit, or a sharp precipice that would tear the young field of probability apart The stark differences between his derivations and those of Bayes make it unlikely that he relied on Bayes initial work. However, Laplace's legendary, regrettable tendency to pirate and then disparage the work of others forces us to keep this an open question.

Laplace moved forward with the development of the probability generating function [3] in addition to the Laplace transforms.

Even his warmest admirers acknowledged Laplace's vanity and selfishness. He was contemptuous of the benefactors of his youth, ungrateful to his policial friends, and his appropriation without acknowledgment of the work of others was reprehensible. Yet, he was also very protective and generous toward his students. In one case, he suppressed a paper of his own in order that a pupil might have the sole credit of the investigation [4].

He worked for a short time as the Minister of the Interior, but was quickly discharged because of his inability to keep small administrative problems simple. In his later years, he debated the true meaning of God, and died in Paris on March 5, 1827, at 78 years of age.

Why Probability
From Whence it Came – An Early History of Probability
Probability and the Renaissance
Blaise Pascal
Pierre Fermat
Abraham de Moivre
Famous Correspondence between Pascal and Fermat
Simon Laplace
Bayes and Price
The Inversion Problem

References

1 . https://en.wikipedia.org/wiki/Pierre-Simon_Laplace last accessed March 19, 2020.
2 Laplace, being Extracts from Lectures delivered by Karl Pearson", Biometrika, vol. 21, December 1929, pp. 202–216
3 Ball, RB. 1908. A Short Account of the History of Mathematics. Front Cover. Walter. Macmillan and Company, limited, 1908 - Mathematics
4. Ball W.W. Rouse (1960). *A Short Account of the History of Mathematics (4th Edition*. New York: Dover Publications.

# Thomas Bayes and Richard Price

In 1763, a paper entitled, "An Essay Towards Solving a Problem in the Doctrine of Chances" appeared in the *Philosophical Transaction of the Royal Society of London* [1]. Few paid attention to the contents of this confusing paper. However, if a dogged reader persisted, they would have learned two remarkable things. First, its author was a reverend. Second, he was dead [2]! The Reverend Thomas Bayes, having published only three manuscripts during his entire life, died two years before the appearance of his final manuscript, and in fact, he never submitted this final manuscript for publication, it being identified and posthumously offered for publication by his friend, Reverend Richard Price.

Similarities and differences mark the relationship of these two men. Both were ministers. Born in 1702, Thomas was the son of the nonconformist minister Joshua Bayes. Nonrebellious, he followed his father into the ministry, becoming a Presbyterian minister in the village of Tumbridge Wells (about twenty-five miles south of London) at the age of twenty-nine. His first paper was on religious matters.[*]

His second written in 1736, entitled, "An Introduction to the Doctrine of Fluxions, and a Defense of the Mathematicians Against the Objections of the Author of the Analyst", was a gentle defense of Newton's calculus, a publication many believed earned him entrance into the prestigious Royal Academy of Science. However, once admitted, he remained quiet and unremarkable, never publishing again for the remaining twenty-five years of his life.

The rebellious, articulate Richard Price, however, was quite another matter. Also born into a family of ministers, Price quickly rejected the strict religious teachings of his family. Openly rejecting the prevalent Christian doctrines of original sin and eternal damnation, he earned the ire of many traditionalists.

Undaunted, he entered the ministry himself, and upon reaching the pulpit, shared his iconoclastic convictions with his parishioners. Contending that individuals had the obligation to use not just the Bible, but their own conscience and the best of their reasoning skills to resolve a moral dilemma made him the focal point of local criticism. Undaunted, he went on to argue against the deity of His Majesty the King.

This outspoken iconoclast and the insular, enigmatic Bayes had two things in common. One was the ministry, although clearly their theological views differed. The second was an affection for mathematics. While very little is left of Bayes work, Price's intense interest in probability is well established. It was through mathematics that the activist Price and the reserved Bayes became friends.

Thus, it was no surprise that, upon Reverend Bayes' death in 1761, his bereaved family would call upon Reverend Price to examine and organize Bayes' papers and notebooks, leading Price to discover the scattered writing of his dead friend.

The rest is speculation. Price recognized the topic of the paper was probability. Maybe the two of them had some earlier conversations about the interesting notion of using the past to

[*] In 1731, Bayes published "Divine Benevolence, or an Attempt to Prove that the Principal End of the Divine Providence and Government Is the Happiness of His Creatures."

predict the future. Nevertheless, Price went about the task of revising Bayes' manuscript draft. Bayes had written an introduction ─ Price replaced it. In addition, Price added an appendix [3]. Others suggest that Price made other substantial changes, amplifying the use of probability in a revolutionary way. Finally, Price read the revised manuscript before the Royal Society on September 23, 1763, two years after Bayes' death, giving full credit to his friend. In 1765, two years after the manuscript was published, Price was himself admitted to the Royal Society for his own work on probability.[*]

The precise contributions of Bayes and Price in the 1763 manuscript remains a mystery. However, we can say that, while it was ultimately Bayes' bombshell, the hand that repacked the explosives and lit the fuse belonged to Reverend Price. The paper itself, like a slowly burning fuse, languish unnoticed for a decade,[†] until its ideas were modified and amplified by Simon Laplace.

The Inversion Problem
Incidence and Prevalence
Physicians and Conditional Probability
Sensitivity and Specificity

1. Bayes T. (1763). "An Essay Towards Solving a Problem in the Doctrine of Chances." *Philosophical Transactions of the Royal Society of London,* **53**:370–418.
2. Fienberg S.E. (2006). When did bayesian inference become "bayesian? *Bayesian Analysis* **1**:1–40.
3. Stigler S.M. (1986). The History of Statistics: The Measurement of Uncertainty before 1990. Cambridge, London. The Belknap Press of Harvard University Press. p. 98.

[*] Price went on to publish several books including *Observations on the Nature of Civil Liberty, the Principles of Government, and the Justice and Policy of War with America in 1776.* His chapel was visited by such notables as John Howard, leader of an influential group of prison reformers, and John Quincy Adams, future President of the United States.

[†] Steven Stigler, a leading historian of statistics, has even suggested that Bayes Theorem was really discovered by Nicolas Saunderson, a blind mathematician who was the fourth Lucasian Professor of Mathematics at Cambridge University. Stephen Hawking is the current holder of this chair.

# Andrey Kolmogorov

The modern, axiomatic treatment of probability can be attributed to Andrey Kolmogorov, one of the greatest probabilitsts of the 20th century. And, with this as a foundation, he went on to make substantial contributions to the fields of stochastic processes, information theory, fluid mechanics, and epidemiologic modeling.

Andrey Kolmogorov was born in 1903 in Tambov, Russia approximately 300 miles south of Moscow. His mother, Kolmogorova, died in childbirth. Little is known about his father; some believe he was deported from St. Petersburg for taking part in protests against the czars and later killed in the Russian Civil War.[1]

In the absence of both parents, Andrey was raised by two aunts at his grandfather's estate. He attended the village school, there demonstrating genuine curiosity about mathematics; the school newspaper printed several of his mathematical and literary works.

In 1910, his aunt adopted him, moving him to Moscow where he attended high school, graduating in 1920. During this time, Kolmogorov became interested and developed a perpetual motion machine. Even his teachers were unable to find its flaws.

Kolmogorov could be quite intense in his work and at the same time, enjoy an easygoing existence. For example, after high school, he wrote a treatise on Newton's law of mechanics, while working as a railway conductor.[2]

He then decided to continue his education, entering Moscow State University. He spent the rest of his career at this university, becoming a faculty member, and then department chair.

However, at first, he was uncommitted to mathematics, devoting energy to metallurgy and Russian history, about which he was passionate.

He first attracted notice as a mathematical intellectual with a paper on set operations, published in 1922. This was an advance on the wok of Suslin, who was advancing the field of Borel sets.[*] Kolmogorov went on to publish eight papers while an undergraduate student.[3], including one in June 1922 in which he constructed a summable function that diverged almost everywhere.

This stunning and unexpected finding in the world of mathematics, boosted him to international acclaim before graduating from Moscow State University in 1925.

He immediately began work under Luzin's supervision, producing in that year his first paper on probability. This was published jointly with Khinchin and contains the "three series theorem" as well as results on inequalities of partial sums of random variables, which would become the basis for martingale inequalities and stochastic systems. By this time, he had eighteen publications including papers on the strong law of large numbers and the law of the iterated logarithm.

In 1929, Kolmogorov earned his doctor of philosophy degree from Moscow State University. In 1931 he became a professor there, devoting himself to a rigorous examination of the underlying tenets of probability. He reformulated probability in a 1933 paper in which he assembled its development from a fundamental collection of axioms, much like Euclid-developed geometry.

That same year, he published his classic book, *Foundations of the Theory of Probability*, laying the modern axiomatic foundations of probability theory and establishing his reputation as the world's leading expert in this field.[4] It was in this work that he developed the concept of

[*] Borel sets is the σ-algebra of all open sets of real numbers.

probability not as a stand-alone field typified by unique relationships but wholly encompassed in the larger field of measure theory (i.e., probability is just one of many types of measures).

He demonstrated intense interest in problems of differentiation and integration and measures of sets. In every one of his papers dealing with such a variety of topics, he introduced an element of originality, a breadth of approach, and depth of thought.

In 1935, Kolmogorov became the first chairman of the Department of Probability Theory at the Moscow State University.

In a 1938 paper, Kolmogorov "established the basic theorems for smoothing and predicting stationary stochastic processes"—a paper that would have major military applications during the Cold War. In 1939, he was elected a full member (academician) of the USSR Academy of Sciences.

During this time, Kolmogorov contributed to the field of ecology. In fact, his study of stochastic processes (random processes), especially Markov processes, led him and the British mathematician Sydney Chapman to independently develop the pivotal set of equations in the field, which have been give the name of the Chapman-Kolmogorov equations. These equations have been instrumental in the mathematical development of the spread of disease.

Later on, Kolmogorov changed his research interests to the area of turbulence, where his publications beginning in 1941 had a significant influence on the field. In classical mechanics, he is best known for the Kolmogorov-Arnold-Moser theorem (first presented in 1954 at the International Congress of Mathematicians). He was a founder of algorithmic complexity theory, often referred to as Kolmogorov complexity theory, which he began to develop around this time.

Kolmogorov married in 1942. Active not only in mathematics, he devoted time to working with gifted children. In addition, he pursued interests in literature and in music.

Kolmogorov served his alma mater, Moscow State University, in different faculty positions and department chairs. However, he retained an abiding interest in his students. He commonly invited students to take long outdoor walks with him, discussing concepts in mathematics.

Kolmogorov died in Moscow in 1987. His remains can be found in the Novodevichy cemetery.

Bernhard Riemann
Henri Lebesgue

Introduction to Measure Theory (nonmathematical)
Measurable Functions
Measure and Integration

[1] https://en.wikipedia.org/wiki/Andrey_Kolmogorov, accessed 1-14-2020
[2] http://mathshistory.st-andrews.ac.uk/ Biographies/Kolmogorov.html, last accessed 1-14-2020
[3] https://en.wikipedia.org/wiki/Andrey_Kolmogorov
[4] http://arbor.revistas.csic.es/index.php/arbor/article/viewFile /551/552, last accessed 1-14-2020

# Bernhard Riemann

Monastyrsky wrote
*It is difficult to recall another example in the history of nineteenth-century mathematics when a struggle for a rigorous proof led to such productive results.*

Georg Friedrich Bernhard Riemann is the father of integral calculus. He was also an influential German mathematician who made lasting contributions to analysis, number theory, and differential geometry, some of them enabling the later development of general relativity.[1]

Riemann was born in 1826 in the kingdom of Hannover, which would become part of Germany, and showed an early interest in mathematics and history. Encouraged by his family, he entered preparatory school in Hannover, later moving to Lüneburg [2]

Riemann was born in 1826 in the kingdom of Hannover, later to become part of Germany. Bernhard was the second of their six children. His father, a Lutheran minister acted as teacher to his children and he taught Bernhard until he was ten years old.

Encouraged by his family and a teacher named Schulz, who assisted in his education, Bernhard entered preparatory school in Hannover, later moving to Lüneburg [2]

There, Bernhard seems to have been a good, but not outstanding, pupil who worked hard at the classical subjects such as Hebrew and theology.

However, his remarkable mathematical talent was noticed by Schmalfuss, the director of the gymnasium. Six days after giving Riemann a textbook on number theory by Legendre, Riemann returned the 859-page book, saying, "That was a wonderful book! I have mastered it."

In 1846, Riemann matriculated at Göttingen University. In accordance with his father's wishes, he began in the faculty of theology, but he soon transferred to the faculty of philosophy to pursue science and mathematics.[3]

With this experience Bernhard, always close to his family, asked his father if he could transfer to the faculty of philosophy so that he could study mathematics.[4] Being very close to his family, he would not change courses without his father's permission.

Receiving his father's blessing, Bernhard then took courses in mathematics from Moritz Stern and the mathematical giant Carl Frederick Gauss. However, there is no evidence that at this time, Gauss, quite unsociable, ever had any personal contact with Riemann.[5]

Riemann studied the work of Cauchy, who had created the $\varepsilon-\delta$ method of calculus, and his work on integration through the development of the Riemann integral is still taught today.

His ability did draw the attention of another Göttingen mathematician, Moritz Stern. After a year Riemann moved to the University of Berlin, where he could benefit from the teaching of Jacobi, Steiner, Dirichlet, and Eisenstein. It was Dirichlet who influenced Riemann the most, and was to become his collaborator. In 1850 Riemann returned to Göttingen, where he would spend the rest of his career.

Until Riemann's work, the mathematical process of integration was not an accepted field of study. The process of integration was seen as simply the reverse of finding the derivative of a function, so essential in differential calculus (co-discovered by Isaac Newton and Gauss). Riemann developed the powerful tool of studying limits using the $\varepsilon-\delta$ method of examining a function's behavior across very small regions.

He then developed the theory of the integral on its own (separate and apart from derivatives) through a limiting process of what has come to be known as Riemann sums.

This work established Riemann as an important mathematician. In addition, he developed a very powerful geometric theory that resolved a number of outstanding problems. He is associated with among the most important but unproved statements in number theory, the Riemann hypothesis.*

In 1859 Dirichlet, who had succeeded to Gauss's professorship in 1855, himself died after a serious illness. Riemann was appointed his successor. In the same year he was elected a corresponding member of the Berlin Academy of Sciences. As a newly elected member, Riemann was required to send a report of his most recent work to the Academy.

His report, titled "On the number of primes less than a given magnitude", was of fundamental importance in number theory. Riemann showed that various results about the distribution of prime numbers were related to the analytic properties of the Riemann zeta function. Some of his results were rigorously established by Hadamard and Vallée-Poussin in 1896, but the Riemann hypothesis remains unproved.

Riemann married in July 1862 and later that year, developed tuberculosis. In order to recuperate, he travelled to Italy several times, befriending the mathematicians Betti and Beltrami.[6] He died in the Italian village of Selasca, where he spent his last weeks with his wife and three-year-old daughter.

Andre Kolmogorov
Henri Lebesgue

Introduction to Measure Theory (nonmathematical)
Measurable Functions
Measure and Integration

[1] https://www.usna.edu/Users/math/meh/riemann.html, last accessed 1-14-2020
[2] https://www.usna.edu/Users/math/meh/riemann.html
[2] https://www.usna.edu/Users/math/meh/riemann.html
[3] https://www.usna.edu/Users/math/meh/riemann.html
[4] http://mathshistory.st-andrews.ac.uk/Biographies/Riemann.html, last accessed 1-14-2020
[5] https://www.usna.edu/Users/math/meh/riemann.html
[6] https://www.usna.edu/Users/math/meh/riemann.html

[*] This involves the Riemann zeta function, which is a function ζ(*s*) of a complex variable *s* defined as follows. If the real part of *s* is greater than 1, define ζ(*s*) to be the sum of the convergent series $\sum_{n \ \&ge\ 1} n^{-s}$, then extend ζ(s) to the whole complex plane by analytic continuation. The Riemann hypothesis states: “If ζ(s) = 0 and the real part of s is between 0 and 1, then the real part of s is exactly 1/2.” This seemingly esoteric condition is of fundamental importance for the distribution of prime numbers.

# Georg Cantor

The one mathematician above all who is responsible for catapulting set theory from an arcane finite and useful contrivance to the basis of modern mathematics is Georg Cantor.

He probably died for it.

The oldest of six children, Georg Cantor was known at an early age for his abilities as a violinist.[1] Although born in the western merchant colony of St. Petersburg, Russia, his family moved to Germany in part to escape the brutal Russian winters. In 1862 Cantor, pleased that his father gave him permission to study mathematics, began his study in Zurich. However, this was interrupted by his father's death a year later. [2] After receiving a substantial inheritance, Cantor moved his work to Berlin where he completed his dissertation on number theory in 1867.

Cantor accepted a position at the University of Halle, where he spent his entire career, married Vally Guttmann and started a family that eventually expanded to six children.

During this time, Cantor entered into correspondence with Richard Dedekin who through a series of challenges and questions, helped propel Cantors work on rational and irrational numbers. The work was illuminating, and to some shocking. Gösta Mittag-Leffler, in response to one of Cantor's submissions to his journal jested, saying that Cantor's writing was "about one hundred years too soon." Cantor, sensitive about criticism, was injured by this comment.[2]

This was also the reaction to Cantor's work on set theory.

Before Cantor, set theory was an interesting but boring back eddy in mathematics. The number of set elements was always finite, and with that the field was concise but constrained with no room for growth.

In ten years, Cantor turned that idea on its head.

Between 1874 and 1884, Cantor focused on the concept of infinity, which up until that time had been more the philosopher's purview than the mathematicians.

It was well known for example that the number of whole numbers was infinite, and it followed that there must be an infinite number of rational numbers as well (since whole numbers are themselves rational). However, there is an infinite number of rational numbers in $\left[0,1\right]$. Adding that infinite set to the infinite set of whole numbers must mean, it was thought, that there were more rational numbers than whole numbers. Yet, both sets were infinite. How could there be more numbers than those contained in an infinite set?

Cantor placed his efforts here. He first defined finite and infinite sets, then divided the infinite sets into "denumerable" or countable versus nondenumerable or uncountable sets. He introduced the power set of a set $A$, which is the set of all possible subsets of $A$. He later proved that the size of the power set of $A$ is strictly larger than the size of $A$, even when $A$ is an infinite set; this result soon became known as Cantor's theorem. These are cornerstone of modern set theory.

However this was also a bombshell in a world that conflated high mathematics with divinity. At the time there was one and only one concept of infinity, and according to the religious culture of the day, infinity was were God lived.

Critics concluded that Cantor's work denied the "one God, one infinity" assumption. They then pushed, saying that he denied the existence of one God, and that the multiple infinity concept − since it must imply multiple Gods − meant Cantor was a pantheist.

Cantor understood that the concept of the existence of an actual infinity was an important shared concern within the realms of mathematics, philosophy and religion; preserving the relationships was important to him. But, he was driven to continue with his development of an entire theory and arithmetic of infinite sets, called cardinals and ordinals, which extended the arithmetic of the natural numbers.

His notation for the cardinal numbers was the Hebrew character ℵ with a natural number subscript. This represents the size of an infinite set. For example, $\aleph_0$ is the cardinality of the whole numbers. He affirmed that there could be an infinite number of sets that each have greater number of elements than other infinite numbers of sets.

Infinite sets could be of different sizes and therefore different cardinalities. Despite our imaginations, infinity was far more complex than anyone imagined. Yet the denouncements continued.

Cantor continued his pioneering work, even in the face of growing admonitions and negativism from one of his long standing colleagues - Dedekin. He developed the one-to-one concept which is a foundation of set theory. By developing his famous Cantor set, he demonstrated the existence of sets of real numbers that have the same cardinality as the real numbers, but were nowhere dense (just as the natural numbers are nowhere dense). Rational sets on the other hand, were dense but countable.

These distinctions caused havoc with a bedrock of mathematics – the real number line, and the biting remonstrations continued.

Cantor suffered his first known bout of depression in 1884 after a damaging series of attacks on his work by Kronecker. Doubting whether he would ever return to mathematics, he was place in a sanatorium in 1899. He recovered within a few weeks, but shortly thereafter, his youngest son died.

After a paper denouncing his work was presented by König at the Third International Congress of Mathematicians to an audience including Cantor's colleagues, wife, and daughters, Cantor relapsed. He suffered from chronic depression for the rest of his life. Retiring in 1913, he lived out the rest of his life in poverty until he died in a sanatorium in 1919.

[1] https://en.wikipedia.org/wiki/Georg_Cantor

[2] http://mathshistory.st-andrews.ac.uk/Biographies/Cantor.html

# Thomas Joannes Stieltjes

Stieltjes was born in Zwolle, Holland in 1856.[1] His father was a well renowned civil engineer and a member of the Dutch Parliament, permitting his son to gain entrance to the university at the Polytechnical School in Delft in 1873. However Stieltjes repeated failed his technical exams, preferring to read the works of Gauss and Jacobi.

His father help his gain employment at the Leiden Observatory. While there, Stieltjes began a lifelong correspondence with Charles Hermite in celestial mechanics and mathematics, devoting his spare time to mathematical research. He began a lifelong correspondence with Hermite, writing 432 letters in 12 years [2]

However, in 1883, Stieltjes besieged the director of the observatory to release him from his obligatory observational work so that he could devote more time to mathematics. His wife agreed to support him, and Stieltjes devoted himself to mathematics.

Stieltjes proposed an important generalization of the integral for studying continued fractions. Combined with Bernhard Riemann's definition and now known as the Riemann-Stieltjes integral, it is widely used for applications in physics. Commonly theoreticians affix the name of this mathematician to integration. Riemann integration is sometimes referred to as Riemann-Stieltjes integration, and as in this treatise, Lebesgue integration is referred to as Lebesgue-Stieltjes integrals.

In addition, his work on continued fractions was the first general treatment of the subject as a part of complex analysis and laid the groundwork for the development of Hilbert spaces—infinite-dimensional vector spaces, developed by the German mathematician David Hilbert, that were later used in formulating quantum mechanics.[3]

After many years, and the intervention of Hermite, Stieltjes received an honorary doctorate from Leiden University, enabling him to become a professor where he first started his training.

References

[1] http://www.britannica.com/biography/Thomas-Jan-Stieltjes
[2] http://mathshistory.st-andrews.ac.uk/Biographies/Stieltjes.html
[3] https://en.wikipedia.org/wiki/Thomas_Joannes_Stieltjes

# Henri Lebesgue

*Reduced to general theories, mathematics would be a beautiful form without content. It would quickly die.*

In 1875, Henri Lebesgue (pronounced La-BÁK) was born in Beauvais, France. His father, a typesetter, and mother, a school teacher, created a library for their son that he used from an early age. After his father died of tuberculosis, his mother was forced to support the two of them.

Henri Lebesgue's life might have been quite inconsequential but for one of his teachers. Lebesgue at an early age displayed uncharacteristically high mathematical aptitude. However, his mother had no funds to commit to her son's higher education. A teacher of Lebesgue was also in no position to financially support. However, this teacher, in a remarkable act of self-sacrifice, raised money from the citizens of the local communities for Henri to attend school. With this benevolence, Lebesgue could attend the Collège de Beauvais and then at Lycée Saint-Louis and Lycée Louis-le-Grand in Paris.1

Lebesgue entered the and was awarded his teaching diploma in mathematics in 1897. It was when he matriculated at École Normale Supérieure in Paris that he learned of Émile Borel's work on the rudiments of measure theory, and Camille Jordan's work on Jordan measure. For the next two years he studied in its library where he read Baire's papers on discontinuous functions and realized that much more could be achieved in this area.

Lebesgue's first paper was published in 1898 and was titled "Sur l'approximation des fonctions". It dealt with Weierstrass' theorem on approximation to continuous functions by polynomials. His next papers focused on surfaces applicable to a plane, the area of skew polygons, surface integrals of minimum area with a given bound, and the final note gave the definition of Lebesgue integration for a function $f(x)$.

He was appointed professor at the Lycée Centrale at Nancy where he taught from 1899 to 1902.

Building on the work of others, including that of Émile Borel and Camille Jordan, Lebesgue formulated the theory of measure in 1901 and in his famous paper Sur une généralisation de l'intégrale définie, which appeared in the Comptes Rendus on 29 April 1901, he gave the definition of the Lebesgue integral that generalizes the notion of the Riemann integral by extending the concept of the area below a curve to include many discontinuous functions.

This all occurred before he earned his PhD

This generalisation of the Riemann integral revolutionized the integral calculus. In 1902 he earned his Ph.D. from the Sorbonne with the seminal 130 page thesis on "Integral, Length, Area", submitted with Borel, four years older, as advisor. This work expanded the application of integral calculus to intensely discontinuous functions, revolutionizing the field [2]

Curiously, Lebesgue did not concentrate throughout his career on the field which he had himself started. This was because his work was a striking generalisation, yet Lebesgue himself was fearful of generalizations. He made major contributions in other areas of mathematics, including topology, potential theory, the Dirichlet problem, the calculus of variations, set theory, the theory of surface area and dimension theory.

By 1922 when he published Notice sur les travaux scientifique de M Henri Lebesgue he had written nearly 90 books and papers. After 1922 he remained active, but his contributions were directed towards pedagogical issues, historical work, and elementary geometry.

Andre Kolmogorov
Bernhard Riemann

Introduction to Measure Theory (nonmathematical)
Measurable Functions
Measure and Integration

[1] . https://en.wikipedia.org/wiki/Henri_Lebesgue
[2] http://mathshistory.st-andrews.ac.uk/Biographies/Lebesgue.html

# Giuseppe Vitali

Giuseppe Vitali (26 August 1875 – 29 February 1932) was an Italian mathematician who made several contributions to several branches of mathematical analysis. He is most famous for the Vitali set with which he was the first to give an example of a non-measurable subset of real numbers.

Cared for at home by his mother Zenobia Casadio who looked after him and his four siblings,[1] his early schooling took place in Ravenna, where his prowess in mathematics was not particularly encouraging. However, his performance in secondary school improved so much that his instructor, Giuseppe Nonni advised his son's training, urging him to continue in mathematics.

Vitali then studied for two years at the University of Bologna, beginning in the autumn of 1895. His main teachers at Bologna were sufficiently impressed by Vitali that they supported his application for a scholarship to study at the Scuola Normale Superiore in Pisa.

Vitali was awarded the scholarship and began his studies at Pisa in the autumn of 1897. He published three papers in 1900, and earned his teaching diploma in 1902.

After the award of his teaching diploma, Vitali left university level mathematics to teach high school. This was likely due to financial difficulties. He ultimately returned to mathematics after a period of political activity.

His significant mathematical discoveries include a theorem on set-covering, the notion of an absolutely continuous function and a criteria for the closure of a system of orthogonal functions. Since he worked very much on his own, his work involves some rediscovering of known results but also some remarkably original discoveries.

His most significant output took place in the first eight years of the twentieth century when Lebesgue's concepts of measure and integration were revolutionizing the principles of the theory of functions of real variables. He provided examples of nonmeasurable subsets of real numbers. [2]

In 1926, Vitali developed a serious illness and, with a paralyzed arm, he could no longer write. Nevertheless about half his research papers were written in the last four years of his life after the illness struck.

On 29th February 1932 he delivered a lecture at the University of Bologna and was walking in conversation with fellow mathematician Ettore Bortolotti when he collapsed and died in the street. He was 56 years of age.

References

[1] http://mathshistory.st-andrews.ac.uk/Biographies/Vitali.html

[2] https://en.wikipedia.org/wiki/Giuseppe_Vitali

# The Bernoullis

Leon Bernoulli was a doctor in Antwerp, which at that time was in the Spanish Netherlands. The family, of Belgium origin, were refugees fleeing from persecution by the Spanish rulers of the Netherlands. Nicolaus Bernoulli was an important citizen of Basel, being a member of the town council and a magistrate. He had three sons

- Jacob Bernoulli (1654–1705; also known as James or Jacques) Mathematician after whom Bernoulli numbers are named.
- Nicolaus Bernoulli (1662–1716) Painter and alderman of Basel.
- Johann Bernoulli (1667–1748; also known as Jean) Mathematician and early adopter of infinitesimal calculus.*

Originally from Antwerp, they settled in Basel, Switzerland. He counseled one to go into theology, the other, medicine. However, both went their own ways, choosing mathematics over his strenuous objections. This resistance was not uncommon in the Bernoulli family, as they chose against the more lucrative business careers advocated by their father.

## Jacob Bernoulli

During the time that Jacob Bernoulli was taking his university degrees he began studying mathematics and astronomy against the wishes of his parents.

However, until 1676 he followed his family’s wished and completed a degree in theology. Moving to Geneva, Jacob, worked as a tutor, then traveled to France where he spent two years studying under Descarte. In 1681 Bernoulli travelled to the Netherlands where he met many mathematicians including Hudde, then he went to England where he met Boyle and Hooke. As a result of his travels, Bernoulli began a correspondence with many mathematicians which he carried on over many years.

Having discovered his true love for mathematics and theoretical physics, he lectured and taught on these topics regularly. In 1683, Jacob returned to Switzerland, delivering a collection of important lectures on solids and liquids at the Univeristy of Basel in 1683. A year later, he married Judith Stupanus. They were to have two children, a son who was given his grandfather's name of Nicolaus and a daughter. These children, unlike many members of the Bernoulli family, did not go on to become mathematicians or physicists.

One of the most significant events concerning the mathematical studies of Jacob Bernoulli occurred when his younger brother, Johann Bernoulli, began to work on mathematical topics. Johann was told by his father to stay away from mathematics and to study medicine, but while he was studying that topic he asked his brother Jacob to teach him mathematics. They also studied the publications of von Tschirnhaus. The Bernoullis were the first to try to understand and apply the rather abstract ideas of Leibniz's on the calculus.

Jacob Bernoulli's first important contributions were a pamphlet on the parallels of logic and algebra published in 1685. In addition, Jacob Bernoulli published five treatises on infinite

* Infinitesimal calculus is a synonym for what we know as simply “calculus”

series between 1682 and 1704. Bernoulli also studied the exponential series which came out of examining compound interest.

In May 1690 in a paper published in *Acta Eruditorum,* Jacob Bernoulli showed that the problem of determining the curve along which a particle will descend under gravity from any point to the bottom in exactly the same time, After finding the differential equation, Bernoulli then solved it by what we now call separation of variables. Jacob Bernoulli's paper of 1690 is important for the history of calculus, since the term integral appears for the first time with its integration meaning. [1]

Jacob Bernoulli's most original work was *Ars Conjectandi* published in Basel in 1713, eight years after his death. The work was incomplete at the time of his death but it is still a work of the greatest significance in the theory of probability.

## Johann Bernoulli

His parents tried to set Johann Bernoulli on the road to a business career but, like his other brother Jacob, he disliked it intensely. While taking courses in medicine, he studied mathematics with his brother Jacob, who lectured on experimental physics there, where they both immersed themselves in studying Leibniz's calculus papers

Making his way to Paris, he met de l'Hôpital, where they engaged in many deep mathematical conversations, during which he instructed de l'Hôpital in the calculus. After Johann left Paris, de l'Hôpital published the first calculus book, which includes what we now call l'Hôpital's Rule, a publication that without an acknowledgment of Johann’s own lectures, distressed him.

However, Johann began a lengthy series of conversations with Leibniz which was to prove very fruitful. In 1694 he investigated series using the method of integration by parts. He summed series, and discovered addition theorems for trigonometric and hyperbolic functions using the differential equations they satisfy. Johann also had great success in integrating differential equations. As a result he was offered the chair of mathematics at Groningen.

Johann married and had several children, three of which also became mathematicians: Nicolaus, Daniel, and Johann. Oddly, just as his father had done, Johann tried to force his own son Daniel from going into mathematics and physics.

Johann did compete against his brother Jacob, a competition that became contentious, and bitter. Yet, it was at the request of his father-in-law that, Johann began a sea voyage back to his home town of Basel in 1705. It was just after he began this voyage that Johann learned of his older brother’s death.[2]. Johann was given his brother’s chair of mathematics, where he made important contributions to mechanics with his work on kinetic energy.

It was in 1713 that Johann became involved in the Newton-Leibniz controversy. He strongly supported Leibniz and added weight to the argument by showing the power of his calculus in solving certain problems which Newton had failed to solve with his methods. It was Johann’s influential support that delayed acceptance of Newton's physics in Europe.

Johann received great acclaim during his lifetime, and was called the "Archimedes of his age". This is inscribed on his tombstone.

## Daniel Bernoulli

Daniel Bernoulli was the son of Johann Bernoulli. Like his father, he rebelled against parental direction for his career. Unfortunately, there was also interfamily rivalry, jealousy and bitterness.

Daniel was sent to Basel University at the age of 13 to study philosophy and logic. He obtained his baccalaureate examinations in 1715 and went on to obtain his master's degree in 1716. Johann was determined that Daniel should become a merchant and he tried to place him in an apprenticeship.

In Venice Daniel was severely ill and so was unable to carry out his intention of travelling to Padua to further his medical studies. However, while in Venice he worked on mathematics and his first mathematical work was published in 1724 when, with Gold Bach's assistance, Mathematical exercises was published.

Bernoulli's interests were eclectic. While in Venice, Daniel he designed an hour glass to be used at sea so that the trickle of sand was constant even when the ship was rolling in heavy seas. He submitted his work on this to the Paris Academy and in 1725, the year he returned from Italy to Basel, he learnt that he had won the prize of the Paris Academy.

Daniel had also attained fame through his work Mathematical exercises and on the strength of this he was invited to take up the chair of mathematics at St Petersburg. Meanwhile, his brother Nicolaus(II) Bernoulli was also offered a chair of mathematics at St Petersburg so in late 1725 the two brothers travelled to St Petersburg.

Within eight months of their taking up the appointments in St Petersburg Daniel's brother died of fever. He thought of returning to Basel and wrote to his father telling him how unhappy he was in St Petersburg. Johann Bernoulli was able to arrange for one of his best pupils, Leonard Euler, to go to St Petersburg to work with Daniel. Undoubtedly the most important work which Daniel Bernoulli did while in St Petersburg was his work on hydrodynamics.

This work contains for the first time the correct analysis of water flowing from a hole in a container. [3]One other remarkable discovery appears in Chapter 10 of Hydrodynamica where Daniel discussed the basis for the kinetic theory of gases. He was able to give the basic laws for the theory of gases and gave, although not in full detail, the equation of state discovered by Van der Waals a century later.

He and his younger brother left St Petersburg in 1733, making visits to Danzig, Hamburg, Holland and Paris before returning to Basel in 1734. Although Daniel had left St Petersburg, he continued his collaboration with Euler; and the two exchanged many ideas on vibrating systems

Intense father-son rivalry attended Daniel Bernoulli's entry regarding astronomy for the 1734 Grand Prize of the Paris Academy, His father also submitted an entry and the father-son pair were declared joint winners of the Grand Prize. Their joint victory intensified the rivalry between them.[4]

Daniel Bernoulli's accepted of many of Newton's theories and his use of these together with the results generated from Leibniz's calculus fueled his work

Daniel Bernoulli was much honored in his own lifetime. He was elected to most of the leading scientific societies of his day including those in Bologna, St Petersburg, Berlin, Paris, London, Bern, Turin, Zurich and Mannheim.

In addition to Jacob, Johann, and Daniel, the Bernoulli family produced many notable artists and scientists, with particular emphasis on mathematics. in:

- Nicolaus I Bernoulli (1687–1759) Mathematician.
- Nicolaus II Bernoulli (1695–1726) Mathematician; worked on curves, differential equations, and probability.
- Daniel Bernoulli (1700–1782) Developer of Bernoulli's principle and St. Petersburg paradox.
- Johann II Bernoulli (1710–1790; also known as Jean) Mathematician and physicist.
- Johann III Bernoulli (1744–1807; also known as Jean) Astronomer, geographer, and mathematician.
- Jacob II Bernoulli (1759–1789; also known as Jacques) Physicist and mathematician.

Bernoulli Distribution and Introduction to Random Variables

Skewness and Kurtosis for the Bernoulli Distribution.
Moment Generating and Probability Generating Functions

References

1 . https://en.wikipedia.org/wiki/Jacob_Bernoulli
2 . https://en.wikipedia.org/wiki/Johann_Bernoulli
3 https://www.britannica.com/biography/Daniel-Bernoulli
4 https://en.wikipedia.org/wiki/Daniel_Bernoulli

# Alexandre-Théophile Vandermonde

The founder of the theory of determinants, Alexandre Théophile Vandermonde was born in Paris on February 28, 1735. His work in mathematics was limited to only two years.

He was plagued by ill health his entire life, and as a boy took the advice of his parents. His father, a physician who after spending twelve years in the Orient and now practicing medicine in Paris, encouraged his son to avoid the physically demanding medical profession and instead study music which his bright son thoroughly enjoyed.

He pursued a music career, but abruptly changed direction at thirty five years of age. At this point, he devoted himself to mathematics, and from 1771-1773 presented four papers to the Académie des Sciences in 1771, elected to the academy after the first one. These four papers represent his total mathematical output.

The first of these four papers presented a formula for the sum of the $m^{\text{th}}$ powers of the roots of an equation. It also presented a formula for the sum of the symmetric functions of the powers of such roots, using original techniques separate from those of others who published beforehand.

In his second paper Vandermonde considered the problem of the knight's tour on the *This paper is an early example of the study of topological ideas. Vandermonde's consideration of the intertwining of the curves generated by the moving knight were extended first by Gauss and then Maxwell to be useful in generating the ideas of electrical circuits.

In his third paper Vandermonde studied combinatorial ideas. At the time there was no symbol to represent the factorial. He defined the symbol

$$[p]^n = p(p - 1)(p - 2)(p - 3) \ldots (p - n + 1)$$

and

$$[p]^{-n} = 1 / \{(p + 1)(p + 2)(p + 3) \ldots (p + n)\}.$$

In addition, he gave an identity for the expansion of $[x + y]^n$ .

The final of Vandermonde's four papers studied the theory of determinants and is credited with being the first mathematician to prepare a systematic treatment of the theory of determinants.[1] The determinant named after him is one in which the elements of each row or column are: 1, r, $r^{2,}$ ..., $r^{n-1}$ of a geometric progression.

After 1772, Vandermonde dropped out of the mathematics world as precipitously as he entered it. Vandermonde's election to the Académie des Sciences did motivate him to work hard for the Academy and to publish other works on science and music.

Revolution began with the storming of the Bastille on 14 July 1789. The politics of Revolution in France long before this event had been so exciting for Vandermonde, and some speculate that this intense interest in concert with declining health sapped much of his intellectual and physical strength. [2]

* The object of this chess puzzle it to construct the sequence of moves by a knight chess piece on a chessboard, permiting the knight to land on each square of the board one and only one time.

He died on January 1, 1796 in Paris.

References

1. https://en.wikipedia.org/wiki/Alexandre-Th%C3%A9ophile_Vandermonde
2 . http://mathshistory.st-andrews.ac.uk/Biographies/Vandermonde.html

# Augustin Louis Cauchy

*"He met his death with such calm that made us ashamed of our unhappiness"*

The French mathematician Augustin Louis Cauchy (1789-1857) provided the foundation for modern rigor in mathematical analysis.[1]

Born in Paris on Aug. 21, 1789, 38 days after the fall of the Bastille, the first world he knew was one of political upheaval.[2] His father, Louis François, was a parliamentary lawyer, lieutenant of police, and ardent royalist. Sensing the political wind, he moved the family to his country cottage at Arcueil, where they lived for nearly 11 years, themselves schooling their son. The family moved back to Paris in 1800 when the political situation stabilized. Laplace and Lagrange were visitors at the Cauchy family home and Lagrange in particular seems to have taken an interest in young Cauchy's mathematical education, advising Cauchy's father to continue his education.[2]

In 1810 Cauchy took up his first job in Cherbourg to work on the port facilities for Napoleon's English invasion fleet. In addition to his heavy workload Cauchy undertook mathematical researches, proving in 1811 that the angles of a convex polyhedron are determined by its faces.

In his spare time he began to review all mathematics, "clearing up obscurities" and inventing new methods for the "simplification of proofs and the discovery of new propositions", giving the word "determinant" its modern meaning.

Cauchy felt that he had to return to Paris if he was to make an impression with mathematical research, but apparently suffered a severe depressive episode. Recovering he arrived in Paris, he chose to stay in Paris and pursue his career.

Cauchy's work as an academician was transformative. Cauchy was a brilliant academic success, although his countenance as an evangelical Catholic creative difficulty for him.

Cauchy understood the power of the printing press. There were occasions when he would produce two full-length papers in one week. He overwhelmed the community of mathematicians with his published word.

At the age of 27 Cauchy was elected to the Académie des Sciences-an unusual honor for so young a man.

Cauchy cemented the logical foundation of differential calculus with the concept of the limit. His definition of continuity and the derivative in terms of limits presaged future constructions. In addition, he founded complex functional analysis

Cauchy permitted politics to adumbrate his work when he refused to take an oath of allegiance to King Philippe after having sworn and oath to King Charles. Stripped of all his positions, he left his family in Paris, and began a period of self-exile in Switzerland.

Cauchy died on May 23, 1857, after a short illness. The following is a quote from one of his children.

> *"Having remained fully alert, in complete control of his mental powers, until 3.30 a.m.. my father suddenly uttered the blessed names of Jesus, Mary and Joseph. For the first time, he seemed to be aware of the gravity of his condition. At about four o'clock, his soul went to God. He met his death with such calm that made us ashamed of our unhappiness."*
>
> *His last words were "Men die but their works endure."*

Limits and continuous functions

References

[1] https://en.wikipedia.org/wiki/Augustin-Louis_Cauchy

[2] http://mathshistory.st-andrews.ac.uk/Biographies/Cauchy.html

# Siméon-Denis Poisson

Born at Pithviers on June 21, 1781, Simeon Poisson[*] developed many novel applications of mathematics for statistics and physics, contributions that came after a stunning single day coincidence.

Siméon's father had been a private soldier, and on his retirement was given a small administrative post in his native village. When the French revolution broke out, his father assumed the government of the village, and soon became a local dignitary. Commonly required to be away from home most of the day, he left Siméon in the care of a nursemaid, who, caring for the young boy in a countryside home, struggled with her own desperate fear of wild animals nearby. Meanwhile, young Siméon asked to play outside.

She allowed him to play, but only in the manner that permitted her to quell her fear of his being devoured by local beasts. She first spied a nail high up on an outside wall and quickly found a cord. She promptly secured the cord around the nail, and lifting Simeon, she fastened the other end to his clothes, leaving him suspended. Now feeling at ease, she allowed the boy to dangle high enough above the ground and protected from the animals she supposed were lurking nearby. And this he did, hour and hour, day in and day out. [1]

Yet Siméon enjoyed this, entertaining himself by swinging from side to side for hours at a time. It was this activity that he identified years later as the force behind his attraction for the mechanics and process of pendular motion.

Educated by his father, he was both encouraged and pushed to go into medicine. His uncle, a physician himself, offered to teach him the craft, and to give Siméon a feel for the field, started him repetitively pricking the veins of cabbage-leaves with a lancet. [2] Young Poisson soon mastered this and was given permission to begin lancing boils on humans. However, after lancing a boil on his first patient who died several hours later, the distraught Poisson swore he would have nothing more to do with medicine.

When he returned home, he stumbled across a question set among his father's official papers. They were from the Polytechnic school. He was fascinated by their mathematical structure.

This day's horrors and discoveries determined Poisson' career.

At the age of seventeen he entered the Polytechic. One year later, his first paper, focused on a discussion of finite differences impressed his colleagues and quickly appeared in an important journal. As soon as he had finished his studies, he was appointed as a lecturer.

Throughout Poisson's life he held various scientific posts and professorships. He made the study of mathematics his hobby as well as his business, writing between 300-400 manuscripts and books on a variety of mathematical topics, including pure mathematics, the application of mathematics to physical problems, the probability of random events, the theory of electrostatics and magnetism (which led the forefront of the new field of quantum mechanics), physical astronomy, and wave theory.

One of Siméon Poisson's lasting contributions was the development of equations to analyze random events, later dubbed the Poisson distribution.

[*] Taken from http://www.sci.sdsu.edu/~ smaloy/MicrobialGenetics/topics/ mutations/poisson.html, and Wikipedia.

The fame of this distribution is often attributed to the following story. Many soldiers in the Prussian Army died due to kicks from horses. To determine whether this was due to a random occurrence or the wrath of god, the Czar commissioned the Russian mathematician Ladislaus Bortkiewicz to determine the statistical significance of the events.[3] Fourteen corps were examined, each for twenty years providing a wealth of information time. For over half this period, there were absolutely no deaths from horse kicks; for the remaining half, the number death ranged from 1-4, identifying the event as relatively uncommon. Poisson applied his distribution and found that it fit remarkably well.

His excellent writing on celestial mechanics also stand out, and these contributions put him on a level with Pierre-Simon Laplace. In the earliest of these papers, Poisson improved Lagrange's findings on the stability of planetary orbits.

Poisson's memoir was remarkable inasmuch as it reinvigorated Lagrange in his old age to author his own writings, thereby producing one of his greatest manuscripts. In fact he did Poisson's memoir the honor of making a copy with his own hand, that was posthumously found.

In mathematics, his most important works were a series of papers on definite integrals and his advances in Fourier analysis, which paved the way for the research of the German mathematicians Peter Dirichlet and Bernhard Riemann.

Poisson died April 25, 1840.

[1] http://www.sci.sdsu.edu/~ smaloy/MicrobialGenetics/topics/ mutations/poisson.html

[2] https://en.wikipedia.org/wiki/Sim %C3%A9on_Denis_Poisson

[3] https://onlinelibrary.wiley.com/doi/full/10.1111/anae.13261

# Guido Fubini

Born in January 1879, Guido Fubini demonstrated an early aptitude for mathematics demonstrating in secondary school his powerful aptitude for mathematics.*

In 1896 Fubini entered the Scuola Normale Superiore di Pisa where he was encouraged to undertake geometry.[1] Most young doctoral students take a few years to make themselves well known in their area. However, Fubini was lucky for his teacher Bianchi was about to publish an important work on differential geometry, and he included Fubini's thesis in his treatise.

Going against the grain, upon graduation, Fubini put his thesis aside and began work in a completely different topic. Fubini's interests were exceptionally wide moving from his early work on differential geometry towards analysis. In 1908 Fubini moved to Turin where he taught both at the Politecnico and at the University of Turin.

He taught courses on these analysis topics at both the Politecnico and the University in Turin.

During this time his research focused primarily on topics in mathematical analysis, especially differential equations, functional analysis, and complex analysis. However, when World War I began, like many contemporary scientists, he began working in quantitative areas with military applications. However, at the conclusion of the war, he examined issues in acoustics and electrical circuits.

In mathematical analysis Fubini's theorem, is a result which gives conditions under which it is possible to compute a double integral using iterated integrals. As a consequence it allows the order of integration to be changed in iterated integrals. It is applied in the derivation of the Beta distribution and the normal distribution.

He was nearing the end of his career when the political situation in Italy suddenly put him in an exceptionally difficult position. A series of decrees removed Jews from positions of influence in government, banking and education. Fubini was forced to retire from his chair in Turin. Ultimately, he removed his family to New York City.

Taking an interest in the engineering problems that his sons were solving, he wrote a textbook on the mathematical challenges of this work. The textbook appeared posthumously, jointly authored with G Albenge, This last textbook was one of an impressive collection of important textbooks on analysis which included books which described analysis courses which he had given and also books which were collections of problems.

He died in June 1943.

[1] http://mathshistory.st-andrews.ac.uk/Biographies/Fubini.html

* Developed from http://www-groups.dcs.st-and. ac.uk/history/Biographies/ Fubini.html and http://en.wikipedia.org/wiki/Guido_Fubini

# Agner Krarup Erlang

Agner Krarup Erlang (January 1, 1878 – February 3, 1929) was a Danish mathematician, statistician and engineer, who invented the fields of traffic engineering and queuing theory.[1]

Agner Erlang's mother, Magdalene Krarup, broke with the family tradition that all sons became clergymen and all daughters married clergymen when she married Hans Nielsen Erlang, a schoolmaster and parish clerk. Agner was the second of his parents' four children, having an older brother Frederik and two younger sisters Marie and Ingeborg. His sedulous father and mother made a happy if simple home for their family in difficult financial times.

Agner was a bright child, who preferred reading to playing with the other boys. Educated at his father's school when he was young, he became enamored of astronomy, encouraged by his maternal grandfather who also loved it, but went one step further by writing poems about astronomical objects. After his primary education, Agner was tutored at home by his father and another teacher from his father's school. He took his Praeliminaereksamen examination in Copenhagen at the age of fourteen after having to obtain special permission to take the test because he was below the minimum age and passed with special distinction.

After graduating with majors in mathematics and physics, astronomy and he taught high school for the next seven years. He was quite good at it.

With his heavy red full beard and his manner of dressing he appeared more artist then science technician. However, he was extremely modest, preferring the peaceful atmosphere of his study to festivities; he never touched alcoholic liquors nor smoked tobacco. He never married, devoting himself to collecting books on history, philosophy and poetry.

Friends found him to be a good and generous source of information on many topics. He was known to be a charitable man, needy people often came to him at the laboratory for help, which he would usually give them in an unobtrusive way.

At meetings of the Mathematical Association he met Johan Ludwig Jensen, chief engineer at the Copenhagen Telephone Company, who persuaded Erlang to work with him on problems arising from telephone call waiting times. [2]

Erlang published his first paper on the theory of probability and telephone conversations in 1909. In this paper he showed that if telephone calls were made at random they followed the Poisson distribution, and he gave a partial solution to the delay problem. In 1917 he again published, this time giving a formula for loss and waiting time which was soon used internationally. The Erlang distribution is named for him.

In the twenty years that Erlang worked for the Copenhagen Telephone Company he never had to take a day off through illness. However, in January 1929, at the age of 51 he began suffering from abdominal pains and went into hospital for an operation. He died a few days later.

[1] http://www-history.mcs.st-and.ac.uk/Biographies/Erlang.html

[2] http://en.wikipedia.org/wiki/Agner_Krarup_Erlang.

# Carl Friedrich Gauss

Johann Carl Friedrich Gauss, the Prince of Mathematicians was born April30, 1777. His influence on mathematics which he dubbed the "the queen of sciences".[1][*]

Born into a poor family, his illiterate mother never recorded the date of his birth, only remembering that it was on a Wednesday, forty eight days before Easter.[†]

When he, at three years old, informed his father of a mistake in a complicated payroll calculation stating the correct answer, his family recognized the prodigy they had among them. In school, when his teacher gave the problem of summing the integers from 1 to 100 (an arithmetic series ) to his students (who ranged from 6 to 16 years of age), Carl, turned in the answer (5050) on his slate at once.[‡] At age 19, Gauss demonstrated a method for constructing a heptadecagon using only a straightedge and compass which had eluded the Greeks. Gauss also showed that only regular polygons of a certain number of sides could be constructed in that manner (a heptagon, for example, could not be constructed.)

The year 1796 was most productive for both Gauss and number theory. He discovered a construction of the heptadecagon on 30 March. He further advanced modular arithmetic, greatly simplifying manipulations in number theory. On 8 April he became the first to prove the quadratic reciprocity law, permitting mathematicians to determine the solvability of any quadratic equation in modular arithmetic. The prime number theorem, conjectured on 31 May, gives a good understanding of how the prime numbers are distributed among the integers. Gauss also discovered that every positive integer is representable as a sum of at most three triangular numbers on 10 July and then jotted down in his diary the famous note: "ΕΥΡΗΚΑ!". On October 1 he published a result on the number of solutions of polynomials with coefficients in finite fields, which 150 years later led to the Weil conjectures.[2]

He completed *Disquisitiones Arithmeticae*, his magnum opus, in 1798 at the age of 21, though it was not published until 1801. This work was fundamental in consolidating number theory as a discipline and has shaped the field to the present day.

Gauss proved the fundamental theorem of algebra, which states that every polynomial has a root of the form $a+bi.$ In fact, he gave four different proofs, the first of which appeared in his dissertation. In 1801, he proved the fundamental theorem of arithmetic, which states that every natural number can be represented as the product of primes in only one way.

In 1801, Gauss developed the method of least squares fitting, 10 years before Legendre, but did not publish it. The method enabled him to calculate the orbit of the asteroid Ceres, which had been discovered by Piazzi from only three observations. Piazzi could only track Ceres for a few months, following it for three degrees across the night sky. Then it disappeared temporarily behind the glare of the Sun. Several months later, when Ceres should have reappeared, Piazzi

[*] Taken from http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

[†] Gauss would later solve this puzzle about his birthdate in the context of finding the date of Easter, and in the process derived methods to compute dates in both past and future years.

[‡] It is notable that his teacher, known for his apathetic attitude toward his students, took an interest in Gauss, teaching him much about mathematics and offering him encouragement.

could not locate it.* Gauss, who was 23 at the time, heard about the problem and tackled it. After three months of intense work, he predicted a position for Ceres in December 1801—just about a year after its first sighting—and this turned out to be accurate within a half-degree when it was rediscovered by Franz Xaver von Zach on 31 December at Gotha, and one day later by Heinrich Olbers in Bremen. However, after his independent discovery, Legendre accused Gauss of plagiarism.

Gauss published his monumental treatise on celestial mechanics *Theoria Motus* in 1806. He became interested in the compass through surveying and developed the magnetometer and, with Wilhelm Weber measured the intensity of magnetic forces. With Weber, he also built the first successful telegraph.

Gauss is reported to have said "There have been only three epoch-making mathematicians: Archimedes, Newton and Eisenstein. There is also a story (perhaps apocryphal) that in 1807 he was interrupted in the middle of a problem and told that his wife was dying. He is purported to have said, "Tell her to wait a moment 'til I'm through".

Gauss arrived at important results on the parallel postulate, but failed to publish them. Credit for the discovery of non-Euclidean geometry therefore went to Janos Bolyai and Lobachevski. However, he did publish his seminal work on differential geometry in *Disquisitiones circa superficies curves*. The Gaussian curvature (or "second" curvature) is named for him. He also discovered the Cauchy integral theorem for analytic functions, but did not publish it.

Gauss's personal life was overshadowed by the early death of his first wife, Johanna Osthoff, in 1809, soon followed by the death of one child, Louis. Gauss plunged into a depression from which he never fully recovered. He married again, to Johanna's best friend named Friederica Wilhelmine Waldeck but commonly known as Minna. When she died in 1831 after a long illness, one of his daughters, Therese, took over the household and cared for Gauss until the end of his life.

Gauss was an ardent perfectionist and a hard worker. He was never a prolific writer, refusing to publish work which he did not consider complete and above criticism. This was in keeping with his personal motto pauca sed matura ("few, but ripe"). His personal diaries indicate that he had made several important mathematical discoveries years or decades before his contemporaries published them. Many of his results were subsequently repeated by others, since his terse diary remained unpublished for years after his death. This diary was only 19 pages long, but later confirmed his priority on many results he had not published.

Though he did take in a few students, Gauss was known to dislike teaching. It is said that he attended only a single scientific conference, which was in Berlin in 1828. However, several of his students became influential mathematicians, among them Richard Dedekind, Bernhard Riemann, and Friedrich Bessel. Before she died, Sophie Germain was recommended by Gauss to receive her honorary degree.

In 1831 Gauss developed a fruitful collaboration with the physics professor Wilhelm Weber, leading to new knowledge in magnetism (including finding a representation for the unit of magnetism in terms of mass, length and time) and the discovery of Kirchhoff's circuit laws in electricity. It was during this time that he formulated his namesake law. In 1840, Gauss published his influential Dioptrische Untersuchungen, in which he gave the first systematic analysis on the formation of images under a paraxial approximation (Gaussian optics). Among his results, Gauss showed that under a paraxial approximation an optical system can be characterized by its cardinal points and he derived the Gaussian lens formula.

* The mathematical tools of the time were not able to extrapolate a position from such a scant amount of data.

In 1854, Gauss notably selected the topic for Bernhard Riemann's now famous Habilitationvortrag, Über die Hypothesen, welche der Geometrie zu Grunde liegen. On the way home from Riemann's lecture, Weber reported that Gauss was full of praise and excitement.

Gauss died in Göttingen, in the Kingdom of Hannover in 1855 and is interred in the Albanifriedhof cemetery there.

[1] http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

[2] . http://scienceworld.wolfram.com/biography/Gauss.html.

# Sydney Chapman

Sydney Chapmen, and English mathematician, co-discovered the system of difference-differential equations used in queuing theory, operations research, and epidemiology with Andrey Kolmogorov.

Chapman initially engineering at the University of Manchester but become so enthusiastic for mathematics that he stayed for one further year to take a mathematics degree. He held the Beyer Chair of Applied Mathematics at Manchester from 1919 to 1924.[1]

In 1946, Chapman was elected to the Sedleian Chair of Natural Philosophy at Oxford, and was appointed fellow of The Queen's College, Oxford. In 1953, on his retirement from Oxford, Chapman took research and teaching opportunities all over the world, including at the University of Alaska and the University of Colorado, as well as opportuniees in Cairo, Istanbul, Prague, and Tokyo.

Chapman's most noted mathematical accomplishments were in the field of Markov processes. He and Kolmogorov independently developed the difference differential equations so useful in many fields.

In 1970, Chapman died in Boulder, Colorado, at the age of 82.The lunar Crater Chapman is named in his honor.

References

[1] https://en.wikipedia.org/wiki/Sydney_Chapman_(mathematician) last accessed April 13, 2020.

# Author Biography

Dr. Lem Moyé, M.D., Ph.D. is a physician, epidemiologist, and biostatistician. He completed his undergraduate training at Johns Hopkins University in 1974. After receiving his M.D. at the Indiana University Medical School, he completed post-doctoral training at Purdue University and the University of Texas.

Dr. Moyé has conducted federally sponsored research for over 30 years, including 12 years investigating cell therapy for heart disease. He has published over 220 manuscripts, 12 books including two novels (*Saving Grace*, and *Catching Cold: Breakthrough*), and has worked with both the US Food and Drug Administration pharmaceutical companies.

Dr. Moyé has taught graduate classes in epidemiology and biostatistics for three decades and has served as an expert witness in both state and federal court. He served as a volunteer physician during the Hurricane Katrina calamity, and his memories of that experience led his prize winning book, *Caring for Katrina's Survivors.*

He is cancer survivor, living in Chandler Arizona with his wife, Dixie.

**René Descartes-1**