Weighing the Evidence: Duality, Set, & Measure Theory in Clinical Research Analyses

2nd Edition

Lem Moyé

Lem Moyé, MD, PhD
Principal Evidence, LLC
5671 S. Wayne Drive
Chandler, AZ 85249
Probability@PrincipalEvidence.com

ISBN

To Dixie and the DELTS

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*
- *Probability and Statistical Inference: Applications, Computations, and Solutions* (with Asha S. Kapadia and Wen Chan)
- *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators*
- *Statistical Reasoning in Medicine: The Intuitive P–Value Primer- 2nd Edition*
- *Face to Face with Katrina's Survivors: A First Responder's Tribute*
- *Elementary Bayesian Biostatistics*
- *Saving Grace – A Novel*

Preface

How would I analyze a clinical trial if there was no such thing as statistical hypothesis testing?

Before we get to that, let's take a moment to get acquainted.

I became a physician in 1978, earned my PhD in biostatistics and epidemiology in 1987, and immediately began a 32 year career as a faculty member at the University of Texas School of Public Health in Houston, Texas.

That was the heyday of clinical trials.

Public health was a well-organized discipline, receiving full support from the scientific community and the National Institutes of Health. Its scientists, working in the government as well as the private sector, were committed to rubbing out chronic disease with the same energy and zeal that helped eradicate or reduce the prevalence of many infectious diseases. My particular focus was cardiology, where the one-two punch of heart attacks and resultant heart failure stole the lives of millions of Americans each year.

Biostatistics was an important instrument on this mantle, and I immersed myself fully. However, while I have always enjoyed the background mathematics, I became concerned about its application to clinical research. The use of estimation theory was exquisite; I have no issue with the computations of means, event rates, hazard ratios or other quantities based on estimation theory.

It was the inference component that wore at me.

Statistical hypothesis testing not only seemed a poor fit (which is no surprise, because it was not designed specifically for biostatistics), but was becoming an increasingly aggressive tail, unmercifully wagging the clinical trial dog.

My first book, *The P-Value Primer* reflected the sincere attempt of a young scientist to sort out the proper role of biostatistics in clinical research. I believed exposition was key, but, while the book did well, clinical researchers continue to lose ground to *p*-value primacy during a time of new confusion about what these values actually mean.

Consider that, while eminent statisticians call for reducing the threshold of statistical significance from 0.05 to 0.005 [1], the American Statistical Association, for the first time in its 177 year history, felt compelled to issue a statement clarifying for its own membership what *p*-values mean and how they should be used, a clarification that itself had to be explained [2].

How can it be that, approximately 95 years after Ronald Fisher's first writings on statistical inference, statisticians remain confused about the interpretation of a device that is experiencing potentially deeper penetration into clinical research? Tightening a metric suffused with confusion was a poor message to send.

Clinical researchers are by and large the victims of these insidious infiltrations of this style of inference. Why do physician-scientists, otherwise so punctilious about clinical measures e.g., an MRI interpretation, or the small movement of a biomarker level in patients with cancer, willingly turn over their data to statistical hypothesis testing with its continued confusion over the interpretation of *p*-values?

It is a provocative question with a simple answer – they are told that they must.

By mentors and department chairs, by journal reviewers and editors, by grant administrators and the US Federal Food and Drug Agency. And the problem now is worse.

So now I am trying a different tack. Rather than just expatiate these issues. I have asked myself the question "How would I analyze data from a clinical research effort if there were no statistical hypothesis testing tools?"

This provocative question instantiated a four year quest on my part, leading me to develop a new construct and new quantitative tools. They involve a concept that I have termed "duality" and also draw on the topics of set and measure theory in mathematics.

The purpose of this text is to expound on each of these topics and demonstrate that their application to clinical research provides new insight and addresses interpretative conundrums that statistical hypothesis testing cannot.

The audience for this book is clinical trial researchers, biostatisticians, epidemiologists, and of course students of these disciplines. This is a wide swath of expertise, and I have worked to use language that all members of these zones of expertise can understand.

Ok. Let's crack on.

References

1 Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George E, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Hua Ho T, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE. Redefine statistical significance. Nat Hum Behav. 2018 Jan;2(1):6-10. doi: 10.1038/s41562-017-0189-z.

2 . Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016 [Epub ahead of print]. Available from: 10.1080/00031305.2016.1154108.

Table of Contents

Introduction

The thesis of this book is that statistical hypothesis testing does not answer the actual questions that the clinical researcher has posed, but instead answers a question that the researcher 1) has not asked, and 2) has no interest in its answer.

My approach to this dilemma is to answer the question, "If we only had estimation theory and not statistical hypothesis testing, how would we analyze clinical research data?" This approach gives clinical researchers direct access to the answers to their fundamental research question, "Does the experimental exposure help my patients or injure them?"

This book begins quite non-mathematically, discussing the philosophical concerns about the use of the *p*-value, and the acculturation of generations of health care researchers to the use of statistical hypothesis testing even though it was not designed for clinical research from its first principals. Its inculcation has led to the institutionalization of physicians, biostatisticians, and administrators, who frankly would be lost without this single number's presence. [*]

[*] Those readers who are already familiar with this dialogue can skip Chapter 1.

The clinical research community has permitted itself to be caught up in the tidal drift generated by the need for a computational, interpretative tool. While this device added structure to research interpretation in the 1950's, it has, in my view, placed restrictions on research design that have nothing to do with biology, pathophysiology, or even logistics but is instead driven by the need to generate a *p*-value based assessment of the impact of the intervention or exposure.

This is not a conspiracy theory book. None of the *p*-value history that I provide is nefarious. While there have been experienced and prominent members of the statistical community who have been influential in reinforcing *p*-value primacy, there is no statistical hypothesis testing Darth Vader in command. In fact many statisticians conduct statistical hypothesis testing because simply 1) that is what has been asked of them, and 2) they know of no alternative. We are ourselves to blame for this confused miasma. Our answer does not reside in a Star Wars villain but in Shakespeare's *Julius Caesar*.

The book combines a new approach – duality theory – with a well-established approach in mathematics – measure theory – to weigh the evidence in a clinical research effort supporting benefit and supporting harm. Duality theory states that an estimator of an effect in a clinical trial, be it a difference in mean change in diastolic blood pressure, or a prevalence ratio, simultaneously contains evidence of benefit and evidence of harm. The evidence for each is extracted.

However since multiple analyses from the same trial commonly utilizes overlapping sets of observations and variables, the redundancy should be quantified and identified. This is the role of quanta analysis and is based

on set and measure theory. The new developed tool (called quanta analysis), has its foundation in the basics of set and measure theory.

The combination of duality, set, and measure theory appears to be new.

The mathematics of measure theory is commonly taught at advanced levels, it need not always be so; this book and is one of the exceptions.

The first examples offered by this book are almost absurdly simple, yet are necessary for the reader to begin to gain some experience and intuition in the use of quanta analyses. As the examples increase in sophistication, the reader can see how duality/quanta analysis assembles risk and benefit in increasingly complex clinical research scenarios.

In the end, the reader will know the theory and operation of this process as well as its strengths and weaknesses. I finish with some additional embellishments that can be useful – even illuminating – if pursued. Clinical trials are mute on these latter issues because there is little methodology to support their inclusion. However, they are of longstanding clinical interest.

This is my work, so I and I alone am responsible for any and all errors. Fortunately, it is easy to publish new book versions, so should mistakes slip into my writing, please point it out and I will correct and release new editions. This work, like life, is a work in progress and requires midcourse adjustments and corrections.

Finally, I have relied on many teachers and workers as my ideas have developed, going back to my days of driving frigid roads through barren winter landscapes to attend advanced probability courses at Purdue University. Barry Davis at the University of Texas School of Public

Health has been a colleague and friend for over 30 years before I retired. I have learned from him at the Coordinating Center for Clinical Trials there. Robert Hardy, Mort Hawkinsw, and Asha Kapadia were fine mentors. Special thanks goes to Hulin Wu, Deijan Lai and and Hongjian Zhu, conversations with whom about the practical use of measure theory were particularly productive.

Ray Lipicky at the FDA was always pushing me to think anew. He always found room in my complacent bonnet to add yet one more bee.

My colleagues and mentors in the SHEP, SAVE, ALLHAT, CARE, SPOTRIAS, and CCTRN trials have each pulled at my intellectual gravity center, adjusting the trajectory of my thoughts, as have colleagues at the NIA and NHLBI.

And also, my thanks to Ms. Shelly Sayre, Ms. Rachel Vojvodic, Ms. Judy Bettencourt, and Ms. Michelle Cohen who for twelve years patiently bore by blackboard scratching, conversational musings, and energetic remonstrations about the current use of biostatistics in health care research.

And, of course, I am indebted to giants in the field of mathematical analysis, e.g., Georg Cantor, Bernhard Riemann, Henri Lebesgue, Thomas Stieltjes, and Andre Kolmogov whose work was omnipresent, and always ready for my study and absorption. Short biographical sketches of each are included at the end of this book.

Lem Moyé
Chandler, Arizona
January, 2020

The 10,000 foot view

Before we get to the granular details, let's get an overall perspective on my approach; the 10,000 foot point of view.

I believe that clinical investigators are simply interested in answering general questions authoritatively. One of the most important of these such questions is "Are subjects who have received the test intervention better off than those in the control group?" In order to answer this question, we need to identify and weigh the evidence for each of benefit and harm from a collection of analyses.

I am assuming for all examples (except those of the chapter on exploratory analyses) that the researchers have conducted a well-designed, concordantly* executed, two-armed, randomized, controlled clinical trial testing and intervention versus control therapy against prospectively declared outcomes of high precision. The investigators simply want an answer to the question

''Are patients in the intervention group better off than those in the control group?"

Now, there are many analyses that these investigators will conduct to address this question. However, the statistical community's argument of parsimony, i.e., the only analyses that are really persuasive in a clinical trial are

* Executed in accordance with the trial's protocol

the analyses of the primary outcomes, implies that these other outcomes and analyses, while clinically relevant, do not determine the final result of the study.

Thus, this statistical reasoning has reduced the clinical trial endeavor from its full panoply of survival, comorbidity, quality of life, physiologic, and biologic markers findings to the monogramatic "positive", "negative", or "uninformative"[*] commonly based on a single endpoint.

For example, a clinical trial assessing the impact of a new therapy on heart failure quite justifiably will choose total mortality as its primary outcome. However, investigators also assess measures of morbidity (hospitalization for heart failures, number of days patients are not hospitalized for heart failure (hospital free days), and exercise tolerance.

In addition, they will have quantitative assessments of heart function (left ventricular end systolic volume, left ventricular end diastolic volume, sphericity index). And they can in addition include a number of proteomic[†] measures such as brain natriuretic peptide. A comparison of these measures across the two therapy groups provide an assessment of the change in the subject that may have been produced by their therapy.

Yet, these outcome measures beyond the single primary outcome are only considered in a secondary or auxiliary role in the standard clinical trial analyses. While it is not fair to say that they are ignored, they are deemphasized. Why that is so will be discussed Chapter Two.

---

[*] An uninformative result is a finding that is not statistically significant, but underpowered.

[†] the identification of proteins that are produced from specific organs whose presence can indicate the degree of health.

Standard statistical treatments do not permit the quantitative combination of different analysis results in a clinical trial into a single expression of effect. This is not a failure of statistical hypothesis testing as much as it is the inertia in a field that analyses and interprets one outcome at a time.

Yet in health care, physicians must analyze multiple findings simultaneously – we must integrate them. This integration is conducted cerebrally, not mathematically in clinical practice and has historically followed the same development in clinical research. In this book, we will produce an ensemble summary of the results of all analyses executed in a clinical trial that are responsive to a particular question.

Thus, the two methodologic goals of this book are to, from the entire set of analyses conducted in a clinical trial that are carried out to address a particular question, 1) examine the finding of each analysis, parsing out the component of the finding that supports benefit, 2) channel these results into and through a benefit function, and finally 3) accumulate these benefit findings over all of the analyses (accumulation is principally the same as "integration" in mathematics). We will also carry out the same procedure for each analysis, extracting the components of analyses that support harm, using duality theory.

What is duality?

Duality is the property of an estimator in clinical trial that allows it to simultaneously provide evidence of benefit and a finding of harm. While this property can be confusing to traditional statisticians, it is nothing new to clinicians who become accustomed to handling lab tests whose finding are unclear.

As an example, consider a physician who orders a baseline serum creatinine level on a patient who is about to start a course of a nephrotoxic drug. This baseline finding is 0.95 mg/dl.

Completely normal.

The physician then has the measurement repeated after the patient has been on the medication for several days.

The repeat value is 1.08. The upper limit of normal is 1.1

While it is possible for the physician to decide that the creatinine level is still normal, many doctors would give this second estimate some additional consideration. Undoubtedly, the creatinine value is still "normal" and supports the notion that the drug has not been harmful.

However, the single value of 1.08 permits another perspective. There is the variability introduced by the (im)precision of the estimate. In addition, there are physiologic effects that could change the creatinine value, e.g., the patient's hydration state, and, of course, the drug itself. The value of 1.08 might reflect the beginning of toxicity.

Put another way, there is a region around the value of 1.08, which we might call the region of plausible creatinine levels. Part of the interval may reflect normal creatinine values; another part of the interval (that which is greater than 1.1) may reflect abnormal values. The single value of 1.08, because of its imprecision and the myriad influences on it, generates a wide plausibility interval, simultaneously supports normality and abnormality.[*]

[*] Of course the initial value of 0.95 also has a plausible range of values. However one can incorporate this plausible range by identifying the plausible interval for the difference between the baseline and follow-up

This is what is meant by duality – a single estimate can reflect the possibility of benefit (or in this example, no harm) and the possibility of harm.

As another example, consider the current debate about the role of peanut oral immunotherapy. A response to this allergy is the use of peanut immunotherapy where the subject with the allergy is gradually given an increasing dose of peanut paste over a course of weeks to decrease their immunosensitivity to the legume.

A recent meta-analysis* of clinical trials that each examined the role of immunotherapy demonstrated that subjects in the immunotherapy treatment arms had a greater rate of surviving an oral challenge upon concluding treatment than the control group (relative risk 12.42 [95% confidence interval 6.82-22.61] At first blush this appeared to be a success; however, the authors also noted that patients in the immunotherapy groups of these trials had a greater incidence of anaphylaxis, anaphylaxis frequency, and epinephrine use. How could both findings be true?

An examination of the description of the results by the authors and also a commentary [1] revealed the answer. Many individuals in the treatment group were able to complete the exposure program successfully. However, in that same group, individuals fluctuated in their reaction to the peanut paste, sometimes reacting to a dose to which they evinced no allergic reaction previously. The responses to the therapy were not just variable but were complex,

---

creatinine levels without loss of generality. This is not carried out in this example to simplify the presentation.

* Meta analyses that combine studies not designed to be combined, however mathematically elegant, can be briarpits when it comes to interpretation. The purpose of referencing one here is simply to provide an example of duality, not to provide a dialect for or against this methodologic approach.

demonstrating a pathophysiologic intricacy that undermined the contribution of the standard statistical estimator to a helpful understanding of the exposure's effect.

The same therapy produces harm in some individuals and benefit in others. This is the essence of duality. In duality, the estimate of effect reflects a range of values, some consistent with benefit, others consistent with harm.

Plausible versus confidence intervals.

Many readers will recognize the similarity between a plausible interval and a confidence interval, Both are intervals around a statistical estimator (e.g., a sample mean difference) that reflects variability of that estimate. However, this is really the only similarity.

Confidence intervals were developed to reflect the sample to sample variability of the estimator. That is the only variability that they were designed to capture. Plausible intervals capture that variability, but in addition, also commandeer other sources of uncertainty. For example, consider the technician-to-technician variability in the assessment of an MR image. This is not sampling variability; it is the imprecision in the use of the measurement tool itself.*

Secondly, plausible intervals have no formal estimate of confidence. They are not 90% plausible, or 95% plausible. That has no meaning for us here. It is simply a region of values that are believed to be credible based on the

* Precision is the ability of different measures on the same subject at the same time to be as close to each other as possible, Variability is the difference in the measurement across different subjects in different samples.

inaccuracy (imprecision and sample to sample variability) of the estimator, and any bias introduced by the research design and execution. Plausible intervals are in general wider than 95% confidence intervals.

Finally, plausibility regions need not be symmetric.

This book's methodology: parse, channel, accumulate

This book develops three processes and compares the result.

The first process is to parse the plausible interval, into one interval that suggests benefit and another suggesting a harmful effect.

Next, we will channel that benefit interval into and through a benefit function. We will repeat this process for every analysis that the investigators believe is responsive to the question "are patients in the intervention better off than in the control group?"

Finally, we will accumulate these unitless measures through integration. We then carry out the same process for all of the plausible intervals of harm, and then compare the two.

Set theory and quanta analysis

However, this attempted accumulation of estimates of benefit and estimates of harm raises several critical questions.

The first is that many of the analyses in a clinical trial use the same observations and the same variables; they are redundant. Shouldn't subsequent analyses, using many of the same observations and variables as previous analyses, be discounted? After all, those observations and variables (in the guise of other estimators) were already used.

Each analysis is based on a combination of subjects and variables. We need to keep track of not just the raw number of them, but their actual identities in order for us to follow the redundancy.[*]

We call this data, this collection of observations and variables used for an analysis, that analysis' "region of the analysis". We need to compute the size of this region and track its overlap with the region of other analyses.

Set and measure theory permit us to measure the size of this region. We will call this size its ψ – measure (pronounced "psi measure"),

Different analyses will have different regions of analyses (since they use different collections of observations and variables) and therefore different ψ – measures.

When the regions are disparate (that is the collection of analyses use entirely different subjects and variables from each other), their ψ – measures add. However, when these regions have overlapping subjects, the ψ – measure has to be computed differently.

Measure theory suggests that analyses be broken into analysis fragments or quanta, which reflect contributions to the overall ψ – measure that are independent from other quanta. We use set and measure theory to compute this accumulation of ψ – measure over different but intersecting regions of analysis.

[**] Subjects must of course be de-identified to meet with HIPAA rules. Identification here means simply their study ID number/acronym.

This accumulation is the total content (denoted as $\mathbf{\Gamma_q}$) of the analyses used to address question q. We simply write this as $\mathbf{\Gamma_q} = \int_{A_q} d\psi$. (Figure 1).

(Figure 1)

Those readers who do not have strong backgrounds in mathematics should not be frightened away from this notation. Figure 1's integral is nothing more than an announcement of intent. It states that we intend to accumulate all of the analysis content over regions of A_q whose analyses were conducted to answer question q. These regions have their content assessed using $\psi-$ measure.*

What is unique for us in the clinical trial arena is that here we are integrating over not just part of the real line (like in a first calculus course), but over a set of analyses,. This concept is not novel in mathematics, but it typically is not applied to clinical research. This type of integral is special in measure theory going by the moniker Lebesque-Stieltjes†.

We can now accumulate the benefit and harm functions from duality theory and accumulate them with respect to $\psi-$ measure to obtain ensemble measures of benefit and harm. This is what duality-quanta analysis attempts to accomplish.

---

* Note that this integral is not our classic one, where we integrate over a region of the real line (e.g., the area under the Gaussian or normal curve), and use familiar assessments of area such as *dx* or *dz*.

† Pronounced LeBĀk-StillJes

If we were going to say this mathematically, we might begin by identifying a collection of analyses $\{\omega_1, \omega_2, \omega_3, ..., \omega_n\}$, which are to be analyzed sequentially. For each of these analyses, say, the i^{th} analysis, we compute the plausible interval for benefit, $\chi_i^{(b)}$, and apply the benefit function to it, $\mathbf{Y}_b\left(\chi_i^{(b)}\right)$. We then accumulate or measure the benefit function over each of the *n* analyses using its contribution $\psi(\omega)$. In mathematics we would describe this as accumulating the benefit function over all regions of analyses that are included with respect to the ψ – measure, writing it as $\int_{A_q} \mathbf{Y}_b\left(\chi_i^{(b)}\right) d\psi$.

We do the same thing for harm, $\int_{A_q} \mathbf{Y}_h\left(\chi_i^{(h)}\right) d\psi$ and then take their ratio. The construction of this process and managing its complexities and implications is the main topic of this book.

So we have two concepts to balance. The first is the regions of analyses that we must mathematically dissert and manage. The second is the parsing of the plausible intervals into those portions reflecting benefit, collecting them, "measuring them", and accumulating them over the analysis region quanta, and then repeating this process for harm. Finally, we take the ratio of the two.

A second issue involves the relationships between variables. This issue of correlation is easily addressed once we have developed the notation for this quantum approach and is addressed in Chapter 22.

But, before we dive into those details, let's first address the question "Why is this development even necessary?"

References

1. Abbasi J. JAMA. Weighing the Risks and Rewards of Peanut Oral Immunotherapy. 2019 Jul 31. doi: 10.1001/jama.2019.9142. [Epub ahead of print] No abstract available. PMID: 31365041

Statistical Theology

Choosing to walk away from an established guide like the *p*-value – even though that guide is now quite blind – is difficult. *P*-values served well as part of an organizing framework in the 1950's, bringing structure to inchoate clinical investigative protocols and disorderly research findings. Like training wheels on a bike, they helped keep the young clinical trial enterprise upright.

However, we have been riding for seventy years now, and these structures that kept the rudimentary clinical trial infrastructure in place are now too constraining.

Clinical research has been and remains the best hope for the solution to chronic disease, whether that hope reside in genetics, preventive maneuvers, pharmacologic therapy, or biologics. The complexity of clinical trial programs, with multiple treatment arms, interim analyses, assessments of clinical findings of mortality and morbidity, as well as examinations of promising proteomics cannot be brought to bear with their full power and authority when forced to abide by a restrictive *p*-value predominance.

Specifically *p*-values and their attendant statistical hypothesis testing do not permit the full deployment of results produced by the research enterprise. Any tool that requires rigid allegiance even though it itself is ambiguous, and defies a clear definition upon which clinical

investigators, epidemiologists, and biostatisticians can agree has more of the feel of theology than of science.

Those of you who have heard all of the arguments about the problems with *p*-values are encouraged to skip on to Chapter 3 . There will be nothing new for you here. However, others of you, who have accepted without question that *p*-values were important, useful and necessary (perhaps because you were told that they were so) may be illuminated by the following dialectic.

The two minute problem

Much of the world's population does not like mathematics. Finding problems in mathematics, uninteresting, irrelevant and a waste of time, most people are all too happy to turn over "the math" to someone else, whether that math be income taxes, working through some simple geometry for determining how many gallons of paint are required to double coat a wall, determining how long it takes to get to a destination at their current speed, or figuring out a tip for their restaurant server.

However when they are pressed to solve a math problem, the math problem falls into two different categories. The first category is the class of math problem that they can solve instantly. For example, city and state taxes combined add 10% to the initial cost of the car. The initial cost of the car is $26,000. Then the additional tax is $2,600.

Easy as pie.

All other problems fall into a second category; the timeless set. For these problems, it doesn't matter whether the individual is given two seconds, two minutes, or two years to work out the answer. They don't know how to approach the problem, much less solve it.

A major difference between these people and the mathematician is that the mathematician has tools that she can use to help to convert the two year problem into a two minute problem.

One of these tools is simplification.

In Polya's great text of mathematical guidance "How to Solve It",[1] an important tool available to mathematicians faced with a problem to solve is simply – don't.

Instead, solve a related problem.

Commonly that related problem can be a simpler problem. The initial mathematical problem that confronts us is complicated. Maybe it is finding the volume of water in a pool that is not rectangular, but instead has different sculptured shapes. Then a first approach is to assume away the complications, turn the pool into a simple circle or ellipse. Solve the simple problem then work back to the more complex, original one .

Or sometimes, the simplification is enough.

The cutting room floor

The application of statistical hypothesis testing in health care is the process by which a complex health related question has been trimmed, reduced, and distilled until it produces a simple question that can be addressed by statistical hypothesis testing.

The result is a simple assessment of one or a small number of clinical outcomes, leaving much of the research data and results in their richness and complexity behind "on the cutting room floor".

While I don't think that this was the intent of biostatisticians, or the senior clinical research leaders of the 1950's, it most certainly was not the intent of physician-scientists, who in fact collected a wealth of data in order to harvest its findings. However, the collision between the

bountiful products of clinical research on the one hand, and the need by administrators to evaporate this product down to a fine, alpha error- managed distillate has produced a product that seventy years later, has us scratching our heads.

This chapter discusses how we got here. Much of the following is taken from Lee Kennedy-Shaffer's fine article "When the Alpha and the Omega: *p*-values, "Substantial Evidence," and the 0.05 standard at the FDA" [2], as well as from [3] and [4]. There is no villain in this story. We are all complicit.

Two trajectories

From our 2020 perspective, we view clinical research and biostatistics as intricately intertwined. A physician-scientist would not consider conducting publishable health care research without at least contacting a biostatistician, and biostatistics for its part, has devoted itself to new methodologies that are commonly related to applications in health care research.

However this has only been the case since the early 1950's. Prior to that, health care research and epidemiology on the one hand, and probability and statistics on the other hand, in the main followed very different paths. I and others have discussed this topic before. The goal here is not to recapitulate in detail, but to provide the vibrancy and energy that each had developed in order to understand their ultimate calamitous collision, a detonation that has produced our current state of affairs.

This collision was not about the *p*-value. Like the city of Gettysburg, the *p*-value just happened to be in the wrong place at the wrong time. The battle was – and continues to

be − over which perspective – statistical or clinical – governs the conclusions of health care research.

Epidemiology and health care delivery

Epidemiology represents a disciplined, cerebral approach to drawing health care conclusions from facts that must be discerned within a fog of variability.

For thousands of years now, these reasoning men and women focused on applying what they observed and deduced to the sick in their care. They suffered from crude instruments, absent labs, and no way to collaborate between villages, towns, and cities except through what we might describe as a verbal, then later, written crude case report.

Celsus stated that "Careful men noted what generally answered the better, and then began the same for their patients". (circa A.D. 25). [5] For the next 1900 years, advances in clinical medicine occurred through the combined use of careful observations, clear recorded descriptions, and deductive reasoning. Chance observation tested and sometimes overturned standard dogma, e.g., the belief that musket wounds must be permitted to fester to heal. [6]

John Graunt's established the application of deductive reasoning to multiple data points [7], and along with William Petty developed the life table methodology, permitting for the first time, the computation of number of deaths from bubonic plague, consumption and "phthisis" (tuberculosis) could be quantified and followed over time. James Johnson [9] pointed out the value of literature review, the role of confounders, using replicates of treatment to address result variability and study replication. Applying these principals, James Lind on the HMS Salisbury in 1747 worked to defeat of scurvy on the high

seas, and later John Snow discerned the cause of cholera in London (1830-1850).

Principles of causation were elaborated by Sir Austin Bradford Hill, the father of clinical trials [8], providing tenets that were based on a common sense approach to determining causality and are remarkably free from complicated mathematical arguments.

The critical point here is that much of this work was not mathematics-centric. For over nineteen centuries there was little quantitative development in which they could rely, so these thoughtful men and women developed solid, intelligent but essentially, non-mathematical contributions. They would use data when it was available (e.g., the work of Graunt and Petty), but they had developed skills to operate in its absence. This continues with for example, establishing the link between tick bites and Lyme disease.

Probability, statistics, and prediction

Probability and statistics developed from a different foundation whose cornerstone (the random process) was condemned as a capital offense by religious leaders during the Middle Ages.* However, decade by decade, and

* It is easy, here in the 21st century, to be critical of religious dicta, but in this case, they served as a protective, albeit extreme reaction to a lawless and demented time that we know as the Dark Ages. The destruction of the Roman Empire innagurated an unparalleled error of depravity. With Rome and other cities demolished, the only choice individuals had was to survive in the countryside where people were never safe and fledging crops planted by outcast city dwellers most always failed. Life for these unfortunates came down to a decision; either join the rampaging gangs, or join the church. Many flocked to the monasteries, not to be devout but to simply survive.

Inside these protective walls, monks and nuns in turn outlawed gambling and its random event foundation for two reasons. First,

century by century, life softened outside the monastery walls. Villages, towns, and then cities flourished again. The advent of the Renaissance heralded the notion of free thought, and the Industrial Revolution introduced the concept of leisure time.* With leisure time there was a new (now legal) interest in what became the raging national past time – gambling.

Gifted observers began to use the data from these activities. A major advance was produced in the early 1600's by Abraham de Moivre, who developed the theory of the normal distribution as an approximation to the binomial distribution. Fermat also wrote extensively on gambling, which he correctly perceived as a process by which the future behavior (of the game) is predicted from past experience. This was the beginning of modern probability thought.

However the field did not escape criticism. Early tabulators involved in using techniques such as sampling in census counts were said to not be involved in science but in "political arithmetic" , defined as "the art of reasoning by figures upon things related to government.†

As the work of Laplace, Poisson, and others moved probability forward, and delved into the implications of the

---

gambling was commonly conducted for financial gain, a worldly concern more fitting for those who chose lives outside the insulating walls of the monasteries and convents. Secondly, the very concept of randomness suggested that events occurred outside the contol of God. While gambling was simply a crime punished by expulsion. contemplating randomness was blasphemy.

* Up until the offloading of manual work to machines, a person's day was consumed principally by work, time at the Church, or eating/sleeping.

† From Charles D'Avenant, taken from Karl Peasons *The History of Statistics in the 17th and 18th Century*.

work of Thomas Bayes, an interesting debate arose about the contribution of the field to society. Should the quantitative scientists be the best ones to interpret the data that they analyze, or should that be turned over to another.

In 1834, when the Statistical Society of London (later to become the Royal Statistical Society) was formed, they lent their perspective to the debate through their selection of an emblem; a fat, neatly bound sheaf of healthy wheat that represented the abundant data, neatly collected and tabulated. On the binding ribbon was the Society's motto *Aliis exterendum*, which means "Let others thrash it out" [9]. As this sense of the field took hold, Pearson and Gossett pushed the work of what to do with aggregate data up to the brink of statistical inference to the twentieth century.

Eruption

The two fields of epidemiology and statistics did more than peacefully coexists, they in fact worked jointly on major issues, such as demonstrating that early vaccines for smallpox were effective,*

However, a tempest was coming for them all in the 20th century, and it was the clinical/observational scientists and epidemiologists who were first blown off of their progress-trajectory of progress. Yet, this storm's creator was not Ronald Fisher and his notions of significance testing in the 1920's.

---

* One of the first examples of comparing observed results to what was expected. Here, the Bernoulli brothers, using the binomial probability distribution, computed the expected number of cases, and the epidemiologists and health care providers counted the observed, permitting a comparison between the two.

It was Albert Einstein.

While not trained in biology, medicine, or the observational work of epidemiologists, Einstein created the intellectual seismic disturbance that shoved the observationalists' ships perilously close to the rocks.

In deriving the famous equation $E = mc^2$ Einstein developed the principal that positions of reference were relative. Observers from different platforms could observe the same result and come to different conclusions –and both were right. Geologists, biologists, chemists, clinicians, and epidemiologists had struggled over the centuries to develop experimental paradigms that would provide the best and unbiased platform from which to observe a result and therefore draw a correct conclusion.

Einstein told them that this was impossible. By telling observational scientists that what they observed may not be as close to the truth as their training suggested, Einstein, unwittingly, but firmly invalidated them.

Some chose to challenge Einstein.

The book "One Hundred Authors Against Einstein" [10] was an attempt by non-physicists to resist the great physicist's ideas of time dilation, declaring them irrelevant to biologic processes. It countered that little could be learned of the real world by abstract mathematics which itself had lost its connection to "common sense".

Yet the proof of the General Theory of Relativity, removed most major scientific criticism of Einstein's work. Einstein's second compendium on gravity demonstrated to observational scientists not only that they could not trust their instruments, but that mathematics was more trustworthy. In demonstrating that an observation (the bending of light by the sun) had been missed through thousands of years of sol observations, but had been

predicted by mathematics, was not just a *tour de force* of physics, but an insufferable blow to those who believed in the power of observation.

The only encouragement the observationalists received from the new masters of physics was to not trust their eyes, but to instead rest their faith on mathematics. And this mathematics was new, dense, and to them, impenetrable. How were they supposed to, for example, listen to hearts, or interpret chemistry tests at relativistic speeds, and how could that possibly help them in their work or practice?

They were dead in the water.

Meanwhile statisticians, who themselves were struggling with the concept of how to use data to convincingly answer questions, received an unanticipated new wind in their sails from Ronald Fisher.

The 1920's and the appearance of statistical significance

One of Ronald Fisher's earliest writing on the general strategy in field experimentation was his 1925 book *Statistical Methods for Research Workers* [11], and in a short 1926 paper entitled "The arrangement of field experiments" [12]. This work contained Fisher's thoughts on experimental design, and his initial framework for significance testing.

It is also where the first mention of a five percent level of significance first appeared.

Using as an example, the assessment of manure's influence on crop yield, he puzzled over how to compare the yields of two neighboring acres of land, one treated with manure and the other not. It was true that the manure-treated plot produced a 10% greater crop yield than that of the non-treated plot, yet Fisher knew that there was also variability due to other factors (e.g. soil moisture, insect

density, difference in seed quality, etc.). Fisher distilled the question down to an assessment of how likely would one expect to see a 10% increase in crop yield in the absence of the manure by chance alone. He then reasoned:

"...the evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the level at which we can say 'Either there is something in the treatment or a coincidence has occurred such as does not occur more than once in twenty trials.' This level, which we may call the 5 per cent level point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials." [12].

He added

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) or one in a hundred (the 1 per cent point). Personally, the writer prefers to set the low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level) " .[12]

Fisher continued to say that if he had the actual yields from earlier years, and could compute the variability of the yields, then he might use Student's *t*-tables to compute the 5% significance level.

The significance level of 0.05 was born from these rather casual remarks [13].

Observational scientists react

Many statisticians were elated by Fisher's writings. The notion of sample-to-sample variability had been known to

investigators for years. Early 20th century Bayesian and non-Bayesian statisticians alike wrestled with how to combine the two in order to assess the value of a measured experimental effect size. Finally, here was a way to deal with the chronic problem of sample-based research using objective mathematics

Yet, Fisher's point of view on experimental design became the flash point of a new controversy. Many observationalists – believers in the scientific method, welcomed an approach to test a scientific hypothesis. Yet, to them, the significance testing scenario was counterintuitive, representing the unhelpful type of thinking that was likely to be produced by mathematical workers who did not spend sufficient time in the observationalists' world of data collection and deductive reasoning.

To these observationalists, the entire process of statistical hypothesis testing was reversed. The scientific method began with the notion of a hypothesis that the scientist believed, e.g., a new compounded powder will reduce fever in post-partum women. This was then to be supported or disproven by the collected data.

However, statistical hypothesis testing began with the reverse perspective, that the compound would not be effective. It then set up the assessment so that the collected data would either prove or disprove (i.e., nullify) the null hypothesis. This not only was ungainly and complicated, but it was indirect. To traditional observationists, Fisher's significance testing appeared to be just the type of indecipherable, mathematical, reverse logic that had already shaken the foundations of early twentieth-century epidemiology.*

* Statistical hypothesis testing had very much the look and feel of proving that the $\sqrt{2}$ was irrational. One did not actually show that this

Specifically, this new, upside-down paradigm of statistical significance appeared to deny the scientist the ability to prove the hypothesis he believed was correct. Instead, the scientist would be required to replace the strong assertion of his own affirmative scientific hypothesis with the tepid alternative of disproving a hypothesis that he did not believe.

Already bruised by the two decade-old assault on its philosophical opinions by physicists and mathematical theorists*, they gathered their forces. From the epidemiologists' point of view, it was bad enough that they

---

quantity was an irrational number. Instead, one assumed that it was rational and then reasoned to a contradiction. Note that there is no special feature of an irrational number that is revealed in this proof, only that it is not rational.

* Over time, epidemiologists have successfully defended their time-tested methodologic perspective. Of course, the flaw in all of the criticisms regarding the use of observation as a foundation method of epidemiology lies in the difficulty in translating findings that are germane in one field (physics) to that of another (life sciences). While the findings of the relativity laws are in general true, they are most useful in physics. The theoretical physicist may be correct in asserting that every observer is biased and that there is no absolute truth about the nature and magnitude of the risk factor–disease relationship. However, this does not imply that all platforms are equally biased. Epidemiologists never stopped striving to find the most objective position possible. Certainly, if bias cannot be removed it should be minimized. The fact that bias may not be excluded completely does not excuse its unnecessary inclusion.

Second, while mathematicians are capable of predicting results in physics, they have not been able to predict disease in any important or useful fashion. No mathematical models warned obstetricians or their pregnant patients of the impending thalidomide–birth defect link. Similarly, mathematical models did not predict the birth defects that mercury poisoning produced in Japan. While physics often studies processes in which mathematics can reign supreme, real life and its disease processes have proven to be painful, messy, and chaotic affairs. The substantial role of epidemiology is incontrovertible in the development of the most important new healthcare research tool of the twentieth century—the clinical trial. The time-tested tools of epidemiology continue to prove their utility in the present day.

had to sit still where Einstein's advocates criticized their world of observational scientists. However, they did not have to take this from an unknown agrarian statistician.

The field responded vehemently and vituperatively as in

"What used to be called judgment is now called prejudice, and what used to be called prejudice is now called a null hypothesis … it is dangerous nonsense ... "[14]

In the meantime, enthusiasts in the statistics community for the notion of significance testing grew. In the 1930's, Egon Pearson and Jerzy Neyman developed the formal theory of testing statistical hypotheses even further. In addition, they introduced the notion of statistical power [15] while other workers produced the concept of the confidence interval. [16, 17, 18] These developments took place on the searing bedrock of controversy, fueled by the vitriolic criticisms of Berkson [19,20] and vibrant ripostes by Fisher. [21]

Enter the administrators

During these debates, few noticed the explosion in health care research. World War II, with its requirement for new, improved medicine and delivery of healthcare services to both soldiers and refugee populations generated explosive new waves of healthcare research.

Medical groups in the UK and the US developed and tested new medications e.g., antibiotics, and oral hypoglycemic agents. The Medical Research Council (MRC) was conducting one of the first clinical trials to evaluate streptomycin. New investigators were pouring into

the labs of pharmaceutical companies or the facilities of universities.

Few knew anything about the raging controversy concerning significance testing. However, new, statistics-centered thought was beginning its migration to clinical research.

W. Edwards Deming, in his 1943 book *Statistical Adjustment of Data,* suggested using *p*-values in repeated experiments as measures of the quantum of evidence against the null hypothesis. Specifically, he recommended the use of "statistical significance" as an inferential method. However, he also added the monitory that "[s]tatistical 'significance' by itself is not a rational basis for action. [22]

This warning was not headed by leading journals.

A 1950 editorial in the *Journal of the American Medical Association* (JAMA) under the banner question "Are Statistics Necessary?" offered a wholehearted answer; "Yes". According to the journal, investigators developing new therapy must be able to 1) assemble a table comparing treated and control subjects, and 2) be proficient in computing *p*-values. [23]

Another JAMA article required statistical tests to be the basis of evidence for therapeutics, urging clinical studies to use randomization, untreated controls, and significance testing. [24] The *Annals of the New York Academy of Sciences* similarly called for quantification of clinical trial results and the application of statistical reasoning [25].

Reaction to this building momentum for statistical significance was pointed, with much of it negative. In 1960, William Rozeboom wrote of his concern for the use of statistical hypothesis testing. Not only did he resist the notion of accepting or rejecting a scientific hypothesis

completely, he also resisted the notion of the 0.05 significance level, recognizing that there was no scientific underpinning for the 0.05 level [26]

Presciently, in a speech delivered to the International Biometric Society in 1969, the outgoing British Regional President J.G. Skellam warned that significance tests might "exercise their own unintentional brand of tyranny over other ways of thinking."[27]

However, these iconoclastic perspectives were overwhelmed by the medical research literature, which was in a state of tortured turmoil over how to manage data assessments. While there was a time tested format for reporting case reports (journals had, after all, been publishing these for generations) there was no standard for reporting – much less analyzing – a dataset.

All agreed that having more than one observation improved upon a simple case report, it remained an open question as to how to report this data? What are the measures of central tendency? How should dispersion be reported and managed? And how does one compare different experiences when those experiences are segregated by treatment group and control group? These questions plagued journal editors, who were anxious to find an organized and controlling structure.

However, such a move played to the journals' best practical interests as well. At the time, the number of manuscript submissions were swelling with no crest in site. The introduction of a significance testing requisite would function as a constraining factor, modulating the number of articles worthy of submission at a time when journal editors were overwhelmed.

"Rule of thumb"

The US Food and Drug Administration (FDA) now entered this tempestuous arena as they themselves struggled with the absence of analytic rigor in new drug applications, the number of which exploded during the post war era.

From the 1950s into the 1960's, the state of protocol submission to the Agency was abysmal. Many drug applications were submitted to the FDA, not just without a statistical plan, but without a protocol [28]. The Agency soon joined the chorus of voices calling for statistical rigor [29].

There was no statement in the regulations that the 0.05 *p*-value metric had to be used, but there was a understanding that the customary "rule of thumb" in assessing the effect of an experiment would be the Fisherian 0.05 standard.

Thus, by the mid-20th century, overwhelmed journal editors as well as overworked FDA employees received a welcome architecture to both provide research supporting structure and also serve as a bulwark and filter to against for the applications they were receiving. Sometimes tacit and sometimes explicit, there was an understanding that statistical hypothesis testing would be required for drug application submissions at the FDA and for manuscript submissions to journals.

From a practicality perspective one cannot blame these administrators for seizing on statistical hypothesis testing. The field was overwhelmed with poorly designed research and shoddy analysis plans. Introduction of a new methodology promised much needed injection of rigor.

The *p*-value appeared to be the natural solution. After all, it incorporated the size of the sample, the effect size

that was observed in the study as well as its variability, plus an attempt to assess the "generalizability" of the results. And of course it was simple to understand. Values < 0.05 were considered acceptable. Those > 0.05 were not worthy of further consideration. It was certainly compact, and believed to be quickly interpretable.

A more thorough discussion of this complex choice is available [30,31,32,33,34]. The instillation of the *p*-value as a research result metric was not one of malevolence. Instead the decision to move forward was made by observant and overwhelmed administrators who simply wished to practically underwrite and promulgate the most solid research efforts. They hoped that the use of this tool would permit the data to speak for themselves in a structured fashion, free of the bias that an investigator would bring to the research paradigm.

Nevertheless, this decisions by these influential gatekeepers of healthcare research had a profound influence on not just the structure of research reports, but on the climate of research itself.

Almost immediately, there were problems.

In the 1960s, as FDA was incorporating significance testing into its new drug, epidemiology solidified its antithetical approach to the role of hypothesis testing in medicine through the elaboration by Sir Austin Bradford Hill first in 1953[35], then again in 1965 of the well-established epidemiologic tenets of causality. These tenets served as the basis of epidemiologic causal thinking in the mid twentieth century.*

* Free of complicated mathematics, these hallmarks of a causal relationship have twin bases in common sense and disciplined observation. The nine precise Bradford Hill criteria are: (1) strength of

However, Hill's thoughtful, accepted approach was beginning to be supplanted by the following style of reasoning:

> *"Since the study found a statistically significant relative risk ... the causal relationship was considered established [36]."*

While this type of comment was not typical, it did demonstrate the extreme conclusions that were beginning to be based solely on the *p*-value.

The answer to the central question "What is the role of mathematics in drawing conclusions from health care research?" now seemed to be that mathematics was going to play the predominant role, with little need for additional thought. What had been offered by Fisher, Neyman, and Pearson, as an objective sense of the strength of evidence of research result was now being transmogrified into a popular but inadequate substitute for clear, causal thinking, as workers replaced their own, careful, critical review of a research effort with the *p*-value. Was there to be no further role for assessing clinical significance in the absence of statistical significance?*

---

association, (2) temporality, (3) dose-response relationship, (4) biologic plausibility, (5) consistency, (6) coherency, (7) specificity, (8) experimentation, and (9) analogy. These are well elaborated in the literature

* One of my medical school interviews took a disasterous turn for the worse when the interviewer, during his examination of me, received a letter of rejection of his submitted article from a journal because the results did not reach statistical significance. The investigator – interviewer raged at the inappropriate use of statistics which was, after all, meant to describe and not to decide research matters. By the time

Also, more nefariously, scientists began to sculpt and therefore promulgate their findings based on analyses in which the *p*-value was small. *

Pushback

It was inevitable that some workers in health care, as their forbearers had fifty years earlier, vigorously resist this degradation in the scientific thought process. The dispute broke out into the open in 1987, when the prestigious and well-respected *American Journal of Public Health* solicited an editorial arguing that significance testing be purged from articles submitted for review and publication.

Subsequently, the epidemiologist Alexander Walker debated with the statistician T.W. Fleiss over the use of significance testing [37,38,39,40,41], with Fleiss, which was joined by Poole.[42] In addition, Rothman, in an editorial for Annals of Internal Medicine in 1986, wrote that "[t]esting for statistical significance continues today not on its merits as a methodological tool but on the momentum of tradition."204[43]

Doubling down on the *p*-value

In the meantime, cardiology began to get a taste of what was coming from its forced feeding of *p*-value laden meals.

- The suggestion by the Multiple Risk Factor Intervention Trial (MRFIT) that hypertensive men with baseline ECG abnormalities were

---

the interviewer soothed himself, the interview time ended and I was sent on my way.

* This now goes by the sobriquet "*p*-hacking"

harmed and not helped by antihypertensive therapy was a stunning blow to the hypertensive control community, who were just as stunned to later realize that this clinical trial based result was false, based on a *p*-value, was not reproducible. [44]

- The clinical trial based finding that the anti-acetyl cholinesterase therapy vesnarinone could save the lives of patients with heart failure was reversed by subsequent clinical trials that demonstrated the harm of this therapy [45,46].
- A clinical trial demonstrating the mortality benefit of Losartan had its result overturned by a subsequent clinical trial [47,48].
- A clinical trial based subgroup analysis that declared amlodipine could prolong the life of patients with non-ischemic cardiomyopathy was also reversed by a subsequent clinical trial [49,50].
- The diminished and confusing efficacy findings from United Kingdom Prospective Diabetes Trial (UKPDS) [51,52,53,54,55].
- The violation of prospectively declared analysis procedures in the Lipid Research Clinics (LRC) trial [56,57].
- The US Carvedilol program controversy [58, 59, 60, 61, 62, 63, 64] of the late 1990's.*

This collection of results served to undermine the confidence of cardiologists in the interpretation of clinical trial results. Mina Antrim could have been speaking to

* The author played a role in this controversy.

cardiologists when she said in 1901 "Experience is a good teacher, but she sends in terrific bills…*"* [65].

At this critical juncture, clinical trials in cardiology were becoming more complex, with multiple prospectively declared endpoints, multiple treatment arms, subgroup evaluations. analyses, and the very early examination of proteomics results. The diseases that were studied were complicated and investigators – following their natures – wanted to be sure that they captured as much of the dimension of the offered as they could.

If the *p*-value wasn't serving its anticipated role, than other quantitative research tools were required.

The response of the biostatistical community was to "double-down" on the *p*-value, producing the following set of research principals.

The first principal, was the elevation of the protocol as a rulebook of the research. The second principal was that endpoint assessments must be planned with clear and declared assessments of type I error penalties.

The first was a re-enunciation of the established need in the 1950's to have a protocol in place. It formalized the research thinking, stating the investigators' belief in the effect and mechanism of action of the intervention. It also stated in practical details the needs of the research, permitting the investigators to identify equipment, expert committees and computing facilities necessary to conduct research efforts with precision and accuracy.

However the re-enforcement of type I error penalties would have profound implications for clinical research.

Clinical research complexities

In my view, the re-anchoring of *p*-values as a requirement in the dynamic clinical research environment with its new

complexities was a mistake. The effects of their introduction in the 1950's were conflated with the requirement of a structured protocol, a requisite that paid handsome research dividends.

The *p*-value did not add to the stronger research foundation, already protocol fortified. It instead added a confusion metric which has only served to distract investigators from the need for clear causal thinking when studying complex diseases.

However, by the beginning of the 21st century, its effects were to be stultifying.

In clinical research, type I error would now operated in a tightly controlled type I error environment. Specifically, only outcomes that had type I error allocated would be considered as primary* as a response to a situation in which nonprotocol–tethered research actions were producing clinical trials purportedly answering the same questions but with disparate results. By doubling down on the *p*-value, biostatisticians and others were requiring that the investigators focus on a small number of outcomes.

Now, in order to conserve type I error, the number of multiple outcomes that would be considered as a primary would be small; most clinical trials have one primary outcome with relatively fewer having 2-3 primary outcomes. No matter how many outcomes were prospectively declared, the only outcomes that would "matter" were those that had an alpha level declared prospectively.

Since the statistical power decreases for smaller alpha levels *ceteris paribus*, there could not be very many of these outcomes because the overall (or family wise) type I

* A primary outcome is one on which the trial's effect of therapy is classified as positive, null, or harm

error had to be 0.05. Thus investigators were required to select the outcomes that they thought would be positive result in order for the trial to be judged positive.

For some trials, it is clear what the outcome should be. There may be an expectation in the research and clinical community, or the regulators may have outlined what outcome they wish to see.

However, this is the minority of research programs. Most research programs have an idea how the therapy may work, but had no reliable guiding pre-assessment as to what outcome is going to positive in one particular sample of patients. In one sample, it may be left ventricular ejection fraction. In other sample, left ventricular end diastolic volume may be most influenced by therapy. Each has known variability, and each is related to similar clinical consequences, yet the investigator cannot know which of these outcomes will be influenced by the therapy in this sample. Yet, statistical hypothesis testing requires that they choose one, or choose both and pay a severe type I error penalty[*]

Physician-investigators understand that there is a wide breath of knowledge to be gained by analyzing the entire dataset. Insisting on parsimony (that is, focusing on and analyzing only the small number of outcomes that have type I error allocated for them prospectively), means that much of the data will essentially go un-analyzed. [†]

---

[*] They could choose both, but have to divide the type I error between the two, leading to an important increase in sample size.

[†] In 2016, an NIH admistrator said to me that if it were up to him, the only clinical trial outcomes that would be funded would be those that were an affirmative part of the type I error allocation. While this is not representative of NIH policy, it does reveal how penetrating the type I error mentality can be.

Investigators want to cover new ground, and enjoy the exploration process. Exploratory analyses can evaluate the effect of the therapy in subgroups, the effect of the therapy on different endpoints, and the effect of different doses of the medication.

The rationale for identifying a large number of outcomes in clinical research has its own undeniable force of logic (stemming from logistical/financial, epidemiologic, and the need to examine data in new and provocative ways exploratory analysis).

Much of this was cut off at the knees by the need to control the overall type I error. Specifically, the community, rather than rethink the role of the *p*-value, decided to constrain research to fit into its narrow interpretative environment.* By the 21st century, the limitations of statistical hypothesis testing were clear. Complex clinical research endeavors were expanding beyond the *p*-value's ability to guide helpful interpretations.

Over time, these principles were absorbed by the cardiovascular community. Contemporaneous protocol review committees (PRCs), Data Safety and Monitoring Boards (DSMBs), the FDA, top tier journals, and knowledgeable audiences of international cardiology meetings now expect these conditions to be met. However, at what cost.

* I must point out that I was part of this process. Having been trained in the construction, use, and interpretation of p-values, this notion of type I error control I regretfully supported.

The results have been investigator frustration, and tragic alpha error fiascos, e.g., the MERIT-HF program.* This is the product of process in which clinical researchers permitted statistical hypothesis testing to not merely be supportive, but to dominate their efforts.

Wasteland

We had the opportunity to deemphasize alpha errors, and instead embrace the full panoply of findings in clinical trials by abandoning the *p*-value and its restrictions.

Those who called for this in the 1990's could have been better supported by the National Institute of Health with the creation of a new methodologic area for the development of new tools to replace statistical hypothesis testing.†

Bayes procedures, always a useful counterpoint to standard statistical hypothesis testing could also have received new encouragement for development.‡ Such devices when fully engaged permit research efforts to assess all of the data and relevant analyses, providing a summary conclusion.

Instead, the research community acquiesced to the continued enforcement of statistical hypothesis testing implementation, forcing many if not most of the analyses in a clinical research effort outside that effort's interpretative paradigm.

---

* In MERIT-HF, the alpha allocation and endpoint composition was changed at several points in the study, ultimately preducing a failed result.

† See the conclusions for how this could have, and might still be developed.

‡ To its credit the Medical Devices Division at the FDA has affirmatively placed an emphasis on this approach for at least the past decade.

The outcome has been dissatisfying and disappointing. Clinical trial protocols now include intense details about the order of analyses and what type I error function is implemented to control the overall alpha error level. The permitted illumination provided by secondary endpoints is reduced,* and truly new and unanticipated information is extinguished and considered not publishable in many areas. Instead, analyses are highlighted that have limited role in understanding the pathophysiology of a disease, or are the product of complex outcome combinations,† but produce executable sample sizes, again driven by statistical hypothesis testing concerns.

The practical impact of these functions is to exclude the impact of analyses that can shed light on either the breath of the findings or the biologic mechanism on the overall finding of the study by denying them the status of primary since the alpha calculus prohibits too many primary endpoints.

The clinical research mantra used to be "if the study wasn't published, then it might as not have been done". The operational mantra now is "if alpha was not allocated prospectively for the analysis, the analysis is vacated and non-admissible."

This philosophy reduces major clinical trials with well planned, multitudinous analyses to rest on the findings of a single primary endpoint when biology, physiology, and

---

* They are reduced because they are not given the same weight as primary analyses, and there is no standard way to combine primary and secondary analyses into an omnibus measure of effect.

† Combined outcomes are outcomes that have important pathophysiologic rationale in the study, but because of their low event rates cannot stand as the study's primary endpoint because the required sample size would be prohibitive. These endpoints are therefore combined into a more complex endpoint requiring a lower sample size.

pathophysiology and common sense say the findings are best interpreted in the light of multiple endpoints.

The end result in many fields e.g., cardiology is that we now operate in a wasteland of barren research effort, stripped of its epidemiologic richness, relentlessly patrolled by a ruthless *p*-value centric metric.

And it is about to get much worse.

Reproducibility

There are many concerns about the absence of reproducibility in health care research. Several examples have been provided here (*vide supra* vesnarinone, losartan, amlodipine) where the findings of one (expensive) trial were overturned by another. Examples are commonly provided about the reproducibility of research in other sciences, and many wonder what can be done to improve the reproducibility in clinical research endeavors.

In the journal *Nature Human Behavior* [66], a collection of distinguished and experienced statisticians and quantitative scientists reviewed the issue in its complexity.

Their conclusion – the health care field needs a new *p*-value threshold.

Specifically, the *p*-value threshold should be reduced from 0.05 to an 0.005 level of significance.

While some believe that this might be a helpful change when applied retroactively[67] when looking forward, the sample size, financial, and logistical consequences to be borne by the clinical investigators if not insurmountable, are considerable.

To some degree, this statistician-based initiative of reducing the *p*-value threshold to 0.005 was predictable. The initial *p*-value injection into healthcare research was supported by many of us in the 1950's. However, its use in the presence of the new research complexities of the 1980's

and 1990's (subgroups, multiple treatment arms, and multiple endpoints) generated a collection of *p*-value conundrums, yet we were not sufficiently compelled to abandon it.

Instead, we enforced its use, ensuring that type I error was conserved, to the detriment of research design. [*]

However this alpha imposition did not solve the reproducibility issue, so statisticians now suggest a further crackdown, reducing the maximum type I error from 0.05 to 0.005.

We can understand how biostatisticians would be attracted to this. We statisticians have been in the business of generating *p*-values in health care research for almost 70 years. It is a good, consistent business for us. The notion of strengthening the *p*-value can be packaged to produce fine publicity optics, and does not change the trappings of the underlying research enterprise.[†]

However, its clinical and research consequences if implemented would be profound. Clinical research efforts would increase profoundly in sample size. They would take longer to complete. Furthermore, their financial costs would skyrocket during a time of diminished financial resources.

And, of course, it would not solve the problem. The reproducibility problem would likely swell. This is because the problem with reproducibility is not the *p*-value.

This is very practical issue. There are many reasons that research efforts are not reproducible. Sampling error is only one of them, and that is what *p*-values measure and

---

[*] Rather than focus on differences in inclusion/exclusion critiera and differences in endpoint definitions.

[†] It is easy enough to change statistical programs to react to thresholds of 0.005 rather than 0.05.

manage. They do not measure whether patient populations are the same. They do not assess whether exposure to the intervention is identical across studies. They do not evaluate whether follow-up time is equivalent between studies.

In addition, different parts of the country (or the world) recruit different individuals with different phenotypes and from different cultures. Concomitant medical care may be different. Outcomes may be similar across two studies but not identical. Endpoint committees code differently. Imaging machines (whose availability is commonly based on hospital contracts and not research needs) have different precision.

The impact of these issues on research results is enormous, has nothing to do with sampling error, and is not erased by changing a *p*-value threshold from 0.05 to 0.005. This recommendation focuses on straining out the sampling error gnat while swallowing the design inconsistency camel.

However, there is also a philosophical concern. Here is a comment by Jerzy Neyman and Egon Pearson, authored by them in an earnest attempt to help others in the early 1930's understand the heart of significance testing and reproducibility.

> *"But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong."*
>
> *Neyman J, Pearson E.*

On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society, Series A. 1933;231:289-337

This is the heart of statistical significance testing. In order to understand the results of such testing, we must 1) give up knowing whether a single experimental result is right, for 2) the idea of how correct we are in assessing the long run experience of multiple research efforts.

But this is not the model of health care research. It is fine for flipping coins but not for clinical trials.

Consider the ALLHAT clinical trial. It recruited over 42,000 patients to study the impact of alternative antihypertensive agents lipid reductions. This successful but immense and complex study likely never be repeated on its large scale. Similarly for the Women's Health Initiative, which enrolled over 161,000 women.

These studies are one-of-a-kind studies. We cannot afford to, as Neyman and Pearson suggest, give up hope in knowing whether each hypothesis is true or false. We must do all that we can to learn if each of these large health care experiments reached the correct result.

Therefore, in order to be sure that we are interpreting the results fairly – that we have the greatest likelihood of an accurate interpretation – we should examine all of the experiment's germane data. They collected a substantial, sometimes overwhelming amount of data to answer the question. That data was designed and embedded to be analyzed and contributory to the research question, not to be discarded simply because it was not included in the alpha calculus of primary endpoints.

Finally, we recognize that different efforts with different designs executed with different subject populations will produce different results, but perhaps not different conclusions. In order to understand the conclusion in its entirety, one should examine all of the data.

Unfortunately, the statistical colloquium (like their colleagues 25 years ago) just focused on the *p*-value.

The metric for reproducibility is not whether clinical trials can produce the same small *p*-value for the same endpoint. It is instead whether clinical trials designed to answer the same question, recruiting from the sample population base, with the same panoply of endpoints each determined with the same precision, can produce results that demonstrate a consistent effect of therapy on the disease. Desirable clinical and epidemiologic reproducibility has little to do with the slavish statistical hypothesis testing results.

This change in the *p*-value threshold must be resolutely resisted. If it is not, then the *p*-value like a tick will burrow deeper into the tender health care research hide.

But, it will not go quietly into the night. We have to force it out.

Conclusions

This is my assessment of where we are with the use of hypothesis testing in clinical research as well as our path to its place.

Neither I nor any of the coworkers that I have been privileged to work with would argue that mathematics has no place in health care research interpretation. When used correctly it can summarize the findings of complicated research programs.

However, statistical hypothesis testing in general, and the *p*-value in particular fails this test.

By itself (and few people argue that it is useful by itself), statistical hypothesis testing cannot even summarize a simple single outcome measure experiment. It must be accompanied by the effect size, the effect size's standard error, and the confidence interval to provide the assessment of strength of association and also the variability around that strength.

We have much to be thankful for with the introduction of methodologic rigor into health care research efforts that began in the 1950's. We should stay close to these improvements and let their requirement continue to guide our clinical research efforts. Solid dependable protocols, concordantly executed are requisites for health care research.

But not the *p*-value.

Experiments now are much more complicated than in the 1950's when the "0.05 rule" was first enforced. Clinical trial programs now commonly have multiple treatment arms. They can look at dose response. They can react to a protocol mandated discontinuation of the treatment arms. They can contain outcomes assessed over multiple time points, multiple outcomes assessed at single follow-up time point. They contain proper subgroups, complex proteomics and exploratory analyses.

P-values were simply not designed for this complex environment.

However, unfortunately, rather than set them aside when the research enterprise became complex, the statistical and administrative community "doubled down' on them. The new research environment excluded subgroup analyses, secondary endpoints, dose response relationships

(and, yes, exploratory analyses) from being quantitatively included in the assessment of the study, principally because there was no way statistical hypothesis testing could manage all of this.

Rather than discard a constraining metric, they just ignored the complexity of the research program that did not lend itself to the *p*-value, relying on the part of the research program that it deemed interpretable through the type I allocation rule. This is not unlike the hungry man who starves because his weak flashlight does not reveal the feast just out of his view.

We need something better. The following is my idea.

References

1 G. Polya, "How to Solve It", 2nd ed., Princeton University Press, 1957, ISBN 0-691-08097-6.

2 . Kenney-Shafer, L. When the Alpha is the Omega: P-Values, “Substantial Evidence,” and the 0.05 Standard at FDAFood Drug Law J. 2017 ; 72(4): 595–635.

3 . Moyé LA.(2006) *Statistical Reasoning in Medicine – The P value Primer*. 2nd Edition New York. Springer.

4 . Moyé LA (2003) *Multiple Analyses in Clinical Trials: Fundamentals for Investigators* New York. Springer.

5. Bull, J.P. Historical development of clinical therapeutic trials. *Journal of Chronic Disease*.:218–248.

6. Malgaigne, LF. (1947). Weuvres Completes d’Ambrosise Paré, vol. 2. Paris, p. 127. Reported in Mettler, p 845.

7. Sutherland I. (1963) John Graunt: A tercentenary tribute. *Journal of the Royal Statistical Society* (A) **126**:537-556.
8. Hill, B. (1953) Observation and Experiment. *New England Journal of Medicine* **248**:995–1001.
9. Cochran WG. Early development of techniques in comparative experimentation. From Owen D.B. (1976).*On the Hstory of Probability and Statistics*. New York and Basal Marcel Dekker, Inc.
10 . Israel, Hans; Ruckhaber, Erich; Weinmann, Rudolf, eds. (1931). Hundert Autoren gegen Einstein. Leipzig: Voigtländer.
11. Fisher, R A. (1925) *Statistical methods for research workers.* Edinburg. Oliver and Boyd.
12. Fisher RA. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*. September 503 - 513.
13. Owen DB. (1976) On the History of Probability and Statistics. New York. Marcel-Dekker.
14. Edwards A. (1972). *Likelihood.* Cambridge, UK: Cambridge University Press.
15. Neyman J., Peason E.S. (1933) On the problem of the most efficienyt tests of statistical hypotheses. *Philosophical Transactions of the Royal Society (London) Se A.* **231**: 289-337.
16 . Pytkowsi W. (1932) The dependence of the income in small farms upon their area, the outlay and the capital invested in cows, (Polish, English summaries), Monograph no. 31 of series Bioblioteka Pulawska, publ. Agri. Res. Inst. Pulasy, Poland. Wald. A. (1950) *Statistical Decision Functions*, Wiley New York.

17. Neyman J. (1937) Outline of a theoray of statsitical estimation based on the classical theory of probabilitiy. *Philosophical Transactions of the Royal Society (London) Ser A. 236*: 333-380.
18. Neyman, J. (1938) L'estimation statistique traitée comme un problème classique de probabilité. Actual. Sceint. Instust. **739;** 25-57.
19. Berkson J. (1942) Experiences with tests of significance. A reply to R.A. Fisher. *Journal of the American Statistical Association.***37**: 242 - 246.
20. Berkson J. (1942) Tests of significance considered as evidence. *Jounal of the American Statistical Association* **37**:335 - 345.
21. Fisher RA. (1942). Response to Berkson. *Journal of the American Statistical Association* **37**:103 - 104.
22. W. Edwards Deming, WEStatistical Adjustment of Data 30 (1943)
23 Editorial, *Are Statistics Necessary?*, 143 J. AM. MED. ASS'N 1260, 1260 (1950)
24 . Otho B. Ross, Jr., Use of Controls in Medical Research, 145 J. AM. MED. ASS'N 72, 72 (1951).
25 Reid DD, *Statistics in Clinical Research*, 52 ANNALS N.Y. ACAD. SCI. 931, 933 (1950).
26 Bakan D, The Test of Significance in Psychological Research, 66 Psychol bull. 436 (1966)
27 Skellam, 1J. G Models, Inference, and Strategy, 25 Biometrics 474 (1969)].
28 Temple, R. How FDA Currently Makes Decisions on Clinical Studies, 2 Clinical Trials 276, 276 (2005).

29 Carpenter DP, Reputation and power: organizational image and pharmaceutical regulation at the FDA 269–97 (2010)

30. Goodman, S.N. (1999). Toward Evidence–Based Medical Statistics. 1: The *p*-value fallacy. *Annals of Internal Medicine* **130**:995–1004.

31 . Marks HM. The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990. Cambridge, UK: Cambridge Univ Pr; 1997.

32. Porter TM. Trust In Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton, NJ: Princeton Univ Pr; 1995.

33. Matthews JR. Quantification and the Quest for Medical Certainty. Princeton, NJ: Princeton Univ Pr; 1995.

34 Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L. The Empire of Chance. Cambridge, UK: Cambridge Univ Pr; 1989.

35. Hill B. (1953) Observation and experiment. *New England Journal of Medicine* **248**:995–1001.

36. Anonymous (1988) Evidence of casueand efect relationship in major epidemiologic study disputed by judge. *Epidemiology Monitor* **9**:1.

37. Walker AM. (1986) Significance tests represent consensus a and standard practice (Letter) *American Journal of Public Health.***76**:1033. (See also Journal erratum;**76**:1087.

38. Fleiss JL. (1986).Significance tests have a role in epidemiologic research; reactions to A.M. Walker. (Different Views) *American Journal of Public Health*;**76**:559-560.

39. Fleiss JL. (1986)Confidence intervals vs. significance tests: quantitative interpretation. (Letter) *American Journal of Public Health* **76**:587.
40. Fleiss JL. Dr. Fleiss response(Letter)(1986). *American Journal of Public Health*.**76**;1033-1034.
41. Walker AM 1986. Reporting the results of epidemiologic studies. *American Journal of Public Health* **76**:556-558.
42 . Poole C. (1987). Beyond the confidence interval. *American Journal of Public Health*. **77**. 195-199.
43 Rothman, KJ Significance Questing, 105 ANNALS INTERNAL MED. (1986).
44. MRFIT Investigators (1982) Multiple risk factor intervention trial. *Journal of the American Medical Association***248**:1465-77.
45. Feldman A.M., Bristow M.R., Parmley, W.W. et al. (1993). Effects of vesnarinone on morbidity and mortality in patients with heart failure.*New England Journal of Medicine* **329**:149–55.
46. Cohn J., Goldstein S.C., Feenheed S. et al. (1998).A dose dependent increase in mortality seen with vesnarinone among patients with severe heart failure.*New England Journal of Medicine* **339**:1810–16.
47. MRFIT Investigators (1982).Multiple risk factor intervention trial.*Journal of the American Medical Association* **248**:1465–77.
48. Pitt, B., Poole-Wilson P.A., Segal, R., et. al (2000). Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure randomized trial–The losartan heart failure survival study. ELITE II.*Lancet.***355**:1582–87.

49. Multicenter dilitiazem post infarction trial research group (1989) The effect of dilitiazem on mortality and reinfarction after myocardial infarction. *New England Journal of Medicine* **319**:385–392.
50. Packer M (2000) Presentation of the results of the Prospective Randomized Amlodipine Survival Evaluation-2 Trial (PRAISE-2) at the American College of Cardiology Scientific Sessions, Anaheim, CA, March 15, 2000.
51. UK Prospective Diabetes Study Group (1991) UK Prospective Diabetes Study (UKPDS) VIII.Study, design, progress and performance.*Diabetologia***34**:877–890
52. Moyé LA.(2003). *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*. New York. Springer.Chapter 8.
53 UKPDS Study Group.(1998). Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes**.***Lancet* **352**: 837–853.
54. Turner, R.C., Holman, R.R. on behalf of the UK Prospective Diabetes Study Group. (1998). The UK Prospective Diabetes Study. Finnish Medical Society DUOCECIM, *Annals of Medicine* **28**:439–444.
55. UKPDS Study Group.(1998). Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes**.***Lancet* **352**: 837–853.

56. The Lipid Research Clinic Investigators. (1979).The Lipid Research Clinics Program: The Coronary Primary Prevention Trial; Design and implementation. *Journal of Chronic Diseases*;**32**:609-631.
57. The Lipid Research Clinic Investigators.(1984).The Lipid Research Clinics Coronary Primary Prevention trial results. *Journal of the Amerian Medical Association***251**: 351-74.
58. Packer M., Bristow M.R. Cohn J.N. et al (1996) The effect of carvedilol on morbidity and mortality in patients with chronic heart failure *N Eng J Med* 334:1349-55
59. Transcript for the May 2, 1996 Cardiovascular and Renal Drugs Advisory Committee.
60. Moyé LA, Abernethy D (1996) Carvedilol in Patients with Chronic Heart Failure (Letter) *N Eng J Med* 335: 1318-1319.
61. Packer M, Cohn J.N., Colucci W.S. Response to Moyé and Abernethy (1996) *N EngJ Med* 335:1318-1319.
62. Fisher LD, Moyé LA(1999) Carvedilol and the Food and Drug Administration Approval Process: An Introduction. *Controlled Clin Trials* 20:1-15.
63. Fisher LD (1999) Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypothses testing *Controlled Clin Trials* 20:16-39.
64. Moyé LA (1999) *P* Value Interpretation in Clinical Trials.The Case for Discipline.*Controlled Clin Trials* 20:40-49.
65 Antrum M. (1901). Naked Truth and Veiled Allusions, p. 99.

66 . Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Hua Ho T, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE. Redefine statistical significance.Nat Hum Behav. 2018 Jan;2(1):6-10. doi: 10.1038/s41562-017-0189-

67 Johnson AL, Evans S, Checketts JX, Scott JT, Wayant C, Johnson M, Norris B, Vassar M. Effects of a proposal to alter the statistical significance threshold on previously published orthopaedic trauma randomized controlled trials. Injury. 2019 Nov;50(11):1934-1937. doi: 10.1016/j.injury.2019.08.012. Epub 2019 Aug 12.

What do we require of this new approach?

A principal problem with the current standard of use of statistical hypothesis testing in health care research is its strict testing regime. Since alpha is commonly prespecified at 0.05 level, there is actually very little error to be distributed among a collection of primary outcomes.

The problem worsens if there are too many outcomes, placing the investigator in the position of either having an enormous sample size, or enormous effect sizes, or unreasonably tiny standard errors in order to have a chance for a positive finding through statistical hypothesis testing.

These are considerations that have little to do with biology but are the artefactual consequences of applying statistical hypothesis testing and *p*-values to health care research, a field of application for which this class of mathematics was not designed.

Instead, these requisites are the price of admission that the clinical investigators must pay in order to be able to simply evaluate whether a treatment in a clinical trial is beneficial or not. Researchers are forced to place the clinical research square peg in the *p*-value round hole.

A new approach to clinical research assessment should have the following features:

- **Easily adapt to the multidimensionality of clinical outcomes**: A new approach should take advantage of characteristics of research that may not apply to other scientific fields. For example, there are many ways in which a treatment may be deemed beneficial. It can improve the subject's sense of well-being. It can decrease the likelihood that they will die in a given period of time. It can decrease hospitalization. It can improve measures of morbidity. The subject may exercise longer. They may have better organ (e.g., renal, liver, lung, or cardiac function). Benefit is multidimensional. An important feature of a new system is that it considers all assessments that the clinical trial makes concerning benefit – not just a subset of them.

- **It must be integrable**. Many estimates of benefit are available and have their own units. Others are per unit of time. Some are in percentages, others in milliliters. This tool must place all of these measures of benefit on the same scale so that these benefits can be accumulated. It we are going to assess the impact for each outcome, we must have a way to accumulate these benefits and a common scale permits that.

- **It must incorporate statistical estimators of different types.** The estimation field has and will continue to make substantial contributions to effect size estimation and its variability. This area will be wholly embraced. Therefore, our new procedures need to include

estimators from different types of analyses, be they simple, or more complex such as imputation, mixed model regression, or survival analyses.

- **It must acknowledge estimator variability.** The new tools must not only acknowledge, but take advantage of the observation that there is sampling variability in the estimates of an effect size. The relative risk for the reduction in fatal and nonfatal heart attack may be 0.84, but we must always be clear that, while this is the numeral that was computed in the research effort, sample to sample variability as well as precision concerns and sometimes bias influence this estimate.

- **It must be inclusive.** There are many analyses that are conducted in a clinical research effort. There are prospective analyses and retrospective analyses. There are analyses on different subpopulations. Proteomic analyses are of growing importance. We need a tool that will allow us to incorporate these different evaluations. The contributions of these analyses must be modulated by the experience and concerns of epidemiologists and clinical scientists (e.g., the primary of prospectively declared outcomes.

- **The tool must have be interpretable.** Any accumulation of benefit must be translatable for the research community, physicians, regulators, and patients.

- **Finally, it must be easy to use.** This is the 21st century. Investigators should have results produced quickly in easily interpretable tables.

Incorporating all of these features will require us acquire an entire new perspective for organizing analyses.

It begins with set theory.

The Basics of Set Theory

Motivation for this work

In health care research, we have collections; collections of questions to be answered, collection of analyses, collections of outcomes, collections of patients.

Our goal is to operate on these collections, extracting the cues and messages that they provide about the effect of therapy. Thus, we must have new abilities in manipulating collections of analyses, and these capabilities reside in set theory.

Set theory provides exactly the toolkit we need to first see how we can create these collections and then how to manipulate them. In doing so, we will find that the set theory tools that allow us to combine and disassemble sets will mirror the operations in measure theory that permit us to measure or value these sets.

So we will begin with some basic set theory followed by an elementary introduction to measure theory.

Those who already understand set theory may skip to Chapter 5 where we begin a discussion of measurable functions. However, for the rest of us, Before we get enter an exposition into set theory, Let's just talk about what you don't need to know to understand it.

-You do not need a degree in mathematics
-You do not need a statistics degree
-You don't need a calculus background
-You don't need trigonometry or geometry

All you really need is a willingness to understand and an understandable text. If you can bring the former, I commit to provide the latter below.

What are sets?

A *set* is simply a collection of objects. These objects can be physical, simply numbers, or metaphysical. It is defined simply by membership criteria.

For example, the collection of US coin denominations is a set. Let's call that set A and define it as {pennies, nickels, dimes, quarters, half dollars, dollar pieces}. If you have 26 cents in your pocket in nickels and a penny, then your set of coins is {N N N N N P} where N denotes a single nickel, P a single penny. Note that the set is denoted by braces {}.

Each distinct entry in our set is called an *element* in (or of) the set. We denote whether an element s is a member of a set S by the symbol $\in$; $s \in S$ simply means that element s is contained in set S, or "s is a member of S".

The order of elements does not matter in sets; sets that have the same elements but just arranged in different orders are equivalent. Thus {N N N N N P} is the same as {N N N N P N}. This greatly eases our burden in set manipulation.

Introducing relationships between sets

Two sets, denoted by A and B are equal, $A = B$ if they contain the same elements, (again, regardless of order).

Sets are defined by their content; this is equivalent to saying that sets are defined by their membership criteria, since it is the membership criteria permitting us to determine if an element is a member of the set or not.

It is remarkable that so much mathematical development with so many useful applications can be based on this simple and clear concept–set membership, and set comparison.

For example, consider the set of all subjects who are screened for a clinical trial. The membership criteria for this set is that each member has had their demography and comorbidity assessed against the study's inclusion and exclusion criteria. We can easily determine if an individual is a member of the set or is not.

To some degree, we are already familiar with sets in clinical trials; we just aren't used to thinking of these research collections that way. For example, the collection of all subjects accepted into a particular clinical trial is a set (the membership criteria is simply acceptance into that clinical trial). Similarly, individuals who are between 40 and 70 years of age in that clinical trial also comprise a set; in fact it's a subset of those individuals who are in the trial.*

As we just saw, sets can contain other sets. These contained sets are known as subsets Thus, if S is the set of all subjects in a clinical trial, and F is the set of females in the trial, then the statement that "the collection of females

* In clinical trial methodology, we commonly think of this collection of subjectsas a subgroup, but this collection also meets the definition of a set as well.

in the clinical F is part of S $\left(F \in S\right)$ is true. Females represent a subset of S. We can also say that "F is contained in S", or $F \subset S$. Another true statement is that "S contains F", or $S \supset F$

It will be quite useful for us to declare that a set has no elements. A set that has no elements is the "null set", denoted by $\{\ \}$, or more commonly $\varnothing$. Thus, if a clinical trial carried out no imputation analyses, its set of imputation analyses I is said to be null, we may write $I = \{\ \}$ or $I = \varnothing$ denoting that the set of imputation analyses is empty or null.

Set operations

One of the reasons that numbers function so effectively and practically in our society is because we can easily manipulate them. We can add to them, subtract from them, and compare them. We will need that same facility when working with sets. The principal set operations we will be working with are unions, intersections, and complements.

Unions

Let's begin with the set P that contains all subjects included in a health care research study. If the study has n subjects, then the set P contains the same n elements. We can already think of subsets of P such as the subset of individuals with LDL cholesterols greater than 175, or the subset of individuals who were exposed to potassium sparing diuretics.*

* Of course its possible that each of these subsets might be null if it has no subjects with these characteristics.

From the set of all subjects in the study P, let's define A_{40} be the subset of all subjects greater than 40 years old, and M as the set of all males. We can describe the people who are in either set as those subjects who are greater than 40 or are male. This we say is the union of A_{40} and M or $A_{40} \cup M$.

This union combines the elements from both sets into a new set. So, since we know that this union contains all subjects who are greater than 40 years old and addition contains any and all males regardless of their age, then we know that $A_{40} \subset A_{40} \cup M$. We also know that $M \subset A_{40} \cup M$. Now, we expect that there may be redundancy in this union. Males greater than 40 years old are in both A_{40} and M. However, the union counts them once and only once.

Working with unions requires practice. A good rule of thumb is to be absolutely clear about the set membership.

For example, does $A_{40} \cup M = P,$ or is it merely contained in P? It depends on the recruitment for the study. If the study only recruited males greater than 40 years of age, then $A_{40} \cup M = A_{40} = M = P.$

However, if the study contains one female less than or equal to 40 years of age, then $A_{40} \cup M \subset P.$ If we don't know the female's age, then we can say $A_{40} \cup M \subseteq P.$ which means that the set $A_{40} \cup M$ is contained in or equals the set of all subjects in the study, P.

Intersections

Now, continuing with the same example of the sets $A_{40}, M,$ and $P,$ we now ask who are the individuals in both A_{40} and M? If we let ω represent an individual P, i.e., $\omega \in P,$, then what are the characteristics of this individual ω when $\omega \subset A_{40}$ and $\omega \subset M$?

These are individuals in the study who are male and over 40 years old. We call this set the intersection of A_{40} and M symbolized as $A_{40} \cap M$.

This understanding of intersection is all that we need to say that $A_{40} \cap M \subseteq A_{40} \cup M.$ This statement is true because if an individual is in both sets, then that individual is in either of them and therefore, they are in the union. However, an individual could be in the union (e.g., a male who is 29 years of age), and would not be in both of them.

One way to remember this is that members of intersections are members in all sets in the intersection, i.e., if $\omega \in A \cap B,$ then $\omega \subset A$ and $\omega \subset B.$ However, a member of a union only need be a member of at least one of the sets of the union.

Complements

Finally, we have the notion of a complement. The complement of a set B is the set of all members who are not in the original set. The complement of B is denoted as $B^c.$ In our current example, A_{40}^c is the set of individuals who are less than or equal to 40 years of age, and of course, M^c is the set of all females in the study. Note that $A_{40} \cap A_{40}^c = \varnothing$ since individuals must be in one set or the other but not both; this is termed mutual exclusivity. Also,

$A_{40} \cup A_{40}^c = P$ since individual must have an age and that age is either less than or equal to 40 or greater than 40.

Can we subtract sets?
What would $A - B$ look like?

Let's think about what $A_{40} \cap M^c$ must be. We know that $A_{40} \cap M$ is the set of all males greater than 40 years old. Taking the intersection, not with M but with M^c leaves us the set of females greater than 40 years old. Another way to think of this is that the set $A_{40} \cap M^c$ is the set of all individuals greater than 40 years old with the males "removed". This combination of operators acts like we would expect the operator $A_{40} - M$ to operate. So, although one cannot technically subtract sets, the combination of the intersection and complement operation allows us to do accomplish exactly that.

Those who want to examine some more complicated manipulations in set theory can proceed to the intermediate set theory discussion. For the rest of us, Let's consider what we can do with these operations.

Venn diagrams

It does not take much imagination to understand that set operations and manipulations can become quite complex, e.g., $\left(A \cup B^c\right) \cap \left(D^c \cup C\right)^c$ for arbitrary sets *A*, *B*, *C*, and *D*. In order to help with visualization, Venn diagrams are particularly useful in being able to see and study the impact of these operations (Figures 1 and 2).

(Figure 1 here)

Let's consider some arbitrary sets A and B. from Figure 1 we can see that $A \cup B$ can be constructed from overlapping sets e.g., simply $A \cup B,$ or from three non-overlapping, disjoint sets, $A \cap B^c$, $A \cap B,$ and $A^c \cap B.$ *

We can also see that $A = \left(A \cap B\right) \cup \left(A \cap B^c\right).$

observing that $A \cap B$ and $A \cap B^c$ are disjoint sets.

The union operation can produce some interesting results (Figure 2).

(Figure 2 here)

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 that $A \cup B = \left\{A, B\right\} : A \cap B = \varnothing.$

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

Distribution Law of Sets

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

* Note from our earlier discussion that we can consider the set $A \cap B^c = A - B,$ and $A^c \cap B = B - A.$

The use of parentheses, permit us to avoid any ambiguity in the order of operations. For example, the operation $(A \cup B) \cap C$ contains every member of set A, every member of set B, and every member of set C, while the ambiguous statement $A \cup B \cap C$ could be the former, or could be the union of every element in set A with elements that are in both sets B and C. We always conduct the operations in parentheses first. If parentheses are nested, we begin from the inside and work our way out.

Another useful rule is DeMorgan's Law.

DeMorgan's Law

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

DeMorgan's law gives us a way to manage set complements. Taking a minute to think them through provides some insight. For example, the second law begins with $(A \cap B)^c$. This is a collection of individuals who cannot be in the intersection of sets A and B. Thus, they can in set A alone, B alone, or neither of them. This is precisely those who are not in A nor those who are not in B, as DeMorgan's law states,

Unions, intersections, and complements become more complicated when we consider the circumstance of three sets (Figure 3).

Taking the unions, intersections and compliments of sets is a straightforward way for sets to generate other sets, which themselves, through the same operation, produce additional sets. Essentially, these three operations put us in the set generation business.

(Figure 3)

Set generation and σ-algebras

As we have seen, the set operators union, intersection, and complement generate new sets that are related to but different from the original sets. The numbers of these new sets can be far larger than we might expect. For example, $\Omega = \{A, B\}$ we can generate sets as follows.

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

This is 11 sets that are generated from the original two sets. If $\Omega = \{A, B, C\}$, then we could generate many, many more sets. In fact, if there are n elements in Ω, then the number of subsets generated by these three set operations is greater than $n!$

This set generation feature is central to our future use of measure theory Specifically, all of the sets that are generated by these operations of union, intersection, and complement we will call the *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 Σ) that follows certain rules of inclusion, precisely satisfied by taking unions, intersections and compliments.

Creating the σ-algebra is straightforward; we start with a collection of sets, then generate from that collection the

null set, and every possible combinations of unions, and complements.*

The precise definition of the σ-algebra, Σ, of subsets Ω is the following collection of subsets;

a) The null set is a member of Σ, $\varnothing \in \Sigma$

b) If the set $A \in \Sigma$, then $A^c \in \Sigma$.

c) If a countable† number of sets $A_1, A_2, A_3, \ldots$ $A_n, \ldots$ are contained in Σ, then $\bigcup_{i=1}^{\infty} A_i \in \Sigma$.

This formal definition implies that intersections of sets are members of Σ as well. So, defining a σ-algebra in terms of unions and complements also implies that this *σ*-algebra must contain their intersections as well.‡ In the end, we have to simply keep in mind that a *σ*-algebra is nothing more than all of the subsets of elements contained in Ω.

Examples of σ-algebras : Consider a collection of five DVD's with unique titles T_1, T_2, T_3, T_4, T_5. The original set

---

* As an example, consider a set of tracks $T = \{t_i\}, i = 1...n$ be all of your music tracks. Then the $\sigma -$ algebra is all of the possible playlists you can construct from these tracks (including the playlist that contains no tracks at all!).

† By countable, we mean there is an infinite number of sets that correspond with the whose numbers. One can be begin with 1, and procede to infinity without missing any of the sets.

‡ Assume A and B are contained in Σ. Then A^c and B^c, must be contained in Σ. But their union $A^c \cup B^c$ must also be in Σ as must their complement $\left(A^c \cup B^c\right)^c$, which by DeMorgan's law is $A \cap B$.

of them is simply $\{T_1, T_2, T_3, T_4, T_5\}$. We can construct the σ-algebra Σ 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, the *σ*-algebra becomes very large. 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.

Painting: A particularly useful way to consider the role *σ*-algebras would be in painting. Suppose one has 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 "*σ*-algebra" consists of only two elements.

However, suppose you now add three additional gallons of different colors, one each for black, blue, and yellow. The *σ*-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 and so many others are all members of a huge mixture of new colors produced by

combinations of the original set. Since the original set was richer, the σ-algebra has exploded.

Subjects: As a final example, let Ω equal the number of subjects in a clinical research program. Here the σ-algebra is all of the subgroups of this population. This is each subject individually, then all subjects taken two at a time. Then taken three at a time and so on. The total number of ways to gather these individuals into different distinct collections is immense.

Of course the number of subgroups actually analyzed is infinitesimal compared to the total number of subgroups actually used in a clinical trial since only a small number are phenotypically meaningful.

Why we need σ-algebras

The concept of a σ-algebra is meaningful for us, because it will be the basis of an entirely different class of functions.

We typically think of functions as operations that map one number to another number, such as $y = x^2$. However, our new interest will require that we not just map numbers to numbers, but sets to numbers.

This will be a new matter for most of us. We will map items to numbers, and then collections of items to numbers. And the greater and more diverse the items in the original Ω, the richer the σ-algebra of events which will be the argument of our function.

For example, envision a system in which all analyses conducted in a clinical research enterprise are collected. We define this set of analyses Ω and then identify all possible sets and subsets of Ω. This resulting σ-algebra Σ of sets can be mapped to specific numbers based on their common traits or characteristics. For example, analyses

which produce estimators that suggest that the exposure is beneficial can be mapped to one number, while those that suggest harm can be mapped to another.

However, in order to keep good structure and order, we need to follow some mathematical rules in these set mappings. These rules help us to define measurable functions, and then, measure.

Elementary, Set, and Measureable Functions

Here we will describe the development of set functions that can be of use to us in health care research. We will with some very simple examples. This will set the stage for the definition of a measurable function and then of the concept of measure.

Let's continue where we left off. Assume that we have a sample space/σ–algebra complex denoted as (Ω, Σ) and that we have members ω_i such that $\omega_i \subset \Sigma$. Recall that the σ-algebra Σ can be explosive in size. However, our goal is not to simply tabulate elements of sets. Ultimately, we want to assign values to these elements, and then values to sets. In order to do that, we need a special type of function called a set function, and its simpler version called an elementary function,

Let's start with the easiest – an indicator function. It is denoted by $1_{[]}$ This function is defined as either 0 or 1 depending on the condition that is defined in the bracket of the subscript. For example suppose that Ω is the set of all statistical analyses conducted in a clinical research

endeavor and Σ is its σ-algebra. Define $f(\omega_i) = 1_{[\omega_i \text{ is a } t-test]}$. Then in order to assign the value of $f(\omega_i)$, we simply inspect ω_i to determine if it is a *t*-test. If it is, then $f(\omega_i) = 1$. It not, then $f(\omega_i) = 0$.

Note that we are using this indicator function to convert the presence of a condition for ω_i (which in this case is not a number) to a number.*

Also, observe that this function assigned a value to an element of the set (as opposed to an entire set of analyses). We call such a set function an elementary function, and denote it by $e(\omega_i)$. In our example, we would write $e(\omega_i) = 1_{[\omega_i \text{ is a } t-test]}$. An elementary function is a special set function that assigns a number to single element of a set ω_i that is contained in our (Ω, Σ).

For example, if (Ω, Σ) is the sample size and $\sigma-$ algebra of all subjects randomized to a clinical trial. Then we might create the elementary function $h(\omega_i)$ to determine if the i^{th} subject if Hispanic. It would operate on each and element in Ω. Specifically, we would write this function as $h(\omega_i) = 1_{[\omega_i \text{ is Hispanic}]}$. This function essentially inspects each subject's ethnicity and assigns that subject the value 1 is they are Hispanic, and 0 otherwise. We can

* Also, in order for us to even assign the number, the property (in this circumstance, that property is whether the analysis is a *t*-test or not) had to be available for inspection. This availability is an essential feature of a measureable function, which we will discuss later.

imagine other such functions for age, gender, and combinations of demographic factors.

The elementary function is a special indicator function that maps a single element of a set to either 0 or 1; its domain is a singleton element, ω_i of a set.

However, it must also have the characteristic that the property inspected by the elementary function must be a property that is available to be inspected.

Measurability

Measurability is a critical concept in the development of set functions. However, it is a concept that can be easily explained for non-mathematician.

Measurability is a property of a function. A function is either measurable or non-measurable. Our work will concentrate on measurable functions.

Measurable functions have two properties.

The first is that the function itself must return either zero or a positive value. It cannot be negative,* This property is automatically handled by the 0-1 definition of the indicator function. This is straightforward.

The second property involves the inspection of ω_i. For example, the function $h(\omega_i) = 1_{[\omega_i \text{ is Hispanic}]}$ inspects for the ethnicity property. If members of the set Ω have that property be inspected, then the set function is a measurable function on Ω.

* We will compensate for this limitation in value by showing that we can premultiply by a sign, such as $-h(\omega_i)$.

A function that would not be measurable on Ω would be $a(\omega_i) = \mathbf{1}_{[w_i \text{ is an Android phone user}]}$. The type of smart phone that a subject in a clinical trial possesses (or whether they actually use one) is not available in clinical trial databases. Thus the function cannot operate because the property that it wishes to inspect is not available.

As seen in this example, there is no use for us to develop non-measureable functions for our application to clinical trials. Thus, the functions that we develop will be measurable on (Ω, Σ). However, the availability of this property we must always keep in mind.

For example if (Ω, Σ) is the set of all analyses conducted in a clinical trial, and $r(\omega_i) = \mathbf{1}_{[w_i \text{ is a regression analysis}]}$, then by definition $r(\omega_i)$ is nonnegative and measurable.

As another example, we can define a elementary function for an analysis as 1 if the analysis contains patients with diabetes and 0 if does not. This would be symbolized as $e(\omega_i) = \mathbf{1}_{\omega_i \text{ contains diabetics}}$. If the members of (Ω, Σ) contain information about the morbidity of patients in the clinical trial (a likely set of circumstances) then $e(\omega_i)$ is measureable.

If we have a collection of participants in a clinical study and we want to know how many of them returned for a visit six months into a study, we could define the set Ω as the set of all individuals ω_i in the study and the set B is the event that a participant returned for the six month visit. Then if we define $e(\omega_i) = \mathbf{1}_{\omega_i \subset B}$ for each subject ω_i in the

study, then each ω_i is mapped to either 0 or 1 depending on whether the individual returned for their six month visit. This is measurable or not based on whether each element ω_i contains the follow-up information. For example, if (Ω, Σ) is created at baseline, the function is not measureable because follow-up information is not yet available. However, if (Ω, Σ) is created at the end of the study when follow-up information is available, then $e(\omega_i)$ is measurable.

Broadening the elementary function

At this juncture, we are comfortable with the notion of an indicator function, and at least know how to determine if it measurable or not. Let's now introduce the concept of the more general set function.

A set function maps a set (not just the element of a set like a elementary function does) to 0 or 1. Unlike the elementary function, its argument is a more general set (that can contain more than just one element). Thus set-indicator functions are related to, but different from elementary-indicator functions. An elementary-indicator function maps an individual element of a set to 0 or 1 based on whether that element is in another set. The broader set function assigns 0 or 1 to the set itself.

As an example, Let's return to our clinical trial example where (Ω, Σ) contains the individuals in a clinical trial at baseline. Let the set of all males be *M*. Since M is a subset of Ω, then $M \subset \Sigma$. Now, consider two functions. The first will be the elementary function $e(\omega_i) = 1_{\omega_i \subset M}$,

assigning 1 if the ω_i^{th} individual is male, 0 if not. The function $e(\omega_i)$ is our standard elementary function.

Now let's create a new function *g*. Its argument is any subset *A* of individuals in Ω, and $g(A)$ is defined as the number of males in set *A*.

Both $e(\omega_i)$ and $g(M)$ are related to each other since they each involve the set *M*; however they use *M* differently. The function $e(\omega_i)$ uses *M* to identify a characteristic of the individual ω_i and assign a value to ω_i based on the presence or absence of that trait (in this case, the characteristic of male gender). It operates on any $\omega_i \in \Sigma$.

However, the function *g* does not concern itself with a trait of a single member of a set, but of the entire set itself.* So while $e(\omega_1, \omega_2\, \omega_3)$ cannot be computed, $g(\omega_1, \omega_2\, \omega_3)$ makes perfect sense. These two functions are related by $g(M) = \sum_{i=1}^{n} e(\omega_i) = \sum_{i=1}^{n} 1_{[\omega_i \subset M]}$.†

We will find in public health problems, that it is usually easier to start with the elementary function when building a complicated simple function, and then to convert it to a more general set function.

We can also easily see that functions constructed from measurable, elementary functions are also measurable. As an example of why this is true, suppose that we define Ω

---

* Of course one can make the argument that the individual ω_i's are sets themselves, but the distinction here is whether the argument of the function can have more than one element.

† We will see that in general this can be written as $g(M) = \int e(\omega_i)$.

as a collection of individuals on whom demographics including age and race are available, and Σ is the σ-algebra of these individuals. Now consider the function $k(\omega_i) = 1_{age \leq 45} + 1_{Hispanic}$. For any value of $k(\omega_i)$ one can find the age and ethnicity of individuals so that the function value can be assigned. The key to this is noting that since the elementary functions are measureable, their sums and difference are measurable as well.

Summary

Let's take a breather for a moment and figure out where we are.

We began with a very basic definition of collections or sets of items, and simple ways to combine and compare two of them. This led to the construction of a collection of sets reflecting all of the many possible ways the original set's elements could be combined.

One example that we have used in clinical research and will continue with is the collection of all analyses conducted. We call that collection of all analyses the set Ω.

From Ω we have generated the large collection of sets reflecting all possible combinations of analyses. This huge collection of sets are gathered together into a σ-algebra, we term Σ.

Commonly Ω and Σ are combined into their own collection (Ω, Σ). With this as a foundation, we defined the notion of a set function which simply maps members of Σ to a real number. A measurable function is a set function whose criterion for inspection is a property of each element in Σ (i.e., mapping criteria is available in Σ), and whose possible values are only positive.

There is a huge number of measurable functions on the (Ω, Σ) analysis space. For example, if $\{\Omega, \Sigma\}$ represents the set of all hepatocytes in a subjects liver, then one possible measure $v(A)$ might return the ALT enzyme content of that set, and a wholly separate measure $\varsigma(A)$ could return the AST consent 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 bilirubin? For example, the measurable function $g(\omega_i) = 3\mathbf{1}_{\omega_i \subset S}$ where S is the set of all subgroup analyses is a measurable function on our (Ω, Σ) foundation.

However, we will need to discipline ourselves to identify the set of measurable functions that will be the most helpful, and that also reflect the research circumstances. This will provide a foundation that is custom-tailored for health care research analyses.

Now, we will take the measurable approach one (final) step further, and identify, a way to assess the value or the contribution of a set (e.g., a collection of research analyses). This is determining the measure of a set.

Measure and its Properties

Measure theory is based on the idea that the content of a set can be mathematically assessed and valued. At this point in our development, we can manipulate sets; now we begin to study how we can measure them.

It is tempting to think that we can already do this with measurable functions, since after all these map the content of a set to a nonnegative number. However, we will see that measure requires more than just assigning a number. There must be an induced relationship between the value of different sets.

For example, if a first set is wholly contained in a second set, shouldn't the value of the second set be at least as great as the value of the first set? Measurable functions do not of necessity do that.

On the other hand, we will also see that we are already in the habit of providing content for sets, although we are not used to thinking of it that way.[*]

[*] One such example is assigning probability to events. Probability is just one of many different ways of assigning measure.

Here, we will spend some time providing the background for measure that we will be able to use in health care research, our ultimate goal.

The set theory however does get somewhat thick. To help you with its absorption, I have created two paths.

The elementary path is for those who have absolutely no background in measure therory. This path will not make you an expert, but you will learn what you need to know in order to understand the basic properties and that it has many applications.

The intermediate path delves into the mathematical underpinnings of measure theory without a nonmathematical preamble. There are substantial set theory findings on this tack. The advanced path dives straight into the development of a new measure for clinical research.

A useful approach would be to take first the elementary path, and then the intermediate one.

Elementary path

So, let's begin with what measure theory is, and start with a story, or, more to the point, a kidnapping.

> *"Don't make a sound, put your hands in your pockets, and get in the back of that car."*
>
> *Your neck is yanked by a hand whose vice-grip forces your head into a paper bag smelling sweet-sick like old milk. Shoved backwards into a car seat, your arm scraping the rough edge of the cold, torn leather, you hear the door close and feel the car peel away from the curb.*
>
> *After several abrupt turns,, other car horns blaring in protest at this vehicle jerking*

movements, you lose track of the sequence of lefts and rights.

No idea where you are.

The car screeches to a halt, the door opens with a bone chilling squeal, and you are pulled out and shoved into a hot and humid room. Pushed down into a chair, the smelly paper bag is removed from your face.

Blinking your eyes back to focus, you see before you

Money.

Piles of money.

Heaped to overflowing on a round table in front of you.

Shaking your head to separate yourself from the smell of the bag, you inch closer to see that although it is money, it doesn't look exactly like money.

"Yeah, that's right," scratches the reedy voice of a women from behind. "All kinds. Dollars, pounds, sheckles, drachma, francs, rubles, deutschmarks. We even have a few thousand native beads in there. We need to know how much it all is worth. Why do you think we brought you here?"

For us, counting all of that money in its various forms goes to the heart of our use of measure theory.

What is Measure Theory

Measure theory, at its heart is the science of measuring the accumulation of things. Sometimes the rules of accumulation are complicated. Other times (e.g., simple counting) the gathering is easy to follow. In any event, the ideas are always straightforward.

Measure theory and its generalities envelopes much mathematical content; probability is one of its subfields. Developed by the French mathematician Lebesgue, and the Russian probabilist Kolmogorov in the 1920s, measure theory is based on set functions, those functions that we discussed that map sets to numbers. Ultimately, we take a collection of sets, assess each set's value, and accumulate that value over each of the sets. In the end, we have the accumulated total value of all of the sets. And since sets can be arbitrary we have substantial freedom in choosing the sets that are to be valued. However, we have to follow certain mathematical rules in this valuation.

Accumulation

For our purposes, measure theory focuses on the process of accumulation (gathering together, or rounding up, or congregating) the value of items.

We already know how to combine items, if we think of set elements as the items. The operations of unions, intersections, and complements are the manipulations by which set elements are combined or gathered into new sets.*

---

* The operation of computing the probability of an event by 1) showing how that event is the "combination" of of simpler events whose probability is easy to find, and 2) using the rules of set operations to reconstruct the event of interest from the simpler sets, and 3) then applying the rules of probabilty, operation by operation, to build up the

This accumulation process is the heart of measure theory. What can appear complicated about measure theory is that the accumulation process may be complex, using different procedures to measure different quantities. However, what appears at first glance to be a complicated myriad of arbitrary rules turns out upon further inspection to be precisely the combination of procedures required to accurately accumulate the required quantity. Let's start with some easy examples.

Example – measuring wealth
Consider the task of measuring the wealth accumulation for a typical 5-year-old US boy over the course of his life.[*] At the beginning of the process, everything that this five year old possesses (e.g., clothes and toys) is purchased by his parents. The wealth that he has truly earned comes solely from his own weekly allowance or small financial gifts, an allowance that can be measured by simply counting up the value of the coins that he is either paid or that he occasionally finds (e.g., from the tooth fairy). Since collecting coins is the only way by which he accumulates his own independently earned wealth, we are content to only count the value of coins. This is easy – we know how to "measure" coins.

---

probability computation of the more complicated event from the simpler set, comes from measure theory.

[*] Adapted from Kapadia AS. Chen W. Moyé LA. (2005) *Mathematical Statistics with Applications*: New York. Taylor Francis.

However, as the boy grows he starts to accumulate wealth in other ways. One natural development is to earn not just coinage but paper money. This change poses a dilemma for our measurement of his accumulated wealth. If we continued to base our measure of his wealth solely on the value of coins, we would miss an important new (and greater source) of his earned wealth, and consequently, the estimate of this earned wealth would be in error.

We therefore very naturally alter our wealth counting mechanism to now include a new counting procedure – the accumulation of the value of paper money.

Note here that the tools used to count money have changed (from coin value to a combination of coin value and paper money value), but not the goal of measuring his accumulated wealth. Our counting mechanism had to flexibly adapt to the new economic situation if it was to remain accurate. Since accuracy is the key, we change the manner in which we count but we remain true to the process of assessing wealth accumulation. We could say that we adjusted our "measuring tool" to now count not just coins, but dollars as well.

Additional changes in how we accumulate wealth are required as our subject prepares to go to college. How should the process adapt to the mechanism of the boy (now a young man) who uses his own money to buy a car?

Simply continuing to merely count coin and paper money value as a measure of his independently acquired wealth clearly introduces inaccuracies. As he acquires smart devices, computers, cars, obtains a paying job, invests in stocks and bonds, buys and sells homes, etc., our accumulation process, which started with simply recognizing coin denominations, must adapt repeatedly to

include these new forms of wealth. Again our rules of accumulation had to adapt to the changing, increasingly complex reality of the circumstances.* Yet the goal of the process remains the same — the "measure" of the man's wealth.

This is a complex process that produces in the end a fairly complicated function. However, while that function may not be recognizable at first glance, we understand how it was developed, and can use it to "measure" the individual's worth.

This we have done by simply creating sets where the assessment of wealth is the same type of value for each element (coin set, property set, stock set, etc.) and then we assign the right value to each set, and subsequently accumulate the value. Thus the sets have to be evaluable by the measure, and we may need to apply a different measure (or value assessment) to each set

Example - Music Tracks

Many people now manage their songs (or tracks) digitally. Suppose an individual with several thousands tracks wishes to get a sense of this collection's value or worth. How could they do this?

One way would be simply count the number of tracks, beginning with perhaps the oldest and moving to the latest downloads, increasing the count by one for each distinct track. This is simply and naively, "counting measure". In the end, one knows the total number of tracks, and in a sense, one has "measured" them.

* The consideration of depreciation of these material assets over time is yet one more easily added addition to our increasingly complex rules in estimating this individual's wealth.

If we call this collection of tracts A, then the value of the tracks might be as easy as $\mathrm{V}(A)=\sum_{i=1}^{n}\mu(i)=\sum_{i=1}^{n}1=n,$ where the measure of the i^{th} track, $\mu(i)$, is simply "1", and n is merely the number of tracts in the collection of music A.

However, another equally valid way to proceed is to place a value on each track, for example, the number of times that track has been played. Many tracks may have never been played, while others may have been played hundreds of times.

In this example, one accumulates the "size" or "measure" by adding not the track, but the number of plays that it has been played. This will lead to a different measure of the music collection. Let $\eta(i)$ be the number of times that the i^{th} track has been played. Then, for this type of accumulation, $N(A)=\sum_{i=1}^{n}\eta(i),$ which is the accumulation of plays and $N(A)$ is the measure or the value of the music collection A.

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). $D(A)=\sum_{i=1}^{n}d(i),$ would be the accumulation duration playing time for the music collection A.

Which of these "measures" is right?

They are all right. Each (total tracks, number of plays, and total time), is legitimate because each value or measure is based on, or can be traced backed, to a measurable characteristic of the music collection. However, each measure is different, because it emphasizes a different property of the music.

Example: Clinical Research Reimbursement

As a final 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.

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

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

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)

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 further and define measure μ as the reimbursed cost for each visit. So the cost for 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 $\mu_A(V_1 \text{ and } V_2) = \mu_A(V_1) + \mu_A(V_2) = 16.7 + 16.7 = 33.4$. 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 all types of relationships between these measures that are induced by the system of payments.

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

Notation

In order to help us, we will need more notation. Typically, the symbol used in measure theory is $\mu(A)$ which means "the measure of the set *A*". For example, the notation $\mu(A_1 \cap A_2 \cap A_3)$ denotes the measure of the set which is the intersection of the objects in the sets A_1, A_2, and A_3. It says nothing about how we actually take the measure, but instead simply signals our intent to carry out the measure procedure.

We will use the "integral" sign the same way, Like the $\mu(\ )$ notation, $\int$ simply announces that we will be measuring a collection of objects. They may be discrete objects, intervals on the real line, volumes of space, (or combinations of all of these different objects.) For example if A is a collection of analyses in a clinical trial, then the notation $\int_A d\mu$ simply means that we want to accumulate the measure or value of those analyses in set A using measure μ. Now we do not know what measure μ is at this point and will define it later. However, this is how we will use it. This can be a little disconcerting to an enthusiast of integral calculus. *

---

* Calculus studens are used to a collection of formulas denoting how to "integrate", such as $\int \cos(x)\,dx = \sin(x)$, or $\int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t}$.

However, it is useful at this point to take a step back and see what we are doing. The classic way 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 a collection of points, in these circumstances, an interval on 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

$$\int_a^b dx = b-a : \int_a^b \cos(x)\,dx = \sin(b)-\sin(a) : \int_a^b \lambda e^{-\lambda x}dx = e^{-\lambda a} - e^{-\lambda b}.$$

Each of these three integrals does something different with the 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 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.

Working with measure's first three properties

Taking the measure of simple sets is straightforward. However, commonly, simple sets have little interest for us. We are interested instead in the measure of complex sets.

We will see that to find the measure of complex sets, we will build the complex set us from a collection of simple sets, using the set operators of union, intersection and complement. If we are to find the measure of the complex set, we must identify the measure operator equivalents of these set functions operations. We will begin to do this in this chapter.

Measure is typically taught as having four properties. This chapter will focus on measure's first three, leaving countable additivity to its own chapter. These first three properties are quite natural and intuitive, permitting us to develop several examples using the measure concept. With the experience we gain from these examples, we will be able to appreciate the need of the fourth property.

Review of the sample space and sigma algebras

Recall that the sample space Ω is the beginning source of set elements that interest us. The members ω of Ω are the building blocks of sets that hold the greatest interest. The set Ω can have a small number of events (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 Ω reside only within the scope of the problem and the imagination of the worker.

Once Ω is established as the foundation, the σ-algebra Σ is constructed. Think of Σ as a set generator; it is nothing more than the collection of sets built from a combination of the elements in Ω using the elementary set operations of unions, intersections, and complements.

Every element ω_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 (e.g., the cubic nm of extravascular space) then the σ-algebra Σ can be quite overwhelming.

Measure vs. measurable functions. Properties of measure

Once Σ is identified, we are free to create a measurable set function on it. Remember that the only properties that a measurable function must have is that it must be

* If for example, Ω contains three and only three elements, Σ contains over thirty elements.

nonnegative, and that every value that it takes must map back to a set in Σ. Recall that we have tremendous freedom in defining measurable functions.

However, the actual measure of a set requires a more intricate operation than that conducted by a measurable function. Measure assigns content to a set. In order for a measurable function to be a measure, it must have the following four properties (three of which we focus on in this chapter:

We will elaborate on these properties, but first here is there simple listing.

Measure property 1

If set A is a member of Σ , then $\mu(A)$ (called "the measure of A") must be a $\mu(A)$ non-negative real number.

Measure property 2

If μ is a measure on (Ω, Σ) then $\mu(\varnothing) = 0.$

Measure property 3

If sets A and B are both elements of Σ 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 $\mu(B) \geq \mu(A).$

Measure Property 4 Countable additivity

(discussed in the next chapter):

If the infinite sequence of disjoint sets A_n is contained in $\Sigma,$ then $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$

Of these properties, property 1 (nonnegative value) is the only one that measurable functions and measure have in common.

We can now explicate each of these three properties and their implications for measure.

Measure property 1:
If set A is a member of Σ , then $\mu(A)$ (called "the measure of A") must be a $\mu(A)$ non-negative real number.

Just as measurable functions, a measure μ is a set function. This real number, $\mu(A)$ is the measure of, the content of, or the value of the set A. And, again, like a measurable function, its assigned value must be to a nonzero number. How the set is converted into a number is the property of the measure. However, the measure or content itself must be a non-negative number.

Measure property 2: If μ is a measure on (Ω, Σ) 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 $\varnothing$ 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 analyses in a clinical trial, it makes little sense to ask what is the measure or content of "no analysis". The statement $\mu(\varnothing) = 0$ is a mathematical statement of that reality,

thereby making the concept of measure more universally practical.

However, the statement that $\mu(\varnothing)=0$ has other important implications for us. For example, select two sets A and B from a (Ω,Σ) collection of sets such that the two sets are disjoint. Then we know that $A\cap B=\varnothing,$ and therefore, by definition for any measure μ, $\mu(A\cap B)=\mu(\varnothing)=0.$

Thus the measure of the intersection of disjoint sets (however that measure may be defined) must be zero .

We will see that this simple implication from among the most intuitive of measure theory principals has critical implications for our work in clinical research.

Measure property 3: If sets A and B are both elements of (Ω,Σ) such that A is contained in $B,$ then $\mu(A)\le\mu(B).$ Another way to say this is that if B contains A , then $\mu(B)\ge\mu(A).$ This is known as the principal of monotonicity.

It might not be evident at first blush, but property three tells us how to compute the union of sets.

Measure of the union of two sets

This is our first attempt to find the measure of a combination of sets. It is important to understand this concept completely because not only does it require us to review what we know about set theory, but it will provide

much new practice in building up the measure of complicated sets from simple ones.

For this development, Let's assume that we are working from a standard (Ω, Σ) collection of sets and that sets A and B are selected from Σ.

To begin with, we note from property 2 ($\mu(\varnothing) = 0$), that if our two sets are disjoint, then $A \cap B = \varnothing$, and thus $\mu(A \cap B) = 0$.

So the solution for the measure of the intersection of disjoint sets is already available to us. However, what is the measure of their union?

In order to examine this concept, simply, consider a set A_1 which contains a single element ω_1. Then $\mu(A_1) = \mu(\omega_1)$.

Now, consider a set A_2 with two elements $\{\omega_1, \omega_2\}$. What is $\mu(A_2)$?

If ω_2 is the same element as ω_1, then

$$\mu(A_2) = \mu(\{\omega_1, \omega_2\}) = \mu(\{\omega_1, \omega_1\}) = \mu(\{\omega_1\}) = \mu(A_1) = \mu(\omega_1)$$

However, if the element ω_2 is not the same as ω_1, (which means that these elements are disjoint), and if measure is to serve as a process of accumulation, then the natural conclusion is that $\mu(A_2)$ is the sum of the $\mu(\omega_1)$ and $\mu(\omega_2)$. This is how we would expect the accumulation process to work.

Now, let's talk in general about two sets A and B. Another way to write the set A is $A \cup \Omega$. Still another way to write Ω is $\Omega = B \cup B^c$. Then we have

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap \left(B \cup B^c \right) \\ &= \left(A \cap B \right) \cup \left(A \cap B^c \right) \end{aligned}$$

Analogously, the set B can be written as $B = \left(B \cap A \right) \cup \left(B \cap A^c \right)$. Thus

$$\begin{aligned} A \cup B &= \left(A \cap B \right) \cup \left(A \cap B^c \right) \cup \left(B \cap A \right) \cup \left(B \cap A^c \right) \\ &= \left(A \cap B^c \right) \cup \left(B \cap A^c \right) \cup \left(A \cap B \right). \end{aligned}$$

This we have seen from the chapter on set theory. Figure 3,reproduced here.

(Figure 1)

From this formulation, we can see that the set $A \cup B$ is composed of three subsets; 1) the part of A that does not contain B, 2) the part of B that does not contain A, and 3) the elements common to both A and B, namely $A \cap B$.

By this restructuring of $A \cup B$ into the union of three sets, we notice that the terms on the right are pairwise disjoint: for example, for an element to be in $A \cap B^c$ it must be in B^c which excludes it from the sets $A^c \cap B$ and $A \cap B$. Since these sets are disjoint, we can sum their measures.

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

This is the most general solution for the measure of the union of two sets and one that we will take advantage of

Special cases of the union of two sets

With this as background, we can now examine some special cases of $A \cup B$. For example, if $A = B,$ then our intuition tells us that $\mu\left(A\right) = \mu\left(B\right).$ We can show that from our previous formulation of $A \cup B$ as

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

In the case where $A = B,$ then $A \cap B^c = \varnothing, A^c \cap B = \varnothing,$ and $A \cap B = A = B,$ so $\mu\left(A \cup B\right) = \mu\left(A\right) = \mu\left(B\right).$

As another example, if A and B are disjoint, then $\mu\left(A \cap B\right) = \mu\left(\varnothing\right) = 0.$ We can then write $\mu\left(A \cup B\right) = \mu\left(A \cap B^c\right) + \mu\left(A^c \cap B\right).$ Then writing $A \cap B^c = A$ and $A^c \cap B = B,$ we have $\mu\left(A \cup B\right) = \mu\left(A\right) + \mu\left(B\right).$

These have been the simple cases. But what if the two sets are neither equal nor pairwise disjoint, i.e., $A \neq B$ and $A \cap B \neq \varnothing$?

Our background in set theory will help us find and bound $\mu\left(A \cup B\right).$ We know that we can write

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

so that

$$\mu(A) = \mu(A \cap B^c) + \mu(A \cap B)$$
$$\mu(B) = \mu(A^c \cap B) + \mu(A \cap B).$$

We sum these two expressions to find that

$$\begin{aligned}\mu(A) + \mu(B) &= \mu(A \cap B^c) + \mu(A \cap B) \\ &\quad + \mu(A^c \cap B) + \mu(A \cap B) \\ &= \mu(A \cap B^c) + 2\mu(A \cap B) + \mu(A^c \cap B)\end{aligned}$$

However, we know from before that

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

Which, comparing term by term is less than

$$\mu(A \cap B^c) + 2\mu(A \cap B) + \mu(A^c \cap B)$$

We may write in general that

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$$
$$\mu(A \cup B) \leq \mu(A) + \mu(B).$$

Using these same three simple properties of measure, we may find the measure of complements of sets. We begin by writing $\Omega = A \cup A^c$. Since the sets A and A^c are mutually exclusive, we can write $\mu(\Omega) = \mu(A) + \mu(A^c)$, or $\mu(A^c) = \mu(\Omega) - \mu(A)$. If the measure of the sample Ω is finite and known, then we can find the measure of A^c from $\mu(A)$.*

As another example, let's use what we know about manipulating sets and their measure to show that if $A \subset B$, then $\mu(A) \leq \mu(B)$. Again, our intuition tells us that this should be true; if set A is contained in set B, then B contains A plus "something else". If the measure of that 'something else' is not zero, then $\mu(A) \leq \mu(B)$. With this helpful thought process behind us, let's now apply what we know of measure theory to this simple problem.

We know that $\mu(A \cup B) = \mu(A \cap B^c) + \mu(A^c \cap B) + \mu(A \cap B)$. In this case where $A \subset B$, then $A \cup B = B, A \cap B^c = \varnothing,$ and $A \cap B = A.$

Thus $\mu(B) = \mu(A^c \cap B) + \mu(A)$. If $\mu(A) \neq \varnothing,$ then $\mu(A^c \cap B) \geq 0,$ and $\mu(B) \geq \mu(A)$. If $A = \varnothing,$ then this equality reduces to $\mu(B) = \mu(\Omega \cap B) + \mu(\varnothing) = \mu(B)$. This is a demonstration of measure property 3.

---

* If $A = \varnothing,$ then $A = \Omega,$ and $\mu(A^c) = \mu(\Omega)$.

Returning to $\mu(B) = \mu(A^c \cap B) + \mu(A)$, simple subtraction reveals that $\mu(A^c \cap B) = \mu(B) - \mu(A)$, another finding of which we will make use.

Summary

So in this chapter, we have explored three properties of measure. Measure is based on sets; we will always apply measure to sets. Thus, our ability to manipulate measure is tied directly to our ability to manipulate sets. Exploring some of the implications of the first three properties of measure permitted us to develop the measure of the intersection of sets, and the measure of the union of sets.

In fact, from three simple properties of measure, we can find the measure of unions and complements of sets. This is one of the most important components of measure theory that we will use. First, we will construct an (Ω, Σ) collection of sets. Then we will identify the set whose measure we want by building that set up from an intelligent combination of set operations of unions and intersections.

This combination of set operations will be paralleled by adding and subtracting the measure of sets that will get us to the measure of the set we ultimately desire.

We are now ready to examine the fourth property of measure – countable additivity.

Property 4 of Measure: Countable Additivity

Remember that the first three properties of measure, (measure must be non-negative, measure of the null set is zero, and the measure of one set that contains another) allow us to assemble the measure of sets that are built up from other sets.

The motivation for properties 1 – 3 (non-negativity, the zero value of emptiness, and the relative value of sets containing each other) comes from the need and desire to bring useful and intuitive concept to measure. We want it to assess the content or value of an item in a way that matches our intuition (value is never negative and the value of "nothing" is zero), and we want also want to measure to accumulate (and not decrease) over a collection of sets, each one containing the preceding one. These three properties built on a basis of set theory ensure that.

However, property four (countable additivity) has a different motivation. It has little to do with how we assign measure to a set, but is more focused on the actual utility of measure.

Specifically, given that measure is assigned to a collection of sets, how can it be used to assign measure to

other more complex sets that are formed from that original collection? *

For example, suppose that we want to measure a set A. We know that A can be produced from a collection of sets $\{A_i\}$. and that each have known measure. If this collection of sets whose measure we know builds the set A from the union and intersections of the members of $\{A_i\}$, then we can build up to the measure of A by what we know of computing the measure of unions and intersections of $\{A_i\}$ no matter how many sets are contained in $\{A_i\}$. This is one way in which countable additivity is important.

Measure Property 4 Countable additivity:
If the infinite sequence of disjoint sets A_n is contained in Σ then $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$.

This can be proved using an induction argument. †

---

* We have seen a rudimentary example of this with the determination of the measure the union of two sets.

† The veracity of this property can be developed through an induction argument. For $k = 1$, $\mu(A_1) = \mu(A_1)$. If we assume $\mu\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} \mu(A_n)$, then for the k+1st set A_{k+1} is disjoint from

Note that the upper bound of the index in this property is infinity. There is another concept where the upper bound is finite, termed finite additivity i.e., $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} \mu(A_i)$. This is a consequence of the countable additivity property which can be easily demonstrated.*

Non null intersections sets within the sequence of sets $\{A_i\}$ requires us to modify the assertion of property 4 to

---

$A_1, A_2, A_3, \ldots A_k$ and is therefore disjoint from their union. Thus

$$\mu\left(\bigcup_{n=1}^{k+1} A_k\right) = \mu\left(\bigcup_{n=1}^{k} A_k \cup A_{k+1}\right) = \mu\left(\bigcup_{n=1}^{k} A_k\right) + \mu(A_{k+1})$$

$$= \sum_{n=1}^{k} \mu(A_n) + \mu(A_{k+1}) = \sum_{n=1}^{k+1} \mu(A_n),$$

completing the induction argument.

* We note that $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$. Now Let's 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

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{k} A_n \cup \bigcup_{n=k+1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{k} A_n \cup \bigcup_{n=k+1}^{\infty} \varnothing\right) = \mu\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} \mu(A_n).$$

Thus, for this collection of A_n, the infinite union reduces to a finite union of exactly the A_n sets that we want.

say that, if the sets are not disjoint, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{n} \mu(A_i).$$ *

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 their measures for equality. 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, and then using the property of countable additivity to sum the measure of these disjoint sets.

With this as background, we are now ready to consider the "content of a clinical research analyses", in a way that we can apply a measure theoretic approach to its accumulation.

* We have seen this demonstrated in the previous chapter with the union of two sets.

An interlude…

Let's just pause for a second and see where we are.

The preceding set and measure theoretic preamble permits us to compute the measure of combinations of sets. If we start with a collection of sets each of whose measure we know, we can now compute the measure of a more difficult and intricate set by using the rules of measure in parallel with the set operations (unions, intersections, and complements of the original simpler sets) to build up the measure of the final set.

But what does this have to do with clinical research?

We began with the notion of duality, i.e., the idea that a single estimate from clinical research can be evidence both for benefit and for harm. There, we described a process by which a region of plausible values of effect sizes (i.e., a plausible interval) could be parsed into one for benefit, and the other for harm. For benefit, this is $\chi_i^{(b)}$, and its benefit function $\mathbf{Y}_b\left(\chi_i^{(b)}\right)$.

Our concept is that this region and function can be identified for each analysis. However, we need to accumulate them over all analyses, $\int_{A\subset\Omega} \mathbf{Y}_b\left(\chi_i^{(b)}\right)$. But how do we compute this when the individual analyses overlap?

What do we do about the redundancy of observations and variables that are used repeatedly in succeeding analyses?

Specifically, the collections of analyses utilize overlapping collections of observations or variables. An examination of a mean difference, and the assessment of an effect using a general linear model with adjustments for covariates are different analyses, but can have observations in the first analysis also included in the second analysis, and variables in the second analysis contain variables in the first analysis. And, the more analyses that were conducted, the more intense the analysis is likely to be. (Figure 1).

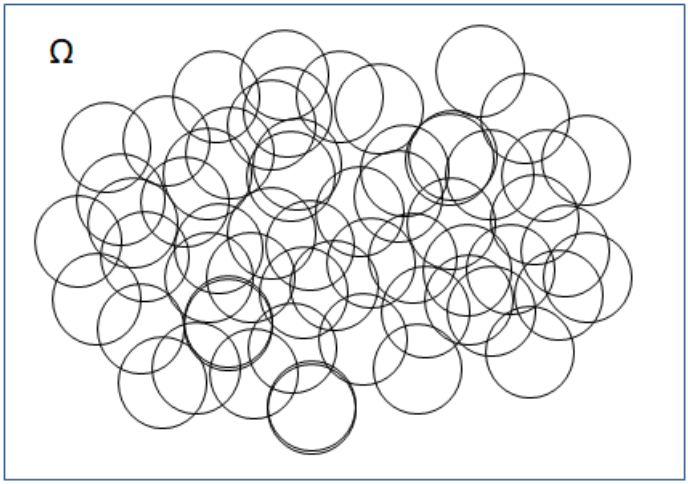

Figure 1. Overlap of observations and variables utilized in analyses in a hypothetical clinical trial.

If we are to in some fashion integrate or accumulate the plausible intervals of benefit and harm over all relevant analyses, how can we conduct this integration with this intense degree of redundancy?

This is where we use our background in set and measure theory. From this perspective, we now know how to create and manage this complex situation.

Specifically, we create a universe containing all of these analyses from a clinical research effort, and call that universe Ω. We then create a σ-algebra of these analyses, Σ, which consists of all of the subsets of Ω.

We know that the sets described in the previous paragraph are members of the σ-algebra because each analysis is a member of Ω and the sets are constructed from our standard set operations of unions, intersections, and complements. In fact, a set (e.g., that of the primary analyses) is but one set of a large collection of analyses that can be assembled and combined in any way that we like.

Of course, there are certainly rules that we will use to direct our attention; we will allow the principals of epidemiology and our intellectual discipline to focus on some sets of analyses, while dismissing others, but, theoretically, each of these analysis sets is available for inspection. However we need a "measure" for each analysis, one that deals sensibly with the redundancy concern.

However, if we can establish such a measure, (our psi measure, $\psi-$ measure), then an entirely different vista opens before us. We will be able not just to measure any analysis, but also measure any set of analyses in Σ. Such an identification permits us, for example, to discount analyses based on their relatively small measure. We can compute the content of the primary analyses in a clinical research effort, or the content of all prospectively declared analyses.

It would allow us to measure "exploratory" analyses, comparing and contrasting their measure to that of

prospective analyses. And the large size of Σ can produce many interesting, heretofore unconsidered analysis sets whose content can be assessed.

Let's assume for a moment that we had a measure μ of each analysis A_i $i = 1, 2, 3, ..., n$ contained in Ω that was based on the number of observations and variables used in each analysis, and we wish to compute $\psi(\Omega) = \psi\left(\bigcup_{i=1}^{n} A_i\right)$. Figure 1 tells us that this would be a complicated operation due to the degree to which the analyses overlap, but we know how to proceed.

In addition, we can follow different paths of analyses to build up this union. For example, we might start in Figure 1's upper right corner, or its lower left corner, or even its center. Out intuition tells us that the solution to $\psi\left(\bigcup_{i=1}^{n} A_i\right)$ will be the same regardless of which path we take, but the contribution of any particular analysis to that union will be path dependent due to the intense overlapping of the analyses. This is an observation of which we will take advantage.

So, with our goal in mind, , and set and measure theoretic background and context in place, we are ready to define an analysis measure, and examine its implications.

Functions and Measures on Analysis Regions

We are now almost ready to develop a quantity that for the moment we will call the content of an analysis.* But before we embark on this, we should review our assumptions.

This reduces to ensuring that we understand the properties of the elements of Ω and Σ. Since we know that any function or measure that we develop based on (Ω, Σ) must be measurable, we should ensure that each element of Ω is imbued with the properties that our functions f and measures μ will inspect and recognize. This is critical because we need to make sure that our assessment of analyses through a function or a measure be measurable, i.e., the elements of Ω have the property that our content can assess.

* Please note that I am staying away from the use of the term information. While that term might be natural and intuitive, I want to avoid any confusion with the well developed science described by the term "statistical information theory", which is related to coding theory, channel theory, and entropy.

In addition, a thorough understanding of Ω also provides a sound foundation for us to develop, manipulate, and ultimately deploy set functions and measures with confidence.

What constitutes an analysis?

A major product of a clinical research effort is its collection of analyses. Let's denote i^{th} analysis of such an effort as ω_i (order does not matter at this point). Then we will define Ω as the congregation of all of these analyses from the study, i.e., the superset $\{\omega_1, \omega_2, \omega_3, ...\omega_n\}$ or $\{\omega_i, i = 1, ..., n\}$.

This we have stated before, but what exactly does this mean? What is an analysis and what are its properties?

An analysis is complicated. While it is easy to think of an analysis as a result (i.e., a hazard ratio), an analysis is a basket of properties associated with a computation. It has many constituent parts.

For example, the analysis has to be designed. What question is the analysis designed to answer? Is the analysis to be prospectively designed or retrospective? On what instrumentation will it be based (e.g., if the analysis was of heart structure, then was that measurement based on magnetic resonance imaging, or echocardiography with contrast?).

The analysis is also based on a specific collection of observations. It uses certain specific variables in the dataset. It utilizes a specific formula classifying it as a type of analysis, e.g., regression analysis, nonparametric analysis, survival analysis, etc.). We can add other characteristics (is the analysis a subgroup analysis?). It generates an estimate of effect. It also produces a standard error and can provide an assessment of bias.

Thus, although it is easy to simply say that each analysis is a member of the space, i.e., $\omega_i \subset \Omega$, there are many characteristics or analysis properties contained in each ω_i. And, each basket of properties is quite rich.

While this may be a new perspective on an analysis, it is not a new concept from our set theoretic perspective. When we assemble, for example a collection of patients in a clinical trial, say, $p_1, p_2, p_3, ...p_n$ and create a space Ω for them, we are collecting individual traits of them (demographics, phenotypes, comorbidity, therapy assignment, compliance, etc.). Each individual is a collector of many different traits of that individual). We are simply applying this familiar concept to the set of analyses.*

This understanding gives us facility in working with the contents of our analysis space Ω. For example, one such property of an analysis is the question that motivated the analysis. Denote this property of ω_i as q_i. Now, this question q_i can be broad, e.g., "what is the effect of the exposure being studies on the overall health of the exposed population as compared to the unexposed population", or narrow, e.g., "what is the effect of a single dose of the exposure on the change in blood pressure over time?"

Since broader questions q can contain more specific questions, it can be anticipated that there will be an interest in focusing on the analysis content over subsets of analyses that address important components of the relevant question. Such a subset can be denoted as $\{\omega_i \,/\, q_i \subset q\}$ or

* In some circles, this basket of items could be considered *metadata*.

$\omega_i \subset A_q = \left\{ \omega_i / q_i \triangleq q \right\}$ which means that the analysis is a member of the set of analyses responsive to question q. Since q_i is a property of the analysis ω_i we can aggregate these analyses, the aggregation being a subset of Ω, resides in Σ. And since they have the properties that we are interested in, they will be measurable and be available for content assessment.

Regions of Analyses

A region of analysis is simply a collection of analyses that has a common characteristic or property.

All of the analyses that provide an answer to a question q constitute a region of analyses. All subgroup analyses is another region. The collection of mixed models assessing gender mediated effects on therapy for systolic blood pressure is another. Our broad definition of ω_i offers a wide latitude in evaluating collections of analyses. Since it contains so many analysis properties, analyses can be assessed using a wide range of measurable functions and measures.

Investigators have broad authority in conducting clinical trial analyses. For example, suppose that the prompting question of the research effort is "Does the provision of allogeneic cells into a subject's heart reverse the progression of heart failure?" The investigator can choose from many candidate variables (e.g., mortality, exercise tolerance, quality of life, and biomarkers). They can also conduct evaluations on different cohorts (e.g., only females, or only patients greater than 60 years of age), as well as implement different but related estimates of the effect of therapy (e.g., nonparametric tests, survival

analyses, general linear model assessments). Each of these is a region of analysis.

So, our purpose will be to accumulate analysis content over these regions of analyses.

But, in addition, we can provide weight to these analyses. The accumulation process over the rich properties of ω_i permit us not to just assess the content of ω_i in multiple dimensions at once. Taking advantage of these analysis features permits the circumstances and characteristics of the analysis to appropriately affect its contribution to the evidence addressing the question at hand. For example, if analysis ω_i is responsive to question q, i.e., $\omega_i \subset A = \{\omega_i \mid q_i \subset q\}$ but is exploratory, and the investigators believe that exploratory analyses are not considered contributory to the answer to question *q*, then that analysis' contribution to answering question *q* can be set to zero. This is easily accomplished by having the Σ measurable function $f(\omega_i)$ that is integrated over the content of ω_i to be zero if ω_i is exploratory. We know $f(\omega_i)$ is measurable since the exploratory characteristic is embedded within ω_i.

As another example, we can simultaneously assess the benefit estimated by ω_i and also the content of ω_i. The benefit we have discussed earlier; we now know that it is a measurable function on (Ω, Σ) because it can be constructed from traits of ω_i.

We can now find that measure, beginning with the concept of the concept of an analysis ω_i.

Defining the content of an analysis

We have discussed how each analysis ω_i is a basket of descriptors that describes and summarizes the analysis. Focusing on two of its items will generate a measure for us.

Let's begin with any one of the analyses $\omega_i \subset \Omega$. The first item we will use is the collection of subjects who contributed data to the analyses. This is not the number of subjects, but the subjects themselves. Each subject's identity is typically denoted as a depersonalized depiction, or ID. It is a unique pointer to a single individual in the analysis.

In order to describe this, we can create a vector or a collection of ID numbers denoting all of the individuals included in the analysis. We will call this collection $\underline{\mathbf{n}}_i$ (the subscript *i* links it to analysis ω_i). We will also denote the number of individuals this represents* as n_i.

* If one thinks of the collection of ID numbers as a vector $\underline{\mathbf{n}}_i$ then n_i is the dimensionality or the tuple of that vector.

We would expect that in two different analyses in the same clinical trial, there would be substantial redundancy in the subjects used for each one. In fact, the collection of individuals for the first would sometimes be the same as the collection for those in the second. In this case, $\underline{\mathbf{n}}_1 = \underline{\mathbf{n}}_2$ and, of course $n_1 = n_2$. Alternatively, if several subjects who took part in analysis 1 did not contribute to analysis 2, then $\underline{\mathbf{n}}_1 \neq \underline{\mathbf{n}}_2$, and $n_1 > n_2$.

We can follow the same procedure for working with the variables that are part of an analysis. Let's define $\underline{\mathbf{v}}_i$ as the collection of variables that are evaluated for the n_i subjects in analysis ω_i. We will also define the total number of variables utilized in analysis ω_i is v_i.

As an example, consider a clinical trial that has randomized 75 subjects to each of a control group or a treatment group. The purpose of the analysis is to address the prospectively asked question, "What is the effect of therapy on the difference in change in mean diastolic blood pressure between the two groups?"

In this circumstance, the vector $\underline{\mathbf{n}}_1$ contains 75 entries, each one being the ID of an individual whose data (variables) were used in the analysis and $n_1 = 75$. Since three variables are involved, (baseline and follow-up DBP, plus the variable denoting the therapy group), then $\underline{\mathbf{v}}_1$ is contains the three variables names (not the data points themselves) and $v_1 = 3$.

The initial content of an analysis

Using what has become our standard definition for ω_i, which is an analysis contained in Σ the σ-algebra of Ω, we define the content of analysis ω_i, as $\psi(\omega_i)$ and write $\psi(\omega_i)$ as

$$\psi(\omega_i) = n_i v_i.$$

The content of an analysis is quite simple; simply the product of the number of participants whose data is included in the analysis and the number of variables that are required for the analysis. Notably, it does not include any of the other components of ω_i e.g., the question that generated the analysis, the design characteristics of the analysis, or the effect size produced by the analysis. Instead, the definition of analysis content is based solely on the data that contributed to the analysis. The unit of ψ is simply subject-variables.

At this point, we are not in a position to claim that $\psi(\omega_i)$ defined as $\psi(\omega_i) = n_i v_i$ is a measure. We will have to examine if it meets all of our four measure criteria. Until we do, we will simply call ψ the content of the analysis.

However, we can explore this concept and appreciate at least some of its implications. As an example, consider a clinical trial in which the analysis being conducted is the comparison of the difference in the change in DBP from baseline to six months between the treatment group and the control group.

Eighty subjects contribute to the analysis and three variables were required (baseline DBP, follow-up DBP and the treatment group identifier). In this case

$\psi(\omega_i) = (80)(3) = 240$. An analysis that evaluates 298 subjects for the effect of therapy on the change in systolic blood pressure (SBP) between two subgroup strata has content (298)(1+3)= 1192.*

From this simple formulation this discussion, we can make the following observations about ψ-content:

1. Every analysis conducted has a content.
2. If the number of variables is fixed, then the larger the set of subjects used in an analysis, the greater the content of the analysis.
3. If the number of participants is fixed, then the greater the number of variables in the analysis, the greater the analysis content.
4. Analysis content is independent of the design features of an analysis.

Since ψ-content is based on simply the number of observations and the number of variables that are included in the analysis, items (1), (2), and (3) are self-evident.

The fourth observation above however, requires attention. The content of an analysis is separate and apart from an assessment determining that analysis' probative value. Analyses of little value (e.g., an analysis conducted in a clinical trial that is irrelevant to the question under consideration) can have high content. Similarly, analyses that have great strategic value may have relatively little

* The 4 variables are 1) baseline SBP, 2) follow-up SBP), 3) therapy assignment, and 4) the one variable on which the data are stratified. Variables e.g., the number of strata divisions, or the function of a variable (e.g., squares or logs) are derivative variables, i.e., their content derives from other variables in the data set.

content. While we will specifically deal with the concept of analysis value in a later chapter, it is clear that other features of the analysis that gauge the analysis' worth must also be integrated.

As developed here, analysis content is separate and apart from analysis contribution. This purpose of ψ - content is to provide a mathematical basis for accumulating overlapping evaluations over their regions of analyses.

However, in order to achieve this goal, we must address the issue of overlapping analyses, the subject of the next chapter.

Analysis redundancy

There is no question but that there is commonly redundancy between different analyses. Thus far, we have not addressed this critical concern.

For example, consider the content of an analysis ω_1 consisting of a general linear model that uses 50 subjects and 2 explainer variables for a dependent variable. From the previous chapter $\psi(\omega_1) = (50)(2) = 100.$

Now consider a second analysis ω_2 which is a general linear model that studies the same 50 subjects and the same 2 explainer variables for the same dependent variable, plus in addition, contains one more explainer variable. We compute $\psi(\omega_2) = (50)(3) = 150.$ Yet, even though the content of the second analysis is greater than that of the first, they both use the same subjects and also have 2 of 3 variables in common. This considerable overlap suggests that the content of the second analyses should be reduced or moderated if the first analysis' contribution has already been considered.

We begin our examination of this by recalling that the content of an analysis ω_i is $\psi(\omega_i) = n_i v_i$ where n_i is the number of subjects and v_i is the number of variables. Note that the content does not depend on the identify of these

subjects, only the number of them. The same is true for the variable component of the analysis' content.

This will not be true for managing the degree to which two analyses overlap.

From our set theory perspective, the caliper of analysis redundancy between two analyses is simply the degree to which two sets (on which those analyses are based) are not disjoint. Thus, the overlap between two analyses is addressed by considering the intersection of these analyses ω_i and ω_j, $\omega_i \cap \omega_j$. Let's denote the content of this intersection as $\psi\left(\omega_i \cap \omega_j\right)$ and define

$$\psi\left(\omega_i \cap \omega_j\right) = n_{ij} v_{ij}.$$

Here, n_{ij} and v_{ij} are the number of subjects and number of variables common to both analyses ω_i and $\omega_{j.}$ Assessing the commonality requires us to focus on not just the number of subjects and variables used in the two analyses, but the degree to which they are the same.

For example, consider two analyses in a clinical trial that has randomized 305 subjects. The first analysis incorporates 298 of these subjects and utilizes 5 variables. The second analysis incorporates 245 of these 298 subjects and utilizes 8 variables, 3 of which are common to the first analysis. Then we may compute

$$\psi\left(\omega_1\right) = n_1 v_1 = \left(298\right)\left(5\right) = 1490$$
$$\psi\left(\omega_2\right) = n_2 v_2 = \left(245\right)\left(8\right) = 1960$$
$$\psi\left(\omega_1 \cap \omega_2\right) = n_{12} v_{12} = \left(245\right)\left(3\right) = 735.$$

We define in general, the content of the intersection of *k* analyses $\omega_1, \omega_2, \omega_3, ...\omega_k$ is

$$\psi\left(\bigcap_{i=1}^{k}\omega_i\right) = n_{i...k}v_{i...k}$$

where $n_{i...k}$ is the number of observations common to all *k* analyses, and $v_{i...k}$ is the number of variables common to all *k* analyses.

Computing the content of analysis unions

The ability to calculate the content of the intersection of analyses is precisely what we need to compute the content of analyses' unions which is our goal. For example, from out previous measure theory development, we know that we can write the content of the union of analyses ω_i and ω_j as $\psi\left(\omega_i \cup \omega_j\right) = \psi\left(\omega_i\right) + \psi\left(\omega_j\right) - \psi\left(\omega_i \cap \omega_j\right)$. In this situation of clinical research program, we can write that the content of the union of two analyses as $\psi\left(\omega_i \cup \omega_j\right) = n_i v_i + n_j v_j - n_{ij} v_{ij}$. In the above example,

$$\begin{aligned}\psi\left(\omega_i \cup \omega_j\right) &= \psi\left(\omega_i\right) + \psi\left(\omega_j\right) - \psi\left(\omega_i \cap \omega_j\right)\\ &= n_i v_i + n_j v_j - n_{ij} v_{ij}\\ &= 1490 + 1960 - 735\\ &= 2715.\end{aligned}$$

However, this relatively simple formulation has important consequences that we must now consider, understand, and accept if we are to work with it.

1. *Analyses have no common content if they have no common subjects.* If there are no common subjects between two analyses, ω_i and ω_j, then $n_{ij} = 0,$ the analyses are disjoint, and the intersection of the two analyses has zero content. This makes intuitive sense because in clinical research, subjects are assumed to operate independently of one another, making separate contributions to analyses. Thus, separate and disjoint collections of participants produce separate content for each analysis, but with no joint contribution there is no joint content. Thus the analyses are disjoint and the content of their insection is zero.

2. *If two analyses have no common variables, although* $v_{ij} = 0$ *the content of the intersection* $\psi\left(\omega_i \cap \omega_j\right)$ *will frequently be, but need not be zero.* At first blush, in the case where $v_{ij} = 0,$ using the reasoning of the previous observation just stipulated, it follows that the content of the intersection of the two analyses should be zero.

   However, there is an exception that is well recognized in clinical research that we must incorportate. Variables – unlike participants – can be and commonly are interrelated, and the degree to which they are related to each other impacts the commonality of the analyses under consideration. For a fixed number of participants, if the collection of variables in analysis is related to the collection of variables in analysis 2, then even though the variables are not explicitly common, their interrelatedness would convey a nondisjoint

interaction. This complicating situation is explicitly managed and incorporated in Chapter 22. At this early point in the development, we will simply say, that, in the absent of interrelationships between the variables, the absent of common variables between two analyses implies that the content of the interaction of the two analyses is zero.

3. *The content of the union of analyses is the sum of the analyses' content when the analyses do not overlap.* This is not unexpected at all, and simply follows from our application of set and measure theory to the circumstance of accumulating the measure of disjoint sets.

4. *Analyses with substantial content can have different purposes and value.* This is a critical observation. The input to the content of two analyses' joint content is strictly mechanical. Therefore, two analyses can have substantial intersection content although from an intellectual or design perspective, they have different motivations and come to conclusions in different noetic dimensions.

   Consider, for example, a clinical trial provides two analyses each involving the same 100 subjects. The first is an assesment of the impact of the randomly allocated therapy on mortality, the primary analysis of the study. The second is an exploratory asssement of the effect of therapy on a newly discovered cell phenotype using those same 100 subjects. Because the analyses use the same 100 subjects, as well as the treatment assignment variable, there is considerable content redundancy.

However, these two analyses have very different purposes and bring different value levels to the overall investigation. Thus, while there may be substantial redundancy in analysis content, the analyses may nevertheless play very different roles in the research enterprise.

Similarly, a large value of $\psi\left(\omega_i \cap \omega_j\right)$ does not suggest that the two analyses ω_i and ω_j address the same research question, only that they have common substrate (i.e., each draws from the same participants and the same variables.)

5. *An analysis' content can be separate and apart from that analysis' value.* This again speaks to the difference between intellectual value and ψ-content. A survival model studying 453 subjects using three variables (time to event, censoring mechanism, and therapy assignment) has less content than a mixed model regression analysis on the same subjects that incorportaes 5 different covariates. However the importance of the mortality evaluation is greater than the regression analysis if the effect of therapy on the death rate in the study was the most important question. The content of an analysis is simply based on the number of observations and number of variables used in the evaluation, and is quite rote. However, an analysis' value depends on other properties of ω_i, $\delta_i(j), j = 1,2,3...$ e.g., the interrogation that the analysis addresses and whether that analysis was prospectively declared. The content of an analysis is separate and apart from its purpose and value.

Given observations (4) and (5) above, why should we

bother developing a mechanical content which has no cerebral input, and is devoid of independent of critical considerations of epidemiology, research design issues, or the investigator determined priority of analyses. Aren't these latter concerns the really interesting and necessary metrics? Why not build a construct based on them?

The answer is yes, those advanced concepts are essential to drawing conclusions from research efforts. However, at this point, we are not yet ready to mathematically include them (this occurs in chapter 24.

We are constructing cornerstones now. Like any foundation, it must be objectively assessable, reproducible, and durable. While intellectual assessments of acceptable analyses change over time* our foundation must be fixed. The "mechanical" ψ-content serves handsomely in this regard since it is incontrovertible that observations and variables are required for analyses.

There are of course other metrics (quality of research design, prospective declaration, etc.) that must and will play an essential role in assessing the quality and value of the analysis, but they will come later.† The absence of their contribution at this point is why we use the neutral term "content" to describe ψ.

But, the question now before us, is whether ψ-content is really a measure at all. This is the topic of the next chapter.

---

* For example, in the 1970's, subgroup evaluations and exploratory analyses were as admissable as primary outcome assessments because they each derived from clinical trials.

† In measure-theoretic language, ψ will ultimately be the measure, while the important epidemiologic and intellectural contributions will be measurable functions that operate on the (Ω, Σ) measure space.

A Final Ingredient: Addimg Variable Interrelationship to the Analysis Content

While the concept of content of an analysis $\psi(\omega_i) = n_i v_i$ where n_i is the number of observations and v_i is the number of variables in analysis ω_i. has some merit, one clear deficiency is the absence of any measure of the relationship between the variables in an analysis. A single analysis may involve two variables, or it may involve more than twenty variables. The measure of the analysis should to some degree include the interrelationships between the variables which our current evaluation does not consider

This chapter focuses on the incorporation of the variable relationships into $\psi -$ content.

We begin with the observation that while linear correlation does not reflect the universe of relationships between variables, it represents the overwhelming majority of them. There are curvilinear relationships in biologic and pathophysiologic processes in health care. However, a

heavy component of these curvilinear relationships is linear. Thus, focusing on the correlationship matrix $\mathbf{R}$ allows us to focus on the majority of the information about interrelationships that we will use.

We will refer to the correlation matrix of a particular analysis ω_i as $\mathbf{R}(\omega_i)$.

Note, for the time being we will assume that each varaible used that is part of the analysis ω_i is incorporated into $\mathbf{R}(\omega_i)$ Now lets redefine our content as

$$\psi(\omega_i) = n_i v_i \left|\mathbf{R}_i\right|^{-1}$$

where $\left|\mathbf{R}_i\right|^{-1}$ is the reciprocal of the determinant of the correlation matrix $\mathbf{R}_i$.

Also, define the measure of the intersection of two analyses ω_i and ω_j,

$$\psi(\omega_i \cap \omega_j) = n_{ij} v_{ij} \left|\mathbf{R}_{ij}\right|^{-1}.$$

whereas before n_{ij} is the number of observations that contain all variables common to the i^{th} and j^{th} analyses, v_{ij} is the number of variables common to both analyses i and j and $\mathbf{R}_{ij}$ is the correlation matrix of only the variables common to ω_i and ω_j, and $\left|\mathbf{R}\right|$ is the determinant of this correlation matrix.

Determinants

Determinants are the reduction of a correlation matrix to a single number. They reflect the degree of the unexplained variability in the system. The maximum value of the

determinant of a positive definite correlation matrix is 1. This occurs when all correlations are zero, which we can interpret as no explained variability; all variability is unexplained.

The minimum value of a correlation matrix's determinant is zero. In this extreme case there is so much variability in the system that at least one of the variables is redundant (one of the variables can be completely recapitulated by a linear combination of the others, a condition known as linear dependency.

We will assume throughout this book that $0 < |\mathbf{R}| \leq 1$.

The greater the dimension of a correlation matrix, in general, the more complicated is its determinant to compute.

Consider the simple example where $\mathbf{R}_2 = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$, then $|\mathbf{R}_2| = 1 - r^2$. The determinant is maximized when $r = 0$. That is when the variables are uncorrelated (no explained variability; all unexplained variability) then $|\mathbf{R}_2|$ is at its greatest. As r^2 increases, the relationship between the two variables becomes tighter, explained variability increases, unexplained variaibility decreases and the determinant decreases.

So $|\mathbf{R}_2|$ is a marker of "variable uncorrelatedness" of the two variables. The larger $|\mathbf{R}_2|$ the greater the degree of uncorrelatedness. Therefore $|\mathbf{R}|^{-1}$ is a marker or correlation or dependency. The greater the correlation, the greater the explained variability, the larger the determinant's reciprocal.

Thus, defining $\psi(\omega_i) = n_i v_i |\mathbf{R}_i|^{-1}$ is the creation of a analysis content that increases with increased dependence between variables. The stronger the correlation structure, the greater the measure of the analysis.

Another way to think of this is that the determinant of the variables in an analysis is large when the explained explained variability is low, and that the content of an analysis is large when the unexplained variabilty is low or the explained variability is high.

This is not an uncommon finding in probability and statsitics. For example, the paired *t*-test uses the correlation between two variables to decrease the variance of the estimate of the mean difference between the two and thereby increase the power of the test *ceteris paribus*.Its power is greater with greater dependency. Multivarite testing e.g., Mahalonobis distance, Hotellings $-T^2$, discriminant function analysis, and multivariate analysis of variance all take advantage of the presence of correlation among the system of variables.

These are common findings in the application of multivariate distribution analysis to clinical research. In this case we have developed a concept of analysis content whose content increases with the intensity of correlation between the variables.

There is one more observation that we need to make about determinants before we proceed with the proof that $\psi-$ content is actually $\psi-$ measure.

Containment Property

Let $\mathbf{R}_p$ be the correlation matrix of p variables and $\mathbf{R}_q$ be the correlation matrix of q variables under the condition

that the q variables of $\mathbf{R}_q$ are actually a subset of the initial p variables. Then $\left|\mathbf{R}_q\right| \geq \left|\mathbf{R}_p\right|$.

If you have a collection of variables and take a subset of them, then the determinant of the correlation matrix of the subset is greater than or equal to the determinant of the correlation of the initial variables.

As an example, consider that there are two variables with the correlation matrix $\mathbf{R}_2 = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$, and a third variable is added such that the correlation matrix for all three variables is $\mathbf{R}_3 = \begin{pmatrix} 1 & r & r_2 \\ r & 1 & r_3 \\ r_2 & r_3 & 1 \end{pmatrix}$. How does the containment property work here?

If we take a simple case, letting $r_2 = r_3 = 0,$ then $\mathbf{R}_3 = \begin{pmatrix} 1 & r & 0 \\ r & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and the $\left|\mathbf{R}_3\right| = 1 - r^2.$ This is the maximum value that $\left|\mathbf{R}_3\right|$ can have. In this case $\left|\mathbf{R}_3\right| = \left|\mathbf{R}_2\right|.$ The only explained variability in this system is conveyed by r.

Now, moving away from this simplified condition, note that nonzero values of r_2 and or r_3 increase the explained variability in the system. Since the determinant is greater when there is more unexplained variability then it follows that $\left|\mathbf{R}_3\right| \leq \left|\mathbf{R}_2\right|$.

We can show this rather inelegantly by brute force algebra.

$$\left|\mathbf{R}_3\right| = \left|\begin{pmatrix} 1 & r & r_2 \\ r & 1 & r_3 \\ r_2 & r_3 & 1 \end{pmatrix}\right| = r_2\left(rr_3 - r_2\right) - r_3\left(r_3 - rr_2\right) + 1\left(1 - r^2\right).$$

Since $\left|\mathbf{R}_2\right| = 1 - r^2$, then we must show $r_2\left(rr_3 - r_2\right) - r_3\left(r_3 - rr_2\right) \leq 0$. Proceeding

$$r_2\left(rr_3 - r_2\right) - r_3\left(r_3 - rr_2\right) \leq 0$$
$$r_3\left(r_3 - rr_2\right) - r_2\left(rr_3 - r_2\right) \geq 0$$
$$r_3^2 - rr_2r_3 - rr_2r_3 + r_2^2 \geq 0$$
$$r_3^2 + r_2^2 - 2rr_2r_3 \geq 0$$
$$r_3^2 + r_2^2 \geq 2rr_2r_3$$
$$\frac{r_3^2 + r_2^2}{2r_2r_3} \geq r$$
$$\frac{1}{2}\left(\frac{r_3}{r_2} + \frac{r_2}{r_3}\right) \geq r.$$

Let $x = \frac{r_2}{r_3}$. Letting the correlations be positive for now $0 < x > \infty$, (0 correlation is quickly handled in a special case where

$$\left|\mathbf{R}_3\right| = \left|\begin{pmatrix} 1 & r & 0 \\ r & 1 & r_3 \\ 0 & r_3 & 1 \end{pmatrix}\right| = -r_3\left(r_3 - 0\right) + 1 - r^2 = 1 - r^2 - r_3^2 \leq 1 - r^2 = \left|\mathbf{R}_2\right|$$

Then

$$\frac{1}{2}\left(\frac{r_3}{r_2}+\frac{r_2}{r_3}\right)=\frac{1}{2}\left(x+\frac{1}{x}\right)=\frac{x^2+1}{2x} \text{ where } x=\frac{r_3}{r_2}.$$

So what are the values of r_2 and r_3 such that $\frac{x^2+1}{2x} > r.$

The minimum value of this function is at $x=1$ where $\frac{x^2+1}{2x}=1$, the maximum value of r. Thus $\frac{1}{2}\left(\frac{r_3}{r_2}+\frac{r_2}{r_3}\right) \geq r$ and $\left|\mathbf{R}_3\right| = \left|\begin{pmatrix} 1 & r & r_2 \\ r & 1 & r_3 \\ r_2 & r_3 & 1 \end{pmatrix}\right| \leq \left|\mathbf{R}_2\right|.$

With this definition of content as $\psi\left(\omega_i\right)=n_i v_i \left|\mathbf{R}_i\right|^{-1}$ and the determinant containment property, we can proceed to demontrating that $\psi-$ content is $\psi-$ measure.

Converting ψ-content to ψ-measure.

This chapter, by necessity, is a fairly technical. For those who want to be convinced that ψ-content is indeed a measure, please read through the material here in detail. However, if you would are comfortable assuming for the moment that this mathematical development is correct, then feel free to skip to the chapter summary.

To recap, we have developed a set function that converts the element of a clinical trial analysis into something we have defined as content. We wanted this content to key off of what was contained in an analysis ω_i to ensure that our content was measurable, but having done that, we had tremendous freedom in the selection of the elements of the analysis ω_i that would be incorporated in our content development. We chose $\psi(\omega_i) = n_i v_i \left|\mathbf{R}_i\right|^{-1}$ and $\psi(\omega_i \cap \omega_j) = n_{ij} v_{ij} \left|\mathbf{R}_{ij}\right|^{-1}$. This was not the only choice of a definition for content (other possible selections are

discussed in Chapter 23) but it is a simple one, and we have explored this structure's properties and weaknesses.

Our next task is to demonstrate that ψ is a measure. Since this is an important step, we will need to develop these arguments formally.

Let Ω as the congregation of all analyses conducted in the clinical trial, i.e., the superset $\{\omega_1, \omega_2, \omega_3, ...\omega_n\}$ or $\{\omega_i, i = 1, ..., n\}$.. Denote Σ its σ-algebra. Remember that the four assertions required to that demonstrate ψ is a measure are:

Assertion 1: For $\omega_i \subset \Sigma$, $\psi(\omega_i) \geq 0$.

Assertion 2: $\psi(\varnothing) = 0$.

Assertion 3: If $\omega_i \subset \omega_j$, **then** $\psi(\omega_i) \leq \psi(\omega_j)$.

Assertion 4: $\psi\left(\bigcup_{i=1}^{n} \omega_i\right) \leq \sum_{i=1}^{n} \psi(\omega_j)$.

We will see that three of these assertions are quite easy to prove, while the fourth is available to us if we apply what we know of set theory diligently and delicately.

Assertion 1: For $\omega_i \subset \Sigma$, $\psi(\omega_i) \geq 0$.

This is quite straightforward. Since any $\omega_i \subset \Sigma$ by definition is an executed analysis, it must be based on a positive number of participants and a positive number of variables. Thus, n_i and v_i are ≥ 0. We assume that $|\mathbf{R}_i|$ is of full rank and therefore ≥ 0. Therefore $\psi(\omega_i) = n_i v_i |\mathbf{R}_i|^{-1} \geq 0$.

Assertion 2: $\psi(\varnothing) = 0.$

If $\omega_i = \varnothing,$ the i^{th} analysis has not been conducted. Thus the contents of the analysis are vacant. Therefore, each of n_i and v_i are equal to 0 and neither $\mathbf{R}$ nor $|\mathbf{R}_i|^{-1}$ does not exist. However we choose to manage a correlation matrix with no variables, since both n_i and v_i are equal to 0, then $\psi(\omega_i) = n_i v_i |\mathbf{R}_i|^{-1} = 0.$

Assertion 3: If $\omega_i \subset \omega_j,$ **then** $\psi(\omega_i) \le \psi(\omega_j).$

This assertion requires the development of the idea of one analysis "containing" another analysis $(\omega_i \subset \omega_j)$, a relatively unexplored concept in clinical trial methodology.

Let's begin with the notion that an analysis ω_i is contained within analysis ω_j when 1) the observations in analysis *i* is the same or is a subset of those used in analysis *j*, i.e., $\underline{\mathbf{n}}_i \subseteq \underline{\mathbf{n}}_j,$ and 2) the variables used in analysis *i* are the same or a subset of the variables used in analysis *j*, i.e., $\underline{\mathbf{v}}_i \subseteq \underline{\mathbf{v}}_j.$ With this definition, then $\omega_i \subset \omega_j$ implies that $n_i \le n_j$ and $v_i \le v_j.$ Also, the determinant containment property ensures that if $\underline{\mathbf{v}}_i \subseteq \underline{\mathbf{v}}_j.$ then $|\mathbf{R}_i| \ge |\mathbf{R}_j|,$ or $|\mathbf{R}_i|^{-1} \le |\mathbf{R}_j|^{-1}.$ Thus $n_i v_i |\mathbf{R}_i|^{-1} \le n_j v_j |\mathbf{R}_j|^{-1},$ and $\psi(\omega_i) \le \psi(\omega_j).$

However, this practical definition has consequences. One important ramification is that even though $\omega_i \subset \omega_j$, the two analyses can be completely unrelated to each other.

For example, an analysis conducted to assess whether the change in blood pressure over time is impacted by a change over time in quality of life can be assessed using a general linear model.

However, the analysis of whether the change in quality of life over time is related to the change in blood pressure over time is also addressed by a general linear model with the same participants and variables. Both analyses utilize the same number of observations and variables and therefore each contains the other; however, the questions motivating the analyses are different. This is another example of how analysis intent is separate and apart from analysis content and must therefore be considered independently (as we will).

In addition, $\psi(\omega_i) \le \psi(\omega_j)$ does not imply that $\omega_i \subset \omega_j$. Consider the situation where analysis ω_1 consists of 100 observations and 10 variables. Its content will be greater than that of an analysis ω_2 from the same trial that utilizes 75 observations and 5 variables. However, if the 10 variables utilized in analysis ω_1 are only demographic variables, while analysis ω_2 utilizes imaging variables only, the analyses do not contain each other even though $\psi(\omega_1) \le \psi(\omega_2)$.

Assertion 4: Let $\omega_i, i = 1,,,n$ be a collection of analyses contained in Σ. Then $\psi\left(\bigcup_{i=1}^{n}\omega_i\right) \le \sum_{i=1}^{n}\psi\left(\omega_j\right)$.

This assertion will be demonstrated for two mutually exclusive and exhaustive cases.

Case 1: $\{\omega_1, \omega_2, \omega_3, ..., \omega_n, ...\}$ are pairwise disjoint. This demonstration is quite straightforward. The disjoint assumption permits $\psi\left(\omega_1 \cup \omega_2\right)$ to be written as $n_1v_1 + n_2v_2 = \psi\left(\omega_1\right) + \psi\left(\omega_2\right)$ since $\psi\left(\omega_1 \cap \omega_2\right) = 0$. In order to demonstrate assertion 4 in this circumstance, we simply need to add additional disjoint analyses one at a time. Adding each additional analysis ω_i into the union only adds the terms n_iv_i to the content.

This can be demonstrated through induction. Assume that $\psi\left(\bigcup_{i=1}^{k}\omega_i\right) = \sum_{i=1}^{k}\psi\left(\omega_i\right)$. Then develop $\psi\left(\bigcup_{i=1}^{k+1}\omega_i\right)$ as

$$\psi\left(\bigcup_{i=1}^{k+1}\omega_i\right) = \psi\left(\left(\bigcup_{i=1}^{k}\omega_i\right) \cup \omega_{k+1}\right) = \psi\left(\bigcup_{i=1}^{k}\omega_i\right) + \psi\left(\omega_{k+1}\right) - \psi\left(\left(\bigcup_{i=1}^{k}\omega_i\right) \cap \omega_{k+1}\right)$$

$$= \sum_{i=1}^{k+1}\psi\left(\omega_i\right) - \sum_{i=1}^{k}\psi\left(\omega_i \cap \omega_{k+1}\right) = \sum_{i=1}^{k+1}\psi\left(\omega_i\right) - 0 = \sum_{i=1}^{k+1}\psi\left(\omega_i\right) = \sum_{i=1}^{k+1}n_iv_i.$$

and the assertion is demonstrated for pairwise disjoint analyses.

Case 2: $\{\omega_1, \omega_2, \omega_3, ..., \omega_n, ...\}$ are not pairwise disjoint.

What made Case 1 so easy was the disjoint nature of the analyses under consideration. We will apply a similar approach to Case 2, but we will need to essentially convert nondisjoint sets to disjoint sets so that the union of the two is the same. Then, we can operate on the union of the disjoint sets. This set methodology will be a common motif in our subsequent development of quanta analysis.

Let's first look at the simplist example to gain some intuition. We need to show in the case of only two analyses, ω_1 and ω_2 that $\psi(\omega_1 \cup \omega_2) \leq \psi(\omega_1) + \psi(\omega_2)$.

We know

$$\begin{aligned}\psi(\omega_1 \cup \omega_2) &= \psi(\omega_1) + \psi(\omega_2) - \psi(\omega_1 \cap \omega_2) \\ &= n_1 v_1 |\mathbf{R}_1|^{-1} + n_2 v_2 |\mathbf{R}_2|^{-1} - n_{12} v_{12} |\mathbf{R}_{12}|^{-1}\end{aligned}$$

Now, let's bound $n_{12} v_{12} |\mathbf{R}_{12}|^{-1}$. The maximum value of $n_{12} v_{12} |\mathbf{R}_{12}|^{-1} = n_2 v_2 |\mathbf{R}_2|^{-1}$ since the maximum number of observations in both analyses ω_1 and ω_2 is n_2, the maximum number of variables is $v_{2,}$ which makes $|\mathbf{R}_{12}| = |\mathbf{R}_2|$. Similarly, the smallest value of $n_{12} v_{12} |\mathbf{R}_{12}|^{-1}$ is when there are no intersecting observations and/or varibles and $n_{12} v_{12} |\mathbf{R}_{12}|^{-1} = 0$ by assertion 2. Thus.

$$0 \leq n_{12} v_{12} |\mathbf{R}_{12}|^{-1} \leq n_2 v_2 |\mathbf{R}_2|^{-1}$$

or

$$0 \leq n_2 v_2 |\mathbf{R}_2|^{-1} - n_{12} v_{12} |\mathbf{R}_{12}|^{-1} \leq n_2 v_2 |\mathbf{R}_2|^{-1}.$$

Thus

$$\begin{aligned}\psi(\omega_1 \cup \omega_2) &= \psi(\omega_1) + \psi(\omega_2) - \psi(\omega_1 \cap \omega_2) \\ &= n_1 v_1 \left|\mathbf{R}_1\right|^{-1} + n_2 v_2 \left|\mathbf{R}_2\right|^{-1} - n_{12} v_{12} \left|\mathbf{R}_{12}\right|^{-1} \\ &\leq n_1 v_1 \left|\mathbf{R}_1\right|^{-1} + n_2 v_2 \left|\mathbf{R}_2\right|^{-1} = \psi(\omega_1) + \psi(\omega_2).\end{aligned}$$

Now lets examine one last specific case for $n = 3$.

In this circumstance we focus on $\psi(\omega_1 \cup \omega_2 \cup \omega_3)$. We write

$$\psi(\omega_1 \cup \omega_2 \cup \omega_3) = \psi(\omega_1 \cup \omega_2) + \psi(\omega_3) - \psi((\omega_1 \cup \omega_2) \cap \omega_3)$$
$$\leq (\psi(\omega_1) + \psi(\omega_2)) + \psi(\omega_3) - \psi((\omega_1 \cup \omega_2) \cap \omega_3).$$

Now we could brute force the set theory computations for $\psi((\omega_1 \cup \omega_2) \cap \omega_3)$ or we could just assess it, similarly to what we did for $n = 2$.

No matter how many observations and variables that there are in $\omega_1 \cup \omega_2$ there are no more than n_3 observations and v_3 variables in $(\omega_1 \cup \omega_2) \cap \omega_3$. If we call $\widetilde{\omega} = (\omega_1 \cup \omega_2) \cap \omega_3$, and $\left|\mathbf{R}(\widetilde{\omega})\right| \geq \left|\mathbf{R}_3\right|$. Thus minimun value for $\psi((\omega_1 \cup \omega_2) \cap \omega_3) = 0$ and its maximum is

$$n_3 v_3 \left|\mathbf{R}(\widetilde{\omega})\right|^{-1} \leq n_3 v_3 \left|\mathbf{R}_3\right|^{-1}$$

Thus $\psi(\omega_3) - \psi((\omega_1 \cup \omega_2) \cap \omega_3)$ is trapped between zero and $\psi(\omega_3)$ and $\psi\left(\bigcup_{n=1}^{3} \omega_i\right) \leq \sum_{i=1}^{3} \psi(\omega_i)$ proving the assertion for $n = 3$.

We now have the basis for an induction proof.

Given that for $k = 1, \psi\left(\bigcup_{k=1}^{1} \omega_i\right) \leq \psi(\omega_i)$. and for $k > 1,\ \psi\left(\bigcup_{i=1}^{k} \omega_i\right) \leq \sum_{i=1}^{k} \psi(\omega_i)$, then prove $\psi\left(\bigcup_{i=1}^{k+1} \omega_i\right) \leq \sum_{i=1}^{k+1} \psi(\omega_i)$.

We follow the structure that we have developed.

$$\psi\left(\bigcup_{i=1}^{k+1} \omega_i\right) = \psi\left(\bigcup_{i=1}^{k} \omega_i \bigcup \omega_{k+1}\right)$$

$$= \psi\left(\bigcup_{i=1}^{k} \omega_i\right) + \psi(\omega_{k+1}) - \psi\left(\bigcup_{i=1}^{k} \omega_i \bigcap \omega_{k+1}\right)$$

$$\leq \sum_{i=1}^{k} \psi(\omega_i) + \psi(\omega_{k+1}) - \psi\left(\bigcup_{i=1}^{k} \omega_i \bigcap \omega_{k+1}\right)$$

And our attention turns to managing $\psi\left(\bigcup_{i=1}^{k} \omega_i \bigcap \omega_{k+1}\right) = \psi(\tilde{\omega})$.

We know $\psi\left(\tilde{\omega}\right) = \tilde{n}\,\tilde{v}\,\mathbf{R}^{-1}\left(\tilde{\omega}\right)$. We know that $0 \le \tilde{n} \le n_{k+1}, 0 \le \tilde{v} \le v_{k+1},$ and $\left|\mathbf{R}\left(\tilde{\omega}\right)\right| \ge \left|\mathbf{R}_{k+1}\right|$, implying that $\left|\mathbf{R}\left(\tilde{\omega}\right)\right|^{-1} \le \left|\mathbf{R}_{k+1}\right|^{-1}$. Therefore

$$0 \le \psi\left(\bigcup_{i=1}^{k}\omega_i \bigcap \omega_{k+1}\right) \le n_{k+1}v_{k+1}\left|\mathbf{R}_{k+1}\right|^{-1}$$

And $0 \le \psi\left(\omega_{k+1}\right) - \psi\left(\bigcup_{i=1}^{k}\omega_i \bigcap \omega_{k+1}\right) \le n_{k+1}v_{k+1}\left|\mathbf{R}_{k+1}\right|^{-1}$

Therefore $\psi\left(\bigcup_{i=1}^{k+1}\omega_i\right) \le \sum_{i=1}^{k+1}\psi\left(\omega_i\right)$ and by induction

$$\psi\left(\bigcup_{i=1}^{\infty}\omega_i\right) \le \sum_{i=1}^{\infty}\psi\left(\omega_i\right).$$

Chapter Summary

We have spent considerable time developing the motivation for and exploring the limitations of ψ-content. While ψ is certainly a measurable function on the analysis space $\left(\Omega, \Sigma\right)$, we need more than that from it. We require the ability to compute the content of a region of analyses in a way that makes some intuitive sense.

What defines "sense" here is the four properties of measure. Thus $\psi\left(\omega_i\right) = n_i v_i$ must not just be a measurable function, but a measure. Having satisfied the four properties of measure, we are assured that this is the case, and can proceed with computing the measure of analysis regions with confidence, since the properties of a measure

are the characteristics that we need to gain an intuitive sense of analysis regions.

Measuring analysis sets (quanta analysis)

The ultimate goal of this overall development is to integrate over a collection of regions A_q a measurable function $f(\omega_i)$ with respect to ψ-measure, $\int_{A_q} f(\omega_i)\, d\psi$, a procedure that will manage the redundancy in the observations and variables in the different analyses that comprise A_q. If for example the investigators wish to assess the totality of evidence in a clinical trial that addressed question *q*, "What is the beneficial effect of therapy in Asian women?", they would deploy an especially derived measurable function $f(\omega_i)$ that accrued benefit and then compute $\int_{A_q} f(\omega_i)\, d\psi$ where $A_q = \left\{\omega_i / q_i \triangleq q\right\}$. However, this integral is difficult to compute directly because the analyses are in general not disjoint, and therefore the veracity of

$\int_A f(\omega_i)d\psi = \sum_{\omega_i \subset A} f(\omega_i)\psi(\omega_i)$ cannot be assumed.*

However, from the statement that $\bigcup_i \omega_i = A_q$ (a statement about the analysis region), and 1) our previous construction of the collection of increasing sets $\{C_i\}$ such that $C_i = \bigcup_{j=1}^{i} \omega_i$ and 2) the collection of disjoint sets $\{B_i\}$ such that $B_i = C_i \cup C_{i-1}^c$, and $\bigcup_i B_i = A_q$, then the $\int_{A_q} f(\omega_i)d\psi$ can be computed exactly as $\int_{A_q} f(\omega_i)d\psi = \int_{\bigcup_i B_i} f(\omega_i)d\psi = \sum_{B_i \subset A} f(\omega_i)\psi(B_i)$. The construction of the disjoint sets $\{B_i\}$ is what converts the integral into a summation.

The sets $\{B_i\}$ are what we call the fractions or quanta of analyses. It only remains to compute the measure of each quanta, $\psi(B_i)$. This is the topic of this chapter.

Computing Quanta Sums

Let's begin with B_1. Begin by writing

$\psi(B_1) = \psi(\omega_1) = n_1 v_1$ as we saw in Chapter 12.

For $i = 2$, we write

---

* In fact we know that $\int_{A_q} d\psi = \psi\left(\bigcup_{\omega_i \subset A_q} \omega_i\right) \leq \sum_{\omega_i \subset A_q} \omega_i$ by the fourth property of measure theory, discussed in Chapter 14.

$$\begin{aligned}\psi(B_2) &= \psi(C_2) - \psi(C_1) \\ &= \psi(\omega_i \cup \omega_2) - \psi(\omega_1) \\ &= n_1 v_1 + n_2 v_2 - n_{12} v_{12} - n_1 v_1 \\ &= n_2 v_2 - n_{12} v_{12}.\end{aligned}$$

This finding was a basis of the observation that $\psi(B_2) \le \psi(\omega_2)$, and $\psi(\omega_1 \cup \omega_2) \le \psi(\omega_1) + \psi(\omega_2)$.

For $i = 3$,

$$\begin{aligned}\psi(B_3) &= \psi(C_3) - \psi(C_2) = \psi(\omega_i \cup \omega_2 \cup \omega_3) - \psi(\omega_i \cup \omega_2) \\ &= \psi(\omega_i \cup \omega_2) + \psi(\omega_3) - \psi((\omega_i \cup \omega_2) \cap \omega_3) - \psi(\omega_i \cup \omega_2) \\ &= \psi(\omega_3) - \psi((\omega_i \cup \omega_2) \cap \omega_3).\end{aligned}$$

Continuing, since

$$\begin{aligned}\psi((\omega_1 \cup \omega_2) \cap \omega_3) &= \psi((\omega_i \cap \omega_3) \cup (\omega_2 \cap \omega_3)) \\ &= \psi(\omega_i \cap \omega_3) + \psi(\omega_2 \cap \omega_3) - \psi(\omega_1 \cap \omega_2 \cap \omega_3) \\ &= n_{13} v_{13} + n_{23} v_{23} - n_{123} v_{123}\end{aligned}$$

then

$$\begin{aligned}\psi(B_3) &= \psi(\omega_3) - \psi((\omega_i \cup \omega_2) \cap \omega_3) \\ &= n_3 v_3 - n_{13} v_{13} - n_{23} v_{23} + n_{123} v_{123}.\end{aligned}$$

We see a pattern beginning to emerge. For B_3 we start with $\psi(\omega_3)$, then subtract off the measure of the dual interactions that involve ω_3, then add the triple interaction.

Continuing for $\psi(B_4)$ (the measure of the union of the analyses $\omega_1, \omega_2, \omega_3, \omega_4$ after removing the union of analyses $\omega_1, \omega_2, \omega_3,$) proceed as

$$\begin{aligned}\psi(B_4) &= \psi(C_4) - \psi(C_3) = \psi(\omega_i \cup \omega_2 \cup \omega_3 \cup \omega_4) - \psi(\omega_i \cup \omega_2 \cup \omega_3) \\ &= \psi(\omega_i \cup \omega_2 \cup \omega_3) + \psi(\omega_4) - \psi((\omega_i \cup \omega_2 \cup \omega_3) \cap \omega_4) \\ &- \psi(\omega_i \cup \omega_2 \cup \omega_3) \\ &= \psi(\omega_4) - \psi((\omega_i \cup \omega_2 \cup \omega_3) \cap \omega_4) \\ &= \psi(\omega_4) - \psi(\omega_1 \cap \omega_4) - \psi(\omega_2 \cap \omega_4) - \psi(\omega_3 \cap \omega_4) \\ &+ \psi(\omega_1 \cap \omega_2 \cap \omega_4) + \psi(\omega_1 \cap \omega_3 \cap \omega_4) + \psi(\omega_2 \cap \omega_3 \cap \omega_4) \\ &- \psi(\omega_1 \cap \omega_2 \cap \omega_3 \cap \omega_4).\end{aligned}$$

Making a simplification of notation of $(nv)_{ij\ldots}$ for $n_{ij\ldots}v_{ij\ldots}$ write

$$\begin{aligned}\psi(B_4) &= \psi(n_4v_4) - \psi((nv)_{14}) - \psi((nv)_{24}) - \psi((nv)_{34}) \\ &+ \psi((nv)_{124}) + \psi((nv)_{134}) + \psi((nv)_{234}) - \psi((nv)_{1234}) \\ &= \psi((nv)_4) - \sum_{j=1}^{3}\psi((nv)_{j4}) + \sum_{j_2=2}^{3}\sum_{j_1=1}^{j_2-1}\psi((nv)_{j_1j_24}) - \psi((nv)_{1234}).\end{aligned}$$

Note that $(nv)_{14} \geq (nv)_{124}$, $(nv)_{34} \geq (nv)_{134}$, and $(nv)_{24} \geq (nv)_{234}$, thus $\psi(B_4) \leq n_4v_4 = \psi(\omega_4)$.

There is an induction argument here. In general

$$\psi\left(B_k\right)=\psi\left(n_k v_k\right)-\sum_{j_1=1}^{k-1}\psi\left(\left(nv\right)_{j_1 k}\right)+\sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{k-1}\psi\left(\left(nv\right)_{j_1 j_2 k}\right)$$
$$-\sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{j_3-1}\sum_{j_3=3}^{k-1}\psi\left(\left(nv\right)_{j_1 j_2 j_3 k}\right)+\ldots$$

Thus the measure of the collection of non-disjoint analyses, $\psi\left(\bigcup_{\omega_i\subset A}\omega_i\right)$ can be assembled into the sum of mutually disjoint combinations of the measures of analysis quanta, which themselves are the measures of discrete fragments of analyses components, permitting $\int_{A_q} d\psi=\sum_{\omega_i\subset A_q}\psi\left(B_i\right)$ where $\sum_{\omega_i\subset A_q}B_i=\bigcup_{\omega_i\subset A_q}\omega_i$ and thus $\int_{A_q} f\left(\omega_i\right)d\psi=\sum_{B_i\subset A_q} f\left(\omega_i\right)\psi\left(B_i\right).$

Strategy in calculating quanta

The fact that the quantity $\psi\left(B_k\right)$ is itself composed of alternating sums and differences of the intersections of increasing numbers of measures of analysis quanta helps us in its calculation. Specifically, $\psi\left(B_k\right)$ is the measure of ω_k minus the sum of all of the dual analysis interactions that involve ω_k plus the sum of the measure of each of the triple interactions that involve ω_k minus the sum of the fourth level interactions that involve ω_k and so on.*

* The number of terms that comprise each of the levels of interactions for $\psi\left(B_k\right)$ are generated from the k^{th} row of the golden triangle.

As an example, in order to compute $\psi(B_7)$, first subtract from $\psi(\omega_7)$ the six binary interactions that involve analysis ω_7 then add the 15 triple intersections terms that include ω_7 then subtract the 20 fourth level interactions that involve ω_7, etc. This process of term counting provides a simple way to compute $\psi(B_k)$ when $\psi\left(\bigcap_{i=1}^{n}\omega_i\right)$ is same constant for all *n* above some value, as demonstrated in examples in some of the later chapters. In addition, the computation is eased when the magnitude of the higher order interactions decreases e.g., when $\lim_{n\to\infty}\psi\left(\bigcap_{i=1}^{n}\omega_i\right)=\lim_{n\to\infty}\left(n_{123\ldots n}v_{123\ldots n}\right)=0.$ *

It's now time for a welcome review of where we are.

---

* Since there is no clinical research effort with an infinite number of participants and variables, this limit argument is not as helpful as we might like. The smallest nonzero value $\psi\left(\bigcap_{i=1}^{n}\omega_i\right)=n_{123\ldots n}v_{123\ldots n}$ can be is one, since an analysis, in order to have positive measure must have at least one observation and one variable.

A breather…

The last chapter covered a lot of new material, so Let's pause for a moment, and recapitulate.

We are interested in measuring the content of analysis. The purpose of this measurement is to provide an assessment of the content of an analysis in clinical research to address a clinical question q.

Since we want to take advantage of many features measure has to offer (including the ability to assess the content of overlapping analyses), we have selected a content function (working to keep it tractable for computing), we have defined the analysis content as the product of the number of participants and the number of variables used in the analysis.

In Chapter 14, we found that this function of an analysis, ψ was not just a content, but a formal measure. This permits us to use ψ as a set function, with some very useful properties (e.g., the ability to compute it on unions of analyses.

Using our set theory, we saw that when the collection of analyses $\{\omega_1, \omega_2, \omega_3 ... \omega_n\}$ are disjoint we could write

$\psi\left(\bigcup_{i=1}^{n}\omega_i\right)=\sum_{i=1}^{n}n_iv_i$. However, our recognition that collections of analyses most frequently have shared participant and variable data, in combination of the work of the previous chapter reveals that when the collections of analyses are nondisjoint, then $\psi\left(\bigcup_{i=1}^{n}\omega_i\right)$ can be assembled into the sum of mutually disjoint ψ - measures not of analyses but analyses fragments or quanta $\{B_1, B_2, B_3 ... B_n\}$. Since these analysis quanta are disjoint, we can write $\psi\left(\bigcup_{i=1}^{n}\omega_i\right)=\sum_{i=1}^{n}\psi(B_i)$ where

$$\psi(B_i)=\psi(n_iv_i)-\sum_{j_1=1}^{i-1}\psi\left((nv)_{j_1i}\right)+\sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{i-1}\psi\left((nv)_{j_1j_2i}\right)$$
$$-\sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{j_3-1}\sum_{j_3=3}^{i-1}\psi\left((nv)_{j_1j_2j_3i}\right)+...$$

We will come back to this computation in a moment. However, focusing on the big picture, we can now answer the question of what is the ψ - measure of a collection of analyses that addresses a clinical research question q. This operation is simply the accumulation of measure.

Recall that such an accumulation is signaled through the use of the integral sign. Thus the integral $\int_{A_q} d\psi$ is the statement of our intent to accumulate the content of the set of analyses A_q using ψ -measure. Thus $\psi(A_q)=\int_{A_q} d\psi$.

However now that we have divided the analyses into

disjoint analysis fractions or quanta, we can go one step further and write

$$\psi\left(A_q\right)=\int_{A_q} d\psi=\sum_{\omega_i\subset A}\psi\left(B_i\right).$$

The content measure of the analysis set A_q is simply the sum of the ψ-measures of the analyses quanta into which the set $A_q=\{\omega_i\}$ is fractionated. If the analyses are pairwise disjoint then the equation simplifies to $\psi\left(A_q\right)=\sum_{\omega_i\subset A} n_i v_i$.

Some helpful observations

Let's make some observation before we proceed.

First, since analyses in clinical research effort commonly use the same subjects of observations and variables, it will be the rare collection of analyses that are pairwise disjoint. Thus, the equation on which we will rely on will be $\psi\left(A_q\right)=\int_{\omega_i\subset A_q} d\psi=\sum_{\omega_i\subset A_q}\psi\left(B_i\right)$.

In addition, the computation of $\psi\left(B_i\right)$ may appear complicated, but it is simply adding and subtracting sums of participant-variable products in a specific sequence. This is readily done in a program like Excel, so we will not let this calculation impede our examination of this content measure's performance.

With this development behind us, we are now in a position to compute not just the total content of a collection of analyses $\{\omega_1,\omega_2,\omega_3,...\omega_n\}$ as

$\psi\left(\bigcup_{A_q}\omega_i\right)=\psi\left(\bigcup_{i=1}^{n}\omega_i\right)=\sum_{i=1}^{n}\psi\left(B_i\right),$ but we can also compute the fraction of the total content of the set of analyses contained by the i^{th} analysis, $\frac{\psi\left(B_i\right)}{\sum_{i=1}^{n}\psi\left(B_i\right)}$. This proportion can function as a weight to be used when we want to assess the role of function of the analysis.

This is useful because ultimately are interested is not just in $\psi\left(A\right)=\int_{A_q} d\psi=\sum_{\omega_i\subset A}\psi\left(B_i\right),$ but in $\int_{A_q} f\left(\omega_i\right)d\psi=\sum_{\omega_i\subset A} f\left(\omega_i\right)\psi\left(B_i\right),$ where the function $f\left(\omega_i\right)$ reflects, for example the duality of an analysis result. In this case the function of interest will be

$$\frac{\int_{\omega_i\subset A} f\left(\omega_i\right)d\psi}{\int_{\omega_i\subset A} d\psi}=\frac{\sum_{\omega_i\subset A} f\left(\omega_i\right)\psi\left(B_i\right)}{\sum_{\omega_i\subset A}\psi\left(B_i\right)}=\sum_{\omega_i\subset A} f\left(\omega_i\right)\left[\frac{\psi\left(B_i\right)}{\sum_{\omega_i\subset A}\psi\left(B_i\right)}\right]$$

In this case, the component $\frac{\psi\left(B_i\right)}{\sum_{\omega_i\subset A}\psi\left(B_i\right)}$ is a weighting function, either increasing or decreasing the contribution $f\left(\omega_i\right)$ based on the proportion of the total content-

measure in the set of analyses A that is explained by individual analysis ω_i.

Now let's look at some simple examples.

A first demonstration

We have so far developed an analysis platform specifically tailored to health care research. It is based on the ultimate source of useful information in clinical research – the participants and their variables.

We have used set theory to assure ourselves that there are clear mathematical rules governing how we manage collections of data in this participant-variable dimension. And we know enough about measure theory to understand that this ψ -content function is a measure. Thus, at least theoretically, we can compute the ψ -content of a single analysis, or a collection of analyses in a reliable and consistent fashion, and now refer to this analysis content as ψ - measure.

This ψ - measure permits us to compute the contribution of an analysis in our construct; this contribution will be the proportion of content-measure contained in the analysis, or $\frac{\psi(B_i)}{\int_A d\psi} = \frac{\psi(B_i)}{\sum_{\omega_i \subset A} \psi(B_i)}$, where the collection of sets $\{B_i\}$ represents the disjoint

contributions of each analysis' content-measure to the overall content of the analysis set $\int_A d\psi$, expressed as a simple proportion.

An examination of the characteristics of this measure will help us to assess its ability to identify potential contributions of different analyses to clinical research results. Our construction of the set $\{B_i\}$ in Chapter 15 demonstrated that it is an individual analysis' quanta, B_i that makes a contribution to the content-measure of the union of the set of analyses $\bigcup_q \omega_i$ separate and apart from the other ω_i, it will therefore be of particular interest to determine when these quanta make substantial versus small contributions to the ψ -measure of this union.

Thus, comparing the content-measures of analysis sets in particular research circumstances will reveal how to use $\psi -$ measure and give us at least an initial sense of how its output compares with the commonly used standard.

Example: A single primary endpoint.
Let's begin with a randomized clinical trial with two treatment arms designed to assess the effect of a new therapy in patients with heart failure. Subjects are randomized into two treatment groups (treatment and placebo), and will have several outcome measures assessed. Each of these outcomes is prospectively declared, and measured with high quality.

For example, the heart's ability to pump effectively will be directly assessed by estimators that assess ejection fraction, end systolic volume, end diastolic volume, or cardiac output. In addition, there are evaluations of the vitality of the individual (e.g., walking distance or quality

of life questionnaires). Each of these measures is taken at baseline and then once again at some pre-specified time in the future (e.g., one year), and their differences obtained and assessed.

In the traditional paradigm, a clinical trial focuses on one (or a small number) of endpoints/analyses, (termed "primary") even though several and sometimes many more endpoints and analyses were prospectively declared; the trial results are based in large part (and sometimes exclusively) on the magnitude of these primary analyses' *p*-values.

Thus, in a trial with a single primary endpoint selected from *k* prospectively declared analyses, the remaining $k-1$ analyses, while interesting and providing additional support for the primary endpoint's finding, in and of themselves do not formerly contribute to the result of the study. The term typically used for these additional, secondary endpoint findings is "supportive".

Our goal is to examine the implications of the application of this ψ -measure in this example.

Assume that there are $\omega_1, \omega_2, \omega_3, ..., \omega_K$ individual analyses to be carried out in this clinical trial, one for each of the *k* prospective declared evaluations, and let ω_1 be the primary analysis.

Let's also assume for simplicity that the same *n* subjects are included in each analysis* and that each analysis consists of three variables (baseline evaluation, follow-up evaluation, and treatment assignment).

The variable denoting treatment assignment is the same across all analyses, but the endpoint measures (baseline and

* The realility of subjects with missing values is easily incorporatated in ψ -measure.

follow-up assessments) are different from analysis to analysis. Our goal is to compute the ψ -measure of the union of these analyses, $\psi\left(\bigcup_{i=1}^{K}\omega_i\right)$ representing the total content of all of these endpoints taken together.

Using our development, we must construct a collection of sets $\{B_1, B_2, B_3, ..., B_K\}$ that are disjoint as in the previous section, and for which $\psi\left(\bigcup_{i=1}^{K}B_i\right)=\psi\left(\bigcup_{i=1}^{K}\omega_i\right)$.

Proceeding, for the first analysis there are *n* subjects and three variables (baseline evaluation, follow-up evaluation, and treatment assignment). Thus, from our formula for ψ -measure, compute

$$\psi(B_1)=\psi(\omega_1)=n_1v_1=3n.$$

In order to compute $\psi(B_2),$ begin with

$\psi(B_2)=n_2v_2-n_{12}v_{12}=3n-n_{12}v_{12}$; n_{12} the number of observations common to the first two analyses which in this example is *n*, and $v_{12}=1$ since there is only one variable that is common to both analyses (the treatment group assignment). Therefore, we can compute

$$\psi(B_2)=n_2v_2-n_{12}v_{12}=3n-n=2n.$$

Let's pause for a moment. Recall that $\psi(B_2)$ is the contribution of analysis ω_2 to the content-measure of $\omega_1\cup\omega_2$, $\psi(\omega_1\cup\omega_2),$ separate and apart from the contribution of analysis ω_1. Note that $\psi(\omega_2)=3n>\psi(B_2),$ that is, the measure of analysis two is greater than its measure when its content commonality with analysis one is removed.

Continuing, write

$$\begin{aligned}\psi(B_3) &= \psi(\omega_3) - \psi(\omega_1 \cup \omega_3) - \psi(\omega_2 \cup \omega_3) + \psi(\omega_1 \cup \omega_2 \cup \omega_3) \\ &= n_3 v_3 - n_{13} v_{13} - n_{23} v_{23} + n_{123} v_{123} \\ &= n(3-1-1+1) = 2n.\end{aligned}$$

In fact, $\psi(B_j) = 2n$ for all $j > 1$.* Thus, each subsequent analysis after the first contributes a separate ψ - measure or quantum of $2n$.
We are now ready to compute that

---

* The quantity $\psi(B_k)$ is itself composed of alternating sums and differences of the intersections of increasing numbers of measures of analysis quanta. Specifically, $\psi(B_k)$ is the measure of ω_k minus the sum of all of the dual analysis interactions that involve ω_k plus the sum of the measure of each of the triple interactions that involve ω_k minus the sum of the fourth level interactions that involve ω_k and so on. The number of terms that comprise each of the levels of interactions for $\psi(B_k)$ are generated from the k^{th} row of the golden triangle. Thus, in order to compute $\psi(B_7)$ subtract from $\psi(\omega_7)$ the six binary interactions that involve analysis ω_7 then add the 15 triple intersections terms that include ω_7 then subtract the 20 fourth level interactions that involve ω_7, etc. This process of term counting provides a simple way to compute $\psi(B_k)$ when $\psi\left(\bigcap_{i=1}^{n} \omega_i\right)$ is same constant for all *n* above some value, as demonstrated in the examples.

$$\psi\left(\bigcup_{i=1}^{k}\omega_j\right)=\sum_{i=1}^{k}\psi\left(B_j\right)$$

$$=3n+\sum_{k=2}^{n}2n=3n+2\left(k-1\right)n=\left(2k+1\right)n.$$

Typically, in clinical trial analyses, the only endpoint that contributes to a quantitative estimate of the effect of therapy is the primary endpoint. However in this formulation, positive measure is available to serve as the basis for a mathematical contribution for each of the *k* prospectively declared endpoints regardless which one is primary.

In fact for three prospectively declare endpoints, the total content-measure is $3n+2n+2n=7n$, the fraction of the total content-measure contained by the primary endpoint is $3n/7n=0.429,$ with 28.6% content-measure remaining in each of the two remaining endpoints.

Thus if this were a trial with 87 subjects, a primary outcome of the placebo corrected effect of left ventricular end systolic volume, and two prospectively declared secondary outcomes of a) the placebo adjusted change in left ventricular end systolic volume, and b) the placebo adjusted change in left ventricular end diastolic volume, then $\psi\left(\omega_1\right)=(3)(87)=261,$ and we can also compute $\psi\left(\omega_2\right)=\psi\left(\omega_3\right)=(2)(87)=174.$

Note again that the raw content measure of an analysis ω_i $\psi\left(\omega_i\right)=n_i v_i$ can be different from that analysis' contribution to $\psi\left(\bigcup_{i=1}^{k}\omega_j\right)$. While $\psi\left(B_1\right)=\psi\left(\omega_1\right)$ by

definition, $\psi(B_2) = 2n \neq \psi(B_1) = 3n$. This is because $\psi(B_2)$ reflects the contribution of analysis ω_2 after taking ω_1 into account. Similarly, $\psi(B_i)$ is the measure of ω_i after taking into account the sequence of analyses $\omega_1, \omega_2, \omega_3, ...\omega_{i-1}$.

Initial analysis sequencing observation

In the previous example with three outcomes and no missing data, we find that, however the three outcomes are sequenced $\psi\left(\bigcup_{i=1}^{3}\omega_j\right) = 7n = \sum_{i=1}^{3}\psi(B_i)$. Each of these two computations is sequence invariant.

However, the value of $\psi(B_i)$ is not. It is easy to see from this simple example that the contribution of the quantum $\psi(B_i)$ is either $3n$ or $2n$ depending on where in the sequence of analyses, $\omega_1, \omega_2, \omega_3, ..., \omega_n$ the analysis ω_i resides.

This finding has important implications for the use of quanta analysis and will now be examined in detail.

Analysis priorities and quanta paths

Our first demonstration of quanta analysis applied to a simple clinical trial scenario provided some intriguing findings.

First, not all content-measure is subsumed by the primary outcome. In fact, in the previous example with 87 subjects and three prospectively declared outcomes, more than 50% of the total content-measure of all three analyses resides with the two non-primary evaluations.

This observation opens the door to the possibility that prospectively declared, non-primary analyses may also be quantitatively considered in summarizing a trial result.

The current state of the art analysis paradigm operates as though all "analysis-measure" is absorbed by the primary evaluation. In this traditional framework, the secondary outcomes may of course be cerebrally integrated into the final result but they are treated as though they "have measure zero".

The quanta analysis opens the door to the possibility of a quantitative combination of primary and secondary outcomes, a concept that will be carefully cultivated and developed later in this book.

Sequencing variant quanta values

A second observation is that the contribution of an individual analyses to the content-measure contained by the union of all analyses depends on where in the sequence of evaluations the particular analysis lies. Thus, while the overall measure of the union of all analyses is the same regardless of sequencing, the contribution that any particular analysis makes to that cumulative content-measure depends on where in the sequence of analyses it is located.

This is an important new finding of the quanta approach, that at first blush takes us aback.

Having an analysis' quantum contribution be based on the sequence of analyses means that there is no unique "solution " to the value of $\psi(B_i)$. If for example, there are three prospectively declared outcomes, the difference in the change in each of 1) left ventricular ejection fraction (LVEF), 2) end systolic volume (ESV) and 3) end diastolic volume (EDV), then there are $3! = 6$ possible sequences of analyses. They are

LVEF ESV EDV
LVEF EDV ESV
ESV LVEF EDV
ESV ESV LVEF
EDV LVEF ESV
EDV ESV LVEF

Each represents a sequence of evalution and evaluate an outcome in a different sequence position.

How do we manage this?

As it turns out, this sequence dependent value of the quantum $\psi(B_i)$ means that additional input is required to

actually set it. The non-uniqueness of $\psi(B_i)$ is precisely what we need.

That input comes from the clinical investigators.

Assigning location to sequence variant analyses

It is the investigators and epidemiologists who provide the missing sequencing information.

For example, if the investigators believe that there is a clear primary endpoint that can be prospectively declared, measured precisely and is both accepted and expected by the research and regulatory community, then this should be the first outcome in the sequence ω_1.

For all other outcomes that use the same participants and variables, then declaring the primary analysis as ω_1 gives the primary outcome the maximum measure. In our previous example, the primary endpoint of left ventricular ejection fraction would be first in the sequence, and as we computed, $\psi(B_1) = \psi(\omega_1) = 261$. The secondary endpoints each had a content measure of $\psi(B_2) = \psi(B_3) = 174$.

If not left ventricular ejection fraction, but left ventricular end systolic volume was the primary endpoint, then it would be left ventricular end systolic volume that had the quanta value of 261. In this elementary example, it does not matter which is the second or third analysis outcome in the sequence because they each have the same quanta value*

In fact, the sequence dependent nature of the quanta construction is the mathematical justification for the selection of the primary endpoint.

* These are sequent invariant quanta in this example.

The motivations for the selection of a primary outcome is multidimensional. The epidemiologic goal is the selection of an outcome as primary is to have an outcome measure that is responsive to the intervention, can be measured with sufficient precision, and is acceptable to the research and regulatory community. Once the selection is made, a decision is made about alpha error expenditure in the traditional paradigm.

From the measure theoretic perspective, making this selection is the same as choosing the first analysis ω_1 in the sequence of analyses; maximum content-measure is assigned to it in accordance with our set theoretic rules for manipulating measure, avoiding any reduction due to intersections with other analysis sets. *

The primary analysis is given the optimal content-measure, *ceteris paribus.* However, this does not mean that the primary outcome has the greatest measure. Consider the circumstance where the primary outcome consists of an analysis based on 100 participants and 3 variables, while the single secondary analysis consists of 100 participants and 8 variables, 3 of which are in common with the first analysis. Then, we compute the measure for the primary outcome as $\psi(B_1) = n_1 v_1 = 300,$ which is less than the ψ -measure for the secondary outcome, which is $\psi(B_2) = n_2 v_2 - n_{12} v_{12} = 800 - 300 = 500.$ In this

* One might argue that the measure theoretic approach of assigning content measure is not unlike asigning alpha prospectively. However content-measure can be assigned to a rich collection of complex intersections and unions of analyses. While probability is a measure (demonstrated by the great Russian probabilist Andrei Kolmogorov), the *p*-value (which is a specific probability of a particular event) is not, limiting its flexibiity as we will see in future chapters.

circumstance, the quantum for the secondary analysis exceeds that of the primary analysis.

By optimal content measure, we simply mean content-measure that is not adjusted for intersections with other analysis sets, i.e., $\psi(B_1) = \psi(\omega_1)$.* The observations and variables that contribute to the primary analysis outcome are first made to the primary analysis. This gives the primary analysis the greatest opportunity to have a large influence on $\psi\left(\bigcup_{i=1}^{k} \omega_j\right)$. Other analyses can be considered, but only if they have positive content-measure after consideration of the primary outcome.

Example: Multiple Primary Outcomes:
Let's now modify the first example, now affirming there are three prospectively declared competitors to be the primary outcomes; 1) the difference in the change in left ventricular ejection fraction, (ΔL), 2) the difference in the change in left ventricular end systolic volume (ΔS), and 3) the difference in the change in left ventricular end diastolic volume (ΔD). We assume the same number of observations and variables as the previous example.

One traditional approach to managing this problem would require the physician investigators to select just one of them as the primary outcome, relegating the other two to playing secondary, supportive roles. This places the investigators in a difficult position because there may be no

* This also demonstrates from a content measure approach the potential value of considering secondary outcomes with the primary analysis which will be discussed later.

scientific justification for the selection of one of these as primary over the other two.

The alternative would be to select two or three of them as primary and distribute type I error in accordance with a multiplicity criteria such as Bonferroni. This choice increases the sample size*.

From the measure-theoretic perspective, we take a different perspective. We already know that for the outcome that is the primary we would assign content-measure of 300 and to each of the other two we would also assign 200 to produce $\psi\left(\bigcup_{i=1}^{3}\omega_i\right) = 700$. We have 6 such sequences of primary outcomes† For each sequence we will plan to conduct a full quanta analysis. Then when completed, there results will be averaged.

Specifically, if the goal of the evaluations is to compute

$$\frac{\int\limits_{\omega_i \subset A} f(\omega_i)\, d\psi}{\int\limits_{\omega_i \subset A} d\psi} = \frac{\sum\limits_{\omega_i \subset A} f(\omega_i)\psi(B_i)}{\sum\limits_{\omega_i \subset A} \psi(B_i)} = \sum_{\omega_i \subset A} f(\omega_i)\left[\frac{\psi(B_i)}{\sum\limits_{\omega_i \subset A} \psi(B_i)}\right],$$

then the final equality on the right will be computed for each of the six primary outcome sequences. Thus we can write our goal as

---

* A multipliicity criteria decreases the type I error for each endpoint selected as primary, and decreased type I error ceteris paribus, increases the sample size.

† $(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)$

$$n_s^{-1}\sum_{s=1}^{n_s}\sum_{\omega_i\subset A} f\left(\omega_i\right)\left[\frac{\psi\left(B_{i,s}\right)}{\sum_{\omega_i\subset A}\psi\left(B_{i,s}\right)}\right]$$

where *s* indexes the sequences, n_s is the number of sequences that have to be considered, and $B_{i,s}$ is the quantum of the i^{th} analysis in the s^{th} sequence.

Since there is no priority in the outcome sequences, they should be weighted the same, permitting each sequence's impact to be equally considered.

This procedure frees investigators from having to choose a single primary outcome from among several absent criteria to inform the selection process.

This approach is easily adaptable to other circumstances. Let's modify the previous example somewhat more. Now the investigators choose a single primary outcome, and the remaining two are secondary. In the traditional paradigm, there is no hierarchy among secondary outcomes. There, the secondary outcomes do not quantitiatively commit to the result of the clinical trial, so no hiearchy is required. Secondary outcomes are reported as merely supportive.

From the quanta perspective, as we have seen, secondary outcomes can provide quantitative support for the trial's answer to the clinical question that motivated it. In the circumstance where there is one primary outcome and two secondary oucomes, one can sequence; however the sequences is retricted.

Specifically, the primary outcome is the first outcome in any sequence, and the secondary outcomes (and only the secondary outcomes) are permuted. Thus, there are only

two that must be considered $(1,2,3),(1,3,2)$, and only these two sequences need have their results averaged.

At what level does averaging take place:
By computing these averages, the investigators are released from having to artificially select a primary endpoint from among candidates primary endpoints which are each, from a precision, culture, and sample size perspective, essentially equivalent. It eliminates the need to make a "best guess" at what the primary endpoint should be. * However, we must be clear where the averaging takes place. We are not averaging ψ -measures, for which there is little justification from our measure-theoretic background. Instead we are averaging at the level of our integral $\int_{\omega_i \subset A} f(\omega_i) d\psi$ itself suitable normed. . How this operates will be clear in the dicussion following the topside function development.

Multiple manuscripts
As a final example in this chapter, consider that a collection of investigators conducts a randomized clinical trial to examine the relationship between a new lipid lowering agent in patients at risk of having a second heart attack. In accordance with the traditional paradigm, they conduct analyses on a single primary endpoint (e.g., combined fatal and nonfatal myocardial infarction) on the overall cohort, conduct similar analyses on a small number of prospectively declared secondary outcomes, and then examine the effect of therapy on the primary outcome for a

* How this operates will be demonstrated after the top side function discussion.

number of proper subgroups (including subjects with diabetes).* These results are published.

As is commonly the case, the investigators then decide to conduct a new sequence of analyses on the subgroup of diabetic patients. These analyses, examine the effect of therapy on the primary and all secondary analyses on the diabetic subgroup, as well as on a collection of evaluations particularly focused on the impact of diabetes mellitus (e.g., amputations, stroke, and deterioration of vision). These additional results in this diabetic cohort are published in a second manuscript.

The question is, how is type I error managed across the two manuscripts? Is there an overall assessment of the type I error rate?

The practitioners of statistical hypothesis testing are relatively mute on the application of the *p*-value arithmetic in this rather simple and common research paradigm. Even though the analyses for the second paper on the diabetic cohort was prespecified, there was no *prima facie* type I error for the analyses of the second paper.

If this were the case, then the investigators would have had to hold some portion of the some portion of the type I error rate aside for the second paper in the traditional paradigm. But then, how would this portion of the type I error have to be apportioned in the analyses on the diabetic subcohort. This would be an awkward alpha calculation at best, demonstrating the relative inflexibility of statistical inference to mange a common problem in the many clinical research efforts that each generate multiple manuscripts.

The quanta evaluation process is not interrupted simply because different collections of analyses are segregated into

* There are also safety analyses. These will be discussed in Chapter 21.

different publications. The investigators simply need to choose their sequence of analyses, and then compute the quanta for the collection of sets $\{B_i,\ i=1,2,3...m\}$ which span the two (or more) papers. It is possible that some of the quanta that are deep in the sequence may be exceedingly small, but if there is no consensus on the sequence of analyses beyond a certain point in the sequence, then one simply uses the function

$$(m-k)\sum_{s=1}^{m-k}\sum_{\omega_i\subset A} f(\omega_i)\left[\frac{\psi(B_{i,s})}{\sum_{\omega_i\subset A}\psi(B_{i,s})}\right]$$ to average over the $m-k$ analyses that are beyond the sequencing ability of the investigators. Recalling that this computation is for a study response to a question q, this response accepts the contribution of each analysis to the question's answer, regardless of which paper in which it appeared.

Subgroup Evaluations

Another situation in which the concept of analysis measure can make a contribution to the evaluation of clinical trials is in subgroup evaluations.

In a clinical trial, a subgroup analysis is the evaluation of a randomly assigned exposure on a sub-cohort based on strata membership determined by participant characteristics at baseline (e.g., gender or age).

These evaluations are historically fraught with concern because they can involve a small number of subjects, the strata specific statistical hypothesis tests do not have adequate statistical power, and the type I multiplicity metric is difficult to address. Therefore, the standing rule for subgroup evaluations is to essentially set aside the

subgroup group finding and be guided by the finding in the overall cohort [1].

It would be interesting to examine this issue from the set and measure theoretic perspective.

As an example, consider a randomized clinical trial assessing the change from baseline to follow-up for a single outcome by therapy assignment. The trial conducts five analyses.

1) ω_m = the effect of therapy on males
2) ω_f = the effect of therapy on females
3) ω_w = the effect of therapy on whites
4) ω_n = the effect of therapy on nonwhites
5) ω_T = the effect of therapy on the overall cohort.

In computing the content- measure of each of these analyses, the number of variables utilized in each analysis will be the same three (baseline measure, follow-up measure, and treatment identity). However, the number of observations varies from analysis to analysis. Since only males are considered in the first analysis ω_m, write $\psi(\omega_m) = 3n_m$, where n_m denotes the number of males. Similarly, for females find $\psi(\omega_f) = 3n_f$. Then in a straightforward manner, we can compute the necessary quanta $\psi(B_i), i = 1...5$. Observe that

$$\psi(B_m) = 3n_m$$
$$\psi(B_f) = 3n_f - 3n_{mf} = 3n_f.$$

Because there are no individuals who are both male and female, and similarly no participants who are both while and nonwhite, the quanta of B_w is computed to be

$$\begin{aligned}\psi(B_w) &= 3n_w - 3n_{mw} - 3n_{fw} + 3n_{wnf} \\ &= (3n_w - 3n_w) + 3n_{wnf} = 0 + 0 = 0.\end{aligned}$$

The contribution of the white race strata is zero after considering the contribution of both gender strata. A similar result is identified for

$$\psi(B_n) = 3n_n - 3n_{mn} - 3n_{fn} + 3n_{mwn} + 3n_{mfn} - 3n_{nfwn} = 0.$$

Thus the contribution of the two racial analyses to the accumulating union, $\bigcup_{i=1}^{4} \omega_i$ is zero after considering the contribution of the two gender analyses.

However, reversing the sequence, the contribution of the two gender analyses to $\psi\left(\bigcup_{i=1}^{4} \omega_i\right)$ is zero after considering the two racial strata analyses.

It can also be shown that the contribution of the analysis of the total cohort is zero after considering either the two gender analyses or the two racial analyses. In addition, the contribution of both the gender and the race strata is zero after first considering the contribution of the total cohort. To continue this example, if the analyses are considered in the sequence $\omega_T, \omega_m, \omega_f, \omega_w, \omega_n$ then each of

the gender strata and the race analyses make no contribution to the measure of $\bigcup_{i=1}^{5} \omega_i$.

Thus, as in the previous example, while $\psi\left(\bigcup_{i=1}^{5} \omega_i\right) = \sum_{i=1}^{5} \psi\left(B_i\right)$ is sequence invariant, the contribution of each of the ω_i (through its computation of B_i) to this measure is sequence dependent.

The subgroup example demonstrates the extreme redundancy in analyses can drive particular analysis quanti B_i to zero. The operation in the subgroup example is a mathematical justification for the discounting of subgroup analyses after the overall cohort has been assessed.*

Notation

In order to incorporate the concept of a priority sequence structure formally into the measure accumulation mechanism, define a function T that oversees the reordering of the sequence of analyses $\{\omega_i\}$, from essentially a random collection of analyses to the ordered set, to the investigator chosen sequence of analyses. Here $T\left(\omega_1, \omega_2, \omega_3, ..., \omega_n\right) = \omega_{[1]}, \omega_{[2]}, \omega_{[3]}, ..., \omega_{[n]}$ where the subscript $[i]$ denotes the i^{th} analysis in the order or priority determined by the investigators. Note that the function T also converts the sequence $\{B_i\}, i = 1, 2, 3, ...$ to

* The general linear model assessment of subgroups could also provide mathematical justification for this assertion. However the demonstgration is analysis dependent. The set theoretic perspective dismisses the post total cohort assesmment subgroup analysis regardless of the perspective.

$\left\{B_{[i]}\right\}, i = 1, 2, 3, \ldots$ with this latter sequence of disjoint sets corresponding to the sequence of ordered analyses. Note that the function T operates on the entire set. To reflect the order of analyses chosen by the investigators, we may write

$$\psi\left(\bigcup_{i=1}^{n}\omega_i\right) = \sum_{i=1}^{n}\psi\left(B_{[i]}\right),$$

and $n_s^{-1}\sum_{s=1}^{n_s}\sum_{\omega_i \subset A} f\left(\omega_i\right)\left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i \subset A}\psi\left(B_{[i],s}\right)}\right]$ where n_s is the number of sequences that were examined by the investigators.

Chapter Summary

In this chapter, we have observed that 1) the mathematical representation of the investigator's choice of their sequence of analyses is an important determinant of the contribution of each analysis to the measure of the union of the collection of analyses and 2) providing the value of $\psi\left(B_i\right)$ for each analysis measures the contribution of that each analysis' quanta to the union of the collection.

Investigators working within the customary design paradigm make decisions about the priority of analyses. These decisions are based on accuracy, precision, and the persuasive power of the endpoint but also include type I error considerations (i.e., how much alpha should be allocated to each analysis). In this traditional framework, analyses that are not included in the alpha spending function make no formal contribution to the overall trial assessment.

The measure theoretic infrastructure requires a selected sequence of analyses that is based only on accuracy, precision, and persuasive force. This opens the door to the inclusion of many different analyses to be included in the assessment of the overall research effort.

We are now ready to consider the "topside function" $f(\omega_i)$.

References

1. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA. 1991 Jul 3;266(1):93-8

.

Topside functions

The work that we have invested thus far in this project has been to essentially develop, through an application of measure theory a 'weighting factor" for a collection of functions. Essentially, we have developed every concept and justified each variable in the formula

$$\frac{\int\limits_{\omega_i \subset A} f\left(\omega_i\right) d\psi}{\int\limits_{\omega_i \subset A} d\psi} = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A} f\left(\omega_i\right) \left[\frac{\psi\left(B_{[i],s}\right)}{\sum\limits_{\omega_i \subset A} \psi\left(B_{[i],s}\right)} \right]$$

except for $f\left(\omega_i\right)$. It is now time to develop this function, and other pertinent functions like it.

As always in this development, we are working with the familiar our analysis space and σ-algebra$\left(\Omega, \Sigma\right)$. This means, that any function in the above formula $f\left(\omega_i\right)$ must be measurable against $\left(\Omega, \Sigma\right)$ which is another way to state that it must 1) not be negative, and 2) must derive its value

from a property of the analysis ω_i that is available for inspection, as previously discussed. This property will be the plausible interval of the estimator that is available from an inspection of analysis ω_i.

Recall that we began this book with a discussion of duality, i.e., the concept that a single result (be it a serum sodium level in a clinical setting, or a research based placebo adjusted difference in exercise tolerance) could simultaneously stand for the occurrence of benefit and of harm.

In this chapter, we return to this idea, now developing the mathematics around it. Once we have completed these new functions' ontogenies, we will combine then with the set-theoretic development thus far. Mathematically, we will develop $f(\omega_i)$ for benefit and for harm. We will then accumulate it using

$$\frac{\int\limits_{\omega_i \subset A} f(\omega_i) d\psi}{\int\limits_{\omega_i \subset A} d\psi} = \sum_{\omega_i \subset A} f(\omega_i) \left[\frac{\psi(B_{[i],s})}{\sum\limits_{\omega_i \subset A} \psi(B_{[i],s})} \right].$$

Finally we will norm this over all analysis paths to compute

$$n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A} f(\omega_i) \left[\frac{\psi(B_{[i],s})}{\sum\limits_{\omega_i \subset A} \psi(B_{[i],s})} \right].$$

to compute a normed value of this function over all analysis sequences.

This family of functions $f\left(\omega_i\right)$ I will call the topside function, because it is the "topside" or numerator to be divided by the quanta component $\frac{n_s}{\psi\left(B_{i,s}\right)}\sum_{\omega_i \subset A}\psi\left(B_{i,s}\right)$.

Return to duality

Remember that we defined duality as the situation where either a lab value, or a single estimator of an exposure's effect in a clinical trial can simultaneously support the concept of benefit and the finding of harm.

Statistical hypothesis testing, it will be remembered, simply rejects a null hypothesis or it does not; the test statistic simply falls into the critical region or it does not. It's mathematics are dichotomous.

Duality is a more complex but realistic assessment of what the state of the result is that is wholly consistent with the interpretation in health care.

As we saw, while clinical investigators can quite naturally be flummoxed by the indirect reasoning of dichotomous statistical hypothesis testing, duality is a concept that reflects their experience in interpretation laboratory and imaging results.

Consider an example where the investigators conduct a clinical trial where the primary outcome is the difference in the change of left ventricular ejection fraction between a group exposed to a new intervention and those in the control group. They determine that this placebo adjusted change in left ventricular ejection fraction (EF) is 6 with a 95% confidence interval of from -2 to 14 based on the standard error.

The standard statistical treatment suggests that the mean placebo adjusted change in EF observed in this sample is consistent with a population change of zero; the result is therefore declared statistically insignificant.

Thus, dualism is not addressed in statistical hypothesis testing. In that realm, results are reduced to "no difference", statistically significant increase", or "statistically significant decrease". The idea that the data can support an increase and a decrease simultaneously is lost.

Duality simply allows us to consider this interval of values as an expanse that simultaneously provides evidence of benefit and of harm. In this case, the range of values consistent with harm is quite small (-2 to 0), while the range consistent with benefit is quite large (0 to 14).

Interval parsing, channeling, and accumulating

What we have just performed is interval parsing. We started with an interval (in this case the standard 95% confidence interval, a concept that we will soon expand), and from that interval, plucked an interval of values consistent with benefit (0-14), and similarly, an interval of values consistent with harm (0-2). This is the parsing component.

What we will do next is channel the benefit interval through a function (0-14) that assesses and norms this interval to that it is unitless. We will perform the analogous operation for the harm interval (-2 to 0).

We then repeat this process for all analyses that are responsive to the clinical research question, accumulating these results for the benefit regions. We repeat this process for the harm region, and then compare the two. It is this accumulation that is accomplished by

$$n_s^{-1}\sum_{s=1}^{n_s}\sum_{\omega_i\subset A} f\left(\omega_i\right)\left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i\subset A}\psi\left(B_{[i],s}\right)}\right]$$

Once complete, we will carry out an analogous collection and integration for each analysis' interval region of harm. In the end we will compare the two.

The duality is incorporated by parsing the plausible intervals of benefit and channeling that region into a benefit function, then conducting the analogous operation for harm.

Our initial concerns

Recall from Chapter 1 that we had two initial concerns about this approach. One was the relationship between the variables being evaluated, i.e., the issue of correlation. This is a straightforward adjustment, but we have not specifically covered this yet.*

However, the second issue of the commonality of observations and variables across analyses we have specifically addressed. ψ - measure is our way of taking into account that the many of the same observations and variables are common to these analyses, and therefore, accumulating the impact of the universe of analyses conducted in the study (our ultimate goal) has to in some way adjust for the multiple use of observations and variables.

The concept of ψ - measure tells us exactly how to discount analyses for the previous use of their observations

* It is discussed in Chapter 22

and variables. This was, if you will, the hard part. All we need to do here is 1) build the plausible interval, 2) parse that interval, and 3) describe the benefit and harm functions through which these parsed intervals are to be channeled.

Beginning construction of the plausible interval

Assume that question q concerns the benefit or harm of an intervention in a clinical trial, e.g., "Does the provision of mesenchymal cells to patients with heart failure ameliorate their signs and symptoms when compared to the experience of controls?" Let's describe the set of analyses that address question q as $A_q = \left\{ \omega_i / q_i \triangleq q \right\}$.

For each analysis ω_i that is a member of A_q, we identify estimate of effect. . For example, it can be the difference between therapy groups of the mean blood pressure change over time, or the relative risk of death associated with an intervention.

Then, for each of these estimators we consider the distorting role of sampling error, bias, and imprecision on this estimate.

Thus, producing the plausible interval for ω_i begins with an inspection of ω_i. One of its elements is the effect size produced by the analysis; this effect size we will notate as e_i. This quantity e_i is any legitimate and well recognized statistical estimator.

Setting the bounds for the plausible interval

Clinical trial analyses produce statistical estimators of an exposure's effect. Investigators, epidemiologists, and statisticians all recognized that a sample-based effect estimate, being a single number, is influenced by factors

unrelated to the effect of the intervention being tested. These factors introduce uncertainty into the effect's true location.

One of these influences is bias. Bias is a systematic influence on the location of the effect size estimator. There are many biases in clinical research.* Fortunately, the degree to which these biases affect a research result can be identified from a detailed examination of the research design. The identification of a particular bias can aid in determining if the observed effect size is too low or too high, allowing one to change the bounds of the plausible interval accordingly.

An additional influence is imprecision. Imprecision is the degree to which the measuring instrument provides a different estimate of an individual's data measurement when the measurement is taken repeatedly. For example, Butler et. al. [1] report that repeat measurements of left ventricular ejection fraction using the same methods in the same patients by experts in echocardiography routinely vary by 7%.

Imprecision is separate and apart from sampling variability, which is the variability introduced by taking a sample of patients from a large population. Although taken from the same population, different samples of that population contains different patients with different life experiences. Thus estimates of an effect vary from sample to sample.

Bias, imprecision, and sampling variability together blur the true location of the effect size estimator, injecting uncertainly into its actual value.

---

* Selection bias, recall bias, ascertainment bias, misclasification bias, immortal time bias are but several of a plethora of biases can influence an effect size estimator.

This blurring of the effect size's location suggests that both larger values and smaller values of the estimator are admissible for consideration. This range of values will be termed the estimator's interval of plausible effects. It is not just the estimator that provides a sense of the effect of the intervention; it is the estimator's interval of plausible values that is most informative about the possible effect size that would be seen in the population.

Experienced workers will see in this concept the familiar idea of a confidence interval. However, while confidence intervals can be used in standard statistical estimation theory as interval estimators, here we will create, then use, and decipher intervals differently.

The principal reason to name our new interval a plausible interval is to differentiate it from the concept of a confidence interval. The confidence interval is based only on sampling error. The plausible interval includes both the influence of bias and the effect of imprecision as well.

Thus the plausible interval is a refraction of the effect location based on the particulars of the individual study design, measurement instrument characteristics, and sampling error. It will be wider than the confidence interval because it includes the additional factors of imprecision and bias.

In this chapter, we will develop this concept generally, leaving actual examples to Chapter 20.

Define the upper e_i^+ and lower e_i^- bounds of the plausible interval of an estimator from an analysis ω_i, and compute

$$e_i^+ = e_i + a_i$$
$$e_i^- = e_i - b_i$$

where a_i and b_i are constants based on bias, imprecision, and variability. Note that this interval need not be symmetric around the actual estimator e_i. The interval of plausible effect is signified as $\left[e_i^-, e_i^+\right]$.

Parsing the plausible interval

This plausible effect interval is to be parsed into two subintervals, one a region of benefit, the other of harm. In order to locate these sub-intervals, knowledge of the value of the statistical estimator's effect that is neutral (i.e., denotes neither benefit nor harm) is required. Define this value of neutral effect as $e_i(0)$. Similarly, let $e_i(b)$ and $e_i(h)$ be the values of the worst possible benefit and the greatest possible harm permitted by the estimator respectively.* Using this notation, then the interval $\left[\min\left(e_i(h), e_i(b)\right), \max\left(e_i(h), e_i(b)\right)\right]$ is the universe of possible values of the estimate.

To compute the plausible interval, consider the case where the greater the benefit, the greater the value of the

---

* The introduction of $e_i(b)$ and $e_i(h)$ is necessary since values of harm need not always be less than values of benefit. For example, if the i^{th} analysis is a total mortality hazard function analysis, then $e_i = 1$ indicates no effect on the time to death , $e_i(h) = \infty$, and $e_i(b) = 0$. Alternatively, if ω_i is an evaluation of changes in mean differences where the greater differences are salubrious, then the value of $e_i = 0$ reflects no mean effect, $e_i(h) = -\infty$, and $e_i(b) = \infty$.

estimator $e_i(b) > e_i(h)$.* We now define the plausible benefit interval $\chi_i^{(b)}$ as;

$$\begin{aligned}\chi_i^{(b)} &= \left[b_i^-, b_i^+\right] \\ &= \left[e_i^-, e_i^+\right] \cap \left[\min\left(e_i(0), e_i(b)\right), \max\left(e_i(0), e_i(b)\right)\right] \\ &= 1_{\left[e_i^-, e_i^+\right]} 1_{\left[e_i(0), e_i(b)\right)} = 1_{\left[b_i^-, b_i^+\right]}\end{aligned}$$

This is the portion of the plausible effect size region that supports benefit. As an example, consider left ventricular ejection fraction. Larger values of left ventricular ejection fraction are considered beneficial *ceteris paribus*; its increases are beneficial and its decreases are harmful. Thus, if the plausible effect region for a change in left ventricular ejection fraction is $[-2, 7]$ and the region of these changes that are beneficial is $\left(e_i(0), e_i(b)\right) = (0, \infty)$, then $\chi_k^{(b)} = [-2, 7] \cap (0, \infty) = (0, 7]$ is the plausible benefit region.

The plausible interval for harm is based on $\left(\min\left(e_i(h), e_i(0)\right), \max\left(e_i(h), e_i(0)\right)\right) = (-\infty, 0)$, and is

$$\begin{aligned}\chi_i^{(h)} &= \left[h_i^-, h_i^+\right] \\ &= \left[e_i^-, e_i^+\right] \cap \left[e_i(h), e_i(0)\right] = 1_{\left[e_i^-, e_i^+\right]} 1_{\left[e_i(h), e_i(0)\right)} = 1_{\left[h_i^-, h_i^+\right]}\end{aligned}$$

* Analagous develop is available for the circumstance for estimators e.g., relative risks, where commonly, the larger the value of the estimator, the greater the harm.

which in this example is $\chi_k^{(h)} = [-2,7] \cap (-\infty,0) = (-2,0]$. We will now construct a function from this parsing of the plausibility function into a plausible interval of benefit and a plausible interval of harm.

What would we like from a benefit function?

If a benefit function is to be compelling, it should increase with increasing benefit, and have that benefit be modulated by the presence of uncertainty. If the plausible interval of benefit does not include the null value for effect* the benefit function should be amplified.

Uncertainly is based on the width of the plausible interval of benefit. The greater the width, the less certain we are of the location of the benefit, and the less convincing the effect size estimate is. In addition, we need the benefit function to be unitless so that it can be easily combined with the benefit functions of other analyses for other analyses $\omega_i \subset A_q = \left\{ \omega_i / q_i \triangleq q \right\}$.

These three features are reflected in the function

$$\mathbf{Y}\left(\chi_i^{(b)}\right) = \mathbf{Y}\left(1_{\left[b_i^-, b_i^+\right]}\right) = r\left[\frac{b_i^- + \frac{b_i^+ + b_i^-}{2}}{\left(b_i^+ - b_i^-\right)}\right] e^{-\rho\left(b_i^+ - b_i^-\right)}$$

This function maps interval of plausible benefit to an assessment of the level of that benefit. Benefit is increased when the plausible interval benefit does not include the null effect value. However, the exponential function discounts the benefit by the benefit interval's length $b_i^+ - b_i^-$. The

* Recall that the null value is the value for which there is no effect, e.g., 1 for a relative risk, or 0 for a mean difference.

separate component in the denominator, $b_i^+ - b_i^-$, makes the denominator unitless. The parameter r is the proportion of the benefit function that makes up the entire plausible interval. Thus, $\mathbf{Y}\left(\chi_i^{(b)}\right)$ penalizes the benefit estimate derived from ω_i for a wide interval, (Figure 1).

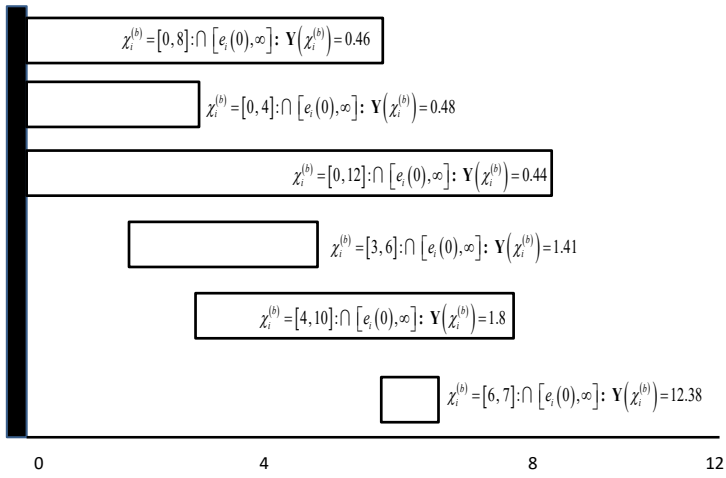

Figure 1. Operation of the benefit function for different benefit plausible regions (r=1)

From Figure 1, the circumstance where $\chi_i^{(b)} = \left[b_i^-, b_i^+\right] = [0,8]$, $\chi_i^{(b)} = \left[b_i^-, b_i^+\right] = [0,4]$ and $\chi_i^{(b)} = \left[b_i^-, b_i^+\right] = [0,12]$ each generate a benefit function value of $\mathbf{Y}\left(\chi_i^{(b)}\right)$ of less than one, reflecting some benefit from this region, but penalizing this strength because their lower bound includes $e_i(0)$, the value of no effect. The

benefit function's value is greater when $b_i^- > e_i^0$, as is the case of the remaining three examples in Figure 1.*

A similar quantity can be computed to assess harm. With the plausible harm interval $\chi_k^{(h)}$ defined as above define $\mathbf{Y}\left(\chi_i^{(h)}\right) = (1-r)\left|\frac{1}{\left(h_i^+ - h_i^-\right)}\left(\frac{h_i^+ + h_i^-}{2} + h_i^+\right)\right| e^{-\rho\left|h_i^+ - h_i^-\right|}$.

Measurable functions of benefit and harm

The notion of benefit and harm can be expanded to an estimate of the size of benefit and the size of harm.

Recall that the plausible benefit interval $\chi_i^{(b)}$ is defined as $1_{\left[b_i^-, b_i^+\right]}$. There are several functions that provide service in assessing the effect of therapy based on that interval. Let $\mathbf{I}$ be the condition where an increase in e_i reflects benefit and $\mathbf{D}$ reflect the circumstance where a decrease reflects benefit. Then one such function is $\mathbf{L}_{\max}\left(\chi_k^{(b)}\right) = \mathbf{L}_{\inf}\left(\chi_k^{(b)}\right)1_{\mathbf{D}} + \mathbf{L}_{\sup}\left(\chi_k^{(b)}\right)1_{\mathbf{I}}$. This represents the assessment of greatest benefit from the plausible interval. It is measurable with respect to (Ω, Σ).

Alternative, one could conservatively estimate benefit as $\mathbf{L}_{\min}\left(\chi_k^{(b)}\right) = \mathbf{L}_{\sup}\left(\chi_k^{(b)}\right)1_{\mathbf{D}} + \mathbf{L}_{\inf}\left(\chi_k^{(b)}\right)1_{\mathbf{I}}$. This serves as an extremely conservative estimate of benefit.

* In reality, it is unlikely that the plauble interval for benefit will not inclue the null value of benefit, due to the cumulative effects of sampling error and precision.

Analogously, $\mathbf{L}_{\max}\left(\chi_k^{(h)}\right) = \mathbf{L}_{\sup}\left(\chi_k^{(h)}\right)\mathbf{1}_{\mathbf{D}} + \mathbf{L}_{\inf}\left(\chi_k^{(h)}\right)\mathbf{1}_{\mathbf{I}}$ is the worst case estimate of harm obtained from $\chi_k^{(h)}$ obtained from the plausible intervals of harm. We will norm this by dividing by the standard error of the estimate.

It is without question that there are other estimates of benefits and harm functions available as discussed in the Limitations chapter. The ones selected here have the advantage of having the required features, and are easy to construct.

Now, let's put it all together.

References

1 Butler J, Anker S, Packer M. Redefining heart failure with a reduced ejection fraction. Journal of American Medical Association. 322:18;761-762.

Putting it all Together

Recall that our purpose for this entire development was to create an analysis platform tailored explicitly (if not exclusively) to clinical research analysis. The required features of this approach were that it 1) include all well designed analyses (without the need for type 1 error considerations) to address a question posed by the clinical investigators, and 2) provide omnibus estimates of effect and harm to address questions in which clinical researchers have the most interest.

The quanta analysis approach, incorporating the concept of duality, is presented as having met these criteria.

We can consider this new method as one of evidence gathering. The clinical researcher, by parsing the plausible intervals of each effect size estimator using duality theory, is identifying all of the evidence from the collection of analyses that support the thesis that the intervention was beneficial.

Performing the analogous evaluation for harm assembles all of the evidence from the germane analyses for harm. This evidence is then accumulated and weighed

using ψ-measure which precisely addresses the question of analysis redundancy.

This measure permits us to integrate much more flexibly than does statistical hypothesis testing, because it both combines the results of many analyses into an omnibus effect and (through ψ-measure) precisely addresses the complications of analysis redundancy (i.e., that different analyses can substantially overlap in the number of observations and the number of variables used in these analyses).

Mathematically, this approach is translated to the following: for each analysis in the set of analyses responsive to the investigator's question q, i.e., $\omega_i \subset A_q = \left\{\omega_i / q_i \triangleq q\right\}$ investigators identify the benefit functions $\left\{\mathbf{Y}\left(\chi_i^{(b)}\right)\right\}$ and the harm functions $\left\{\mathbf{Y}\left(\chi_i^{(h)}\right)\right\}$ The collection of benefit function results $\left\{\mathbf{Y}\left(\chi_i^{(h)}\right)\right\}$ can now be accumulated over all of the analyses $\omega_i \subset A_q = \left\{\omega_i / q_i \triangleq q\right\}$, producing $\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right)$. Since, the analyses are conducted on an overlapping sets of observations and variables, the integral is accumulated with respect to $\psi-$ measure, and $\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) = \int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi$.

But recall from Chapter 19 that since the contribution of each analysis ω_i to this integral is based on the sequence of analyses (i.e., the paths) that the investigators take through the analyses, we developed the summary integral over all possible paths. Thus, we defined integral of a function f with respect of $\psi-$ measure as

$\int_{A_q} f(\omega_i)\, d\psi = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} f(\omega_i) \left[\frac{\psi(B_{[i],s})}{\sum_{\omega_i \subset A} \psi(B_{[i],s})} \right]$. We only have to substitute $\mathbf{Y}\left(\chi_i^{(b)}\right)$ for f in the above expression to see that the integrated measure of benefit $\mathbf{B}_q$, is

$$\mathbf{B}_q = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) \left[\frac{\psi(B_{[i],s})}{\sum_{\omega_i \subset A} \psi(B_{[i],s})} \right].$$

Similarly, the integrated summary of harm, denoted as $\mathbf{H}_q$ can be written as

$$\mathbf{H}_q = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) \left[\frac{\psi(B_{[i],s})}{\sum_{\omega_i \subset A} \psi(B_{[i],s})} \right].$$

These computations beg the calculation of the benefit harm ratio as

$$\mathbf{BHR}_q = \frac{\mathbf{B}_q}{\mathbf{H}_q} = \frac{\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi}{\int_{A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) d\psi}$$

which can be expressed as

$$= \frac{n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) \left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s}\right)} \right]}{n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) \left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s}\right)} \right]} = \frac{\sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) \psi\left(B_{[i],s}\right)}{\sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) \psi\left(B_{[i],s}\right)}$$

$\mathbf{BHR}_q$ is the benefit-harm ratio obtained from a consideration of all of the analyses that the investigators deemed responsive to question q. It is non-negative valued and ranged from zero to infinity. Values less than one suggest more harm than benefit, while values greater than one reflecting the reverse.

Recall also, that we identified an estimate of the least beneficial effect for the intervention based on analysis ω_i, $\mathbf{L}_{\max}\left(\chi_k^{(b)}\right) = \mathbf{L}_{\inf}\left(\chi_k^{(b)}\right)\mathbf{1}_{\mathbf{D}} + \mathbf{L}_{\sup}\left(\chi_k^{(b)}\right)\mathbf{1}_{\mathbf{I}}$. where the plausible benefit interval $\chi_i^{(b)}$ is defined as $\mathbf{1}_{\left[b_i^-, b_i^+\right]}$, $\mathbf{I}$ be the condition where an increase in e_i reflects benefit and $\mathbf{D}$ reflect the circumstance where a decrease reflects benefit. This function can also be integrated over the entire collection of analyses responsive to question q $\omega_i \subset A_q = \left\{\omega_i / q_i \triangleq q\right\}$ as $\Lambda_{q\mathbf{B}}$ where

$$\Lambda_{q\mathbf{B}} = \int_{A_q} \mathbf{L}_{\max}\left(\chi_k^{(b)}\right) d\psi = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{L}_{\max}\left(\chi_k^{(b)}\right) \left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s}\right)} \right].$$

We can think of $\Lambda_{q\mathbf{B}}$ as the normed beneficial effect of the therapy across all analyses responsive to question q, Analogously, we can compute the normed harmful effect across studies, $\Lambda_{q\mathbf{H}}$, as

$$\Lambda_{q\mathbf{H}} = \int_{A_q} \mathbf{L}_{\max}\left(\chi_k^{(h)}\right) d\psi = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{L}_{\max}\left(\chi_k^{(h)}\right) \left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s}\right)} \right].$$

Where, as developed in the previous chapter, $\mathbf{L}_{\max}\left(\chi_k^{(h)}\right) = \mathbf{L}_{\sup}\left(\chi_k^{(h)}\right)\mathbf{1}_{\mathbf{D}} + \mathbf{L}_{\inf}\left(\chi_k^{(h)}\right)\mathbf{1}_{\mathbf{I}}$ is the worst case estimate of harm obtained from $\chi_k^{(h)}$ obtained from the plausible intervals of harm.

The benefit harm ratio, $\mathbf{BHR}_q$, the normed beneficial effect, $\Lambda_{q\mathbf{B}}$ and the normed effect of harm $\Lambda_{q\mathbf{H}}$ are the major products of this work. Together, they permit investigators to determine the strength of the evidence for benefit and harm and compare them using a single value reflecting the integrated finding across all relevant analyses. Similarly, $\Lambda_{q\mathbf{B}}$ and $\Lambda_{q\mathbf{H}}$ provide estimates of the best benefit and worst harm that could be anticipated from exposure to the intervention in the population.

Let's now look at some examples. We will start from the most simple demonstration of quanta theory, and then move to scenarios of increasing complication and controversy. Each of these examples are scenarios that occur in health care research.

In each of the following examples, we assume that the results are research efforts from randomized clinical trials that are prospectively designed and concordantly executed, and that the outcome data have been obtained as precisely as possible. The purpose of these examples is to provide some calibration to duality theory, and then to demonstrate how duality theory is combined with quanta analysis in the circumstance of multiple outcomes.

Example 1: One and only one outcome – no effect size
This is the simplest of examples that demonstrates the data summarization that is provided by duality analysis.

In this circumstance, we have a randomized clinical trial that is assessing the effect of a medication on left ventricular ejection fraction (LVEF). The data shown are the results of sixty hypothetical patients randomized to the medication being tested or to a control therapy. Patients have LVEF measured at baseline and at follow-up. The difference in the change in LVEF between the groups is provided as the effect size.

(Table 1 here)

In this case, the difference in the change in LVEF across the two groups is 0; this estimate's standard error is 4. The plausible effect region (considering both standard error and imprecision is ± 12 absolute EF units.

Using the standard statistical paradigm, the conclusion of this study with its one outcome is that there is no effect of therapy on LVEF.

However, the plausible region (that incorporates the standard error) suggests that there may in fact be a small positive effect in some patients, and a small harmful effect

in other patients.* Duality analysis which considers variability as potential benefit and potential harm, reflects this more nuanced perspective.

With only a single analysis, there is no analysis rotation required (i.e., there is only one analysis path†); thus, this experiment permits us to examine the correspondence between the effect of therapy and duality as reflected in the topside function of Chapter 20.

On a scale of $(-\infty, \infty)$ where positive values reflect benefit and negative values reflect harm, the benefit function's value in this example is 0.22 and the value of harm is -0.22. They are identical since the plausible interval definition is symmetric and the effect size is 0. Not surprisingly, the benefit harm ratio is one, because in this case of one outcome, it reduces to the ratio of the benefit and the harm estimates

The estimate of best benefit is the extreme value of the plausible interval associated with benefit, normed by the standard error. The least benefit estimate is the smallest value of the plausible interval of benefit, also normed. Analogous values of harm are provided. So, we could conclude from this first assessment that both the standard analysis and the quanta analysis support the same

---

* The plausible region reflects, bias (which in this case is zero), imprecision, and the sample variability.

† Our assessments of benefit and harm are based on $\int_{A_q} f(\omega_i)\, d\psi = n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{L}_{\max}\left(\chi_k^{(h)}\right) \left[\frac{\psi\left(B_{[i],s}\right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s}\right)} \right]$ that, in the case of only one analysis reduces to $\int_{A_q} f(\omega_1)\, d\psi = f(\omega_1)$.

conclusion – the absence of evidence that benefit exceeds harm.

The duality analysis comes to the same conclusion, but with a different emphasis.

Example 2: One outcome – moderate effect size

Now, staying with the same research design let's inject some benefit into the finding of this experimental scenario by increasing the estimate of the difference in the change in EF from 0 to 7 (Table 2). This result produces a standard test statistic of $7/4 = 1.75$ which would not meet the 0.05 threshold.

The quanta analysis provides a different assessment. Note that the benefit-harm ratio has increased from 1.00 to 3.30. The conclusion from this one analysis is that the evidence for benefit is greater than the evidence for harm.

This inclination for benefit is also reflected by the best benefit and greatest harm estimates. Note in example 1, the finding of zero effect produce symmetric estimates of best benefit and least harm. In Table 2, that symmetry in the duality analysis disappears.

(Table 2 here)

In the duality analysis, the finding for benefit is greater than that for harm (in absolute value). Thus, the benefit harm ratio has increased from the value of 1 in Table 1 to 3.30 in Table 2. A standard statistical hypothesis test on the Table 2 findings would simply conclude there was no benefit.

However, the duality analysis concludes that while there is evidence of benefit and harm, there is more

evidence of benefit than of harm. The duality analysis also quantifies the best benefit (as a unitless number) of 4.75 and the greatest harm as -1.25.

Note that, in each of these two examples, the least benefit and greatest harm are each zero. This is because the plausible interval crosses zero, the location of null effect. Since the finding is for benefit, the estimate of best benefit is greater than the finding for greatest harm (in absolute value).

Circumstances where the plausible does not cross one, suggesting that there is no evidence that any subject in the study experienced any harm (on the germane outcome measure) is quite rare and is not a circumstances developed in any of the presented analyses in the book.

Example 3: One outcome – large effect size

To consider modifying this example, In Example 3, we increase the effect size from 4 to 9. As we might anticipate at this point, the finding for best benefit has increased with the effect size, and the finding for greatest harm has moved closer to zero. producing a benefit/harm ratio of 18.46. (Table 3)

(Table 3 here)

Example 4: One outcome – overwhelming harm effect
In order to demonstrate that these functions work as expected for a finding of harm an example where the effect size is now negative is provided .

(Table 4 here)

In this circumstance, the difference in the change in LVEF is now negative. Here, the duality analysis reveals there is more evidence for harm than evidence for benefit. The benefit harm ratio is now less than one. In addition, the greatest harm estimate is now -4.25, and the best benefit estimate is only 1.75, also demonstrating a shift to harm.

Conclusions from single outcome examples.
The scenarios of the previous examples in this were to simply provide some calibration for the operation and results of duality analysis in basic clinical trial analyses with a single outcome.

They are not the basis of a supplantation argument – one would not replace statistical hypothesis testing based on these simple comparisons.

We saw that when there was clear benefit, the duality functions demonstrated benefit. They also appropriately revealed that harm was present when the standard statistical estimator suggested harm. When the effect was zero, they demonstrated essentially equivalent benefit and harm.

These findings are what we would expect.

Example 5: Two outcomes with reversed effects
The challenge to now be faced is how the combination of duality theory and the application of measure theory

through quanta analysis functions when there is more than one outcome?

Example 5 examines one of these more complex scenarios. In this prospectively designed, concordantly executed clinical trial, there are two outcomes; the difference in the change in LVEF and the difference in the change in LV end systolic volume (ESV). Neither one of these predominates from a pathophysiologic perspective; there is no natural primary outcome here.

The traditional statistical hypothesis testing domain requires that either one be prospectively chosen as a primary outcome, or that type I error be distributed between the two. A duality/quanta evaluation does not require these actions.

In this circumstance, we have chosen to permit the outcomes to not provide results that demonstrate the same direction of clinical effect (Table 5). The difference in the change in ejection fraction of 7 demonstrates a change in the direction of benefit, yet the change in end systolic volume is in a harmful direction.*

(Table 5 here)

The duality analysis, as we might expect from the previous four examples, recapitulates the standard result; substantially more benefit than harm is seen for the EF outcome, while harm is predominant for the ESV outcome.

Statistical hypothesis testing is of limited use here. If the LVEF outcome was declared as primary, then the study would be viewed from a statistical perspective as positive, a finding that would be contradicted by the ESV findings.

* Increases in end systolic volume, *ceteris paribus*, are harmful.

However, if the ESV outcome were declared as primary, the study would be viewed as demonstrating harm, a result that would ignore the important LVEF finding. If the type I error had been divided in two, with one half apportioned to each of the two outcomes, then neither outcome would be seen as statistically significant, equally unsatisfactory to clinical researchers because that result contradicts the data.

Thus, it would be the artificial process of selecting one or two primary outcomes, and not the data themselves, that determine the study's findings using the traditional statistical hypothesis testing approach. Yet, neither of these conclusions is correct or satisfactory.

Duality theory tells us that while benefit and harm are evident in each of the two outcomes, the magnitude of benefit and harm differ across the two outcomes. We now turn to quanta analysis to accumulate these findings across the two endpoints. This approach permits to integrate the findings of benefit and harm over the two outcomes, and assemble a final number reflecting this combined assessment.

Recall, that since there is more than one outcome, the finding that we take depends on the path or rotation of analysis that is used in the integration. To consider all possibilities, we integrate over all possible paths. (Table 6)

(Table 6 here)

Table 6 demonstrates this path analysis (for orientation, in the rotation, 1 denotes EF, and 2 is ESV) With only two endpoints, there are only two possible rotations or paths; 1, 2 or 2,1.

The first two lines of Table 6 provides the weights or ψ -measures for each analysis in their sequence. For example

$\psi\left(B_{LVEF}\right)=180$ when LVEF is considered first, and $\psi\left(B_{ESV|LVEF}\right)=120$ which is the quantum contribution of the ESV analysis after considering the measure of the EF analysis.

When EF is considered first on the analysis path, its analysis consumes 60% of the total available measure, with ESV making up the additional 40%. The next line provides $\psi\left(B_{ESV}\right)$ and $\psi\left(B_{LVEF|ESV}\right)$, the required quantities when the ESV outcome is the first in the path.

The subsequent data in the table provide the results of the integration as function of the rotation sequence or paths. In each circumstance, both the beneficial findings of the intervention on LVEF and the harmful effects on ESV are considered. However, their contributions depend on the quanta weights. In the first column, the beneficial effect of EF predominates because it is based on 60% of the available measure. However, the **BHR** (Benefit-harm ratio) is not as high as that seen in the duality analysis for EF since it is combined with the finding from ESV.

Similarly, the second column of Table 6 reveals that the harmful effect observed with ESV is modulated by the beneficial effect of EF.

The final summary column averages the duality theory findings over both paths. Note that the benefit harm ratios are substantially modulated from 5.85 observed when LVEF was considered alone and the (0.32) when ESV was considered by itself (Table 5). Overall, both benefit and harm are seen, with a benefit-directional effect. The best and worst benefit and harm cases are provided as well.

In this case statistical hypothesis testing is at best inconclusive, the duality/quanta analysis demonstrates that

there is an overall benefit when considering both LVEF and ESV findings simultaneously.

Example 6: Three outcomes each with small effects
In the next scenario, we, consider the circumstance of clinical trial with three outcomes, LVEF, ESV, and LV end diastolic volume (EDV). In this scenario, there is a modest increase in LVEF with modest decreases in each of ESV and EDV. The standard statistical paradigm suggests that this is a "negative study" (Table 7).

(Table 7 here)

Each of these outcomes reveal that there is an inclination to benefit; however, in the standard statistical paradigm each of these findings would fail the 0.05 *p*-value criteria, much less with any correction for multiplicity.

The duality evaluation for each of these endpoints with its recognition that each of these findings reflects some evidence for benefit and for harm, reveals that the benefit hazard ratio supports benefit for each of the findings

The quanta analysis is more complicated because there are six different paths or sequences to be considered in the assembly of the overall finding. It integrates over the six different paths, and then summarizes (Table 8).

(Table 8 here)

Table 8 demonstrates the findings for each of the six possible paths in the integration. The entries in the upper half of the table demonstrate the quanta weights for each of the endpoints, in each of the six paths (1=LVEF, 2=ESV, 3=EDV). The second half of the table demonstrates the result of the integration for each of the six paths, and then a

summary measure. It is supportive of benefit, a finding that is not surprising since each of the outcomes demonstrated benefit.

Note that each of the six rotations essentially provided the same result – benefit directionality. This is principally because each outcome demonstrated more benefit than harm. Thus, even though the quanta weights differed from rotation to rotation, for each rotation evaluation, there was a finding of benefit and therefore the summary finding across all rotations is one of benefit.

Example 7: Three outcomes with one a disparity

However, suppose one of the outcomes did demonstrated a singular, disparate effect (Table 9).

(Table 9 here)

In this case, the results for LVEF "trends to benefit" (i.e., LVEF moves in the direction of improvement, but do not reach statistical significance). The decrease in ESV is pronounced and also statistically significant. However, the EDV finding moves in the direction of harm.

This circumstance is problematic for clinical investigators, who receive little help from statistical hypothesis testing, a tool too inflexible to manage this complicated set of results coherently and helpfully.

The duality analysis demonstrates the strong ESV finding as a large benefit/hazard ratio of 10.57. The $\mathbf{BHR}_q$ ratios of the other two endpoints are quite small. However, the findings of the quanta evaluations that accumulate the

duality analysis results from each outcome, are unambiguous (Table 10)

(Table 10 here)

The first half of Table 10 is the same as Table 9 since the number of observations and variable have not changed. *The bottom half of Table 10 shows the findings for each rotation. Note that rotations 2,1,3, and 2,3,1, since they begin with ESV, provide the greatest $\mathbf{BHR}_q$ for the overall effect. However, even for these two paths, the $\mathbf{BHR}_q$ is modulated since it is combined with the weaker findings of benefit from LVEF and EDV.

The overall finding of benefit is reduced from the isolated finding of 10.57 to a more moderate findings of 2.78. This is a reasonable finding based on the data.

Statistical hypothesis testing with its requirement to select a single path of analysis (in this case a primary and two secondary endpoints), was much to restrictive since the investigator did not know which single outcome to select as primary.

The quanta analysis avoid this by considering each endpoint in turn as a primary, or first in analysis, each as second, and then each as third, and combining the finding.

With this experience, we can now examine a new take on some outstanding issues in clinical research efforts.

* The quanta computation is based on not just the path of analyses, but also on the number of common observations and variables between the outcomes, and since this has not changed, the quanta are the same.

Quanta analyses and the supremacy of safety

Primum non nocere is the supreme, principal that governs the practice of physicians. It determines their role in the delivery of medical services to individual patients, and sets in place the relationship between health care research and its volunteer subjects.

In classic clinical trial analyses, safety is embedded in the protocol design. It is one of the reasons that we have control groups in clinical trials.[*] It is a rationale for two-sided statistical hypothesis testing in clinical research, and is one of the principal motivations for interim analyses. Each of the prospectively declared outcomes in a clinical trial is assessed for harm as well as for benefit.

Safety is a principal function of the Data Safety and Monitoring Board (DSMB), one of whose tasks is to

[*] One of the arguments that Bradford Hill made to his colleagues in the first clinical trial involving streptomycin was that the presence of a control group would help attribute antibiotic induced adverse effects appropriately to the antibiotic, allowing them to discontinue it if it was unsafe.

monitor the experience of all subjects in a clinical trial, and to react to safety threats by recommending protocol changing, including discontinuing the study.

Institutional Review Boards are a mainstay of local oversight of the conduct of a clinical trials. The United States Federal Food and Drug Administration (FDA) devotes substantial resources to the review of safety data from medication and devices approved in the United States, both prior to and after the intervention's approval.

This all makes sense from public health and epidemiologic perspectives. And certainly, the morality and ethics of placing safety first are beyond reproach.

My purpose here is not to criticize health research workers in their approach to safety in their investigative endeavors, but to instead demonstrate how the philosophy of safety preeminence can play a more central and quantitative role in summarizing the findings of their research efforts.

The safety disconnect in research

The safety experience of clinical research participants predominates in clinical research and contributes to the balanced interpretation of the study. However, there is something of a disconnect in reporting findings from clinical research.

Commonly, the efficacy findings of a research program are detached from the safety findings. For example, efficacy and safety findings appear in different portions of the results section of a manuscript, with efficacy findings appearing earlier than those of safety. There is no attempt to assimilate the two in the results section; this occurs noetically in the discussion section.

While clinical researchers have become accustomed to this, there is no scientific reason why this must be so,

especially since both safety and efficacy must ultimately be integrated in the end, if only cerebrally to have an assessment of risk and benefit.

However, the quantitative separation between efficacy findings and safety findings is convenient because there are no widely used statistical methods through which this integration can take place.

Safety findings and type I error

By the current standard, family wise type I error, so carefully parsed for the collection of prospectively declared efficacy endpoints, is not included in the assessment of safety. In general, safety testing is at the nominal 0.05 alpha level. There are typically no correction for multiplicity for safety evaluations. Essentially type I error for safety endpoints is treated much like that for secondary efficacy endpoints.

This traditional approach is not without its own rationale.* Sharing the type I error apportioned for efficacy with safety analyses decreases the type I error for each assessment. The results of hypothesis testing regarding safety outcomes would need to be more extreme in order to meet the lower alpha level. In addition, strictly requiring the type I error level control for safety findings is not consistent with a "safety first" philosophy.

In addition, we must keep in mind that primary efficacy endpoints are assessed for harm as well as for efficacy, an assessment that occurs under full type I error control, so it

* The argument that safety outcomes are so important that they transcend type I error considerations is vacated by the observation that statistical hypothesis testing is used to assess differences in safety endpoints across therapy groups.

is not as though all safety evaluations across therapy groups are nominally interpreted.

Finally, the remaining lower alpha level for the efficacy endpoints would increase the sample size of the study considerably, because error rates, just like event rates and efficacy, are powerful drivers of the numbers of patients required for the study. This is admittedly a practical, not a safety consideration.

What else can we do?

These rationale are fine, but they do not overcome the argument for a formal and quantitative integration of safety and efficacy findings in clinical research efforts.

The actual explanation for the absence of such integration is that we have no reliable way in biostatistics (and certainly no reliable way using statistical hypothesis testing) to conduct this integration. Since safety is an important occurrence in the use of the study, then a quantitative synthesis of safety and efficacy findings could be useful. However, while desired, it has been unavailable.

However, our duality and quanta analysis approach can accommodate this request. From its perspective in the accumulation of analyses responsive to question analysis result integration responsive to question q, it does not matter whether the analyses that are to be integrated are efficacy analyses or safety analyses, i.e., $\omega_i \subset A_q = \left\{ \omega_i / q_i \triangleq q \right\}$ can include safety analyses are well. We simply need to determine the analysis path.

In addition, the use of the duality principal allows us to assess the evidence of benefit and evidence of harm from the safety outcomes, just as for efficacy, accumulating them using the exact same procedures that we developed for the their accumulation of efficacy and harm from efficacy

outcomes. Thus, these safety analyses are passed through the benefit and harm functions $\mathbf{Y}\left(\chi_i^{(b)}\right)$, and $\mathbf{Y}\left(\chi_i^{(h)}\right)$ and are gathered up with these functions that are evaluated for efficacy analysis in $\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi$ and $\int_{A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) d\psi$ respectively.* No new theory is required; we simply need to expand the same processes of parsing, channeling, and accumulation into and through the safety assessments.

Example: Heart failure therapy and creatinine

For example, consider a clinical trial designed to assess the effect of a new therapy for heart failure. This randomized and double blinded clinical trial will assess the effect of the therapy on well-established outcomes, e.g., left ventricular ejection fraction (EF), left ventricular end systolic volume (ESV) and left ventricular end diastolic volume (EDV). However, since the intervention is anticipated to have a nephrotoxic effect, the difference in the change in creatinine levels across the two groups is also measured.

The results in accordance with the standard analysis and the duality analysis reveal more evidence for benefit than harm with the three efficacy outcomes, and as expected, there is an increase in the mean creatinine level associated with therapy (Table 1).

(Table 1 here)

However, we see that unlike the three efficacy evaluations, there is greater evidence for harm then benefit

* This is also true for the conputation of the estimate of benefit $\Lambda_{q\mathbf{B}}$ and harm $\Lambda_{q\mathbf{H}}$.

for changes in serum creatinine levels. This is where the analysis typically ends in the standard paradigm which is confirmed by the duality analysis.

However, we are in a position to ask what does the integrated result look like. In order to answer this, the one remaining question that we must answer is where in the path the safety analysis should appear. If we place it either before (Safety First) or after (Safety Last) the three efficacy analyses (Table 2).

(Table 2 here)

Table 2 provides the rotation summary. In the first column, we have the result by including safety last in the path. The implication of this last position is that this safety analysis has the smallest ψ -measure. However, even there, we see that there has been an important reduction in the benefit risk ratio despite the benefit findings from the three efficacy endpoints visible from Table 1.

However, when safety is considered first in the quanta analysis path, there is a marked decrease in the benefit/hazard ratio even further, tipping it to harm.

Summary

Thus, not only can safety evaluations be folded into an integrated summary of the result of the study, but it is possible to consider their impact first, before consideration of the impact of the primary endpoints, which is wholly consistent with a safety first mentality. Such a procedure will likely to change the assessment of some clinical trials.

Managing correlation between variables

The development of duality theory and quantum analysis has thus far been silent on the issue of correlation. As clinical investigators understand, while it is reasonable to assume that the individual participants of a clinical research effort make independent contributions to an analysis, the variables utilized in an analysis are commonly correlated among themselves. It stands to reason that this dependency will affect the contribution of a collection of variables to the ψ -measure of an analysis.

Recall that the ψ -measure of an analysis ω_i is simply $\psi(\omega_i) = n_i v_i$ where n_i is the number of observations and v_i the number of variables used in ω_i. In a single analysis, it is possible that v_i could be large. This of course would substantially increase the ψ -measure.

However, if these variables in the analysis of ω_i are correlated, then they are (colloquially expressed) "measuring the same thing". This redundancy in what they measure should decrease the measure of the analysis. This thought process is the motivation for taking a reduction

action on the ψ -measure of this analysis by taking the correlation amongst its variables into account.

This is a technical chapter. For those who are willing to accept that correlation can be incorporated into ψ -measure, then please proceed to the chapter on exploratory evaluations. Those who have a background in multivariable analysis in general and determinants in particular, or are interested in seeing its possible application to ψ -measure, please feel free to charge forward.

Proposed formulation using determinants

A formulation for the measure of an analysis that incorporates this dependency is $\psi(\omega_i) = n_i v_i \left|\mathbf{R}\left(\underline{v}_i\right)\right|$ where $\left|\mathbf{R}\left(\underline{v}_i\right)\right|$ is the determinant* of the v_i by v_i correlation matrix of all of the variables used in the analysis ω_i.

In order to get a sense for the mechanics of this formulation, consider that, when $v_i = 2$, $\psi(\omega_i) = n_i v_i \left|\mathbf{R}\left(v_i\right)\right| = n_i 2 \begin{vmatrix} 1 & r \\ r & 1 \end{vmatrix} = n_i 2\left(1 - r^2\right)$. Thus, when the correlation is minimal, then each of the two variables substantially contribute to the analysis' measure. When the

* The determinant is computed from a matrix of numbers. It is a measure of the degree to which linear combinations of some of the columns can be used to reproduce other columns. When the matrix is not just numbers but the correlation coefficients between variables (i.e., a correlation matrix), then the determination of that correlation matrix is a reflection of redundancy in the system. In a system of *n* variables, the largest the determination can be is 1, reflecting no dependency, and the smallest the determinant can be is zero, reflecting complete redundancy (i.e., at least one of the variables can be reproduced by adding multiples of some of the other variables).

dependency is high, the contribution of the two variables to the measure of the analysis diminishes due to fact that to some degree, these two variables measure the same phenomenon.

Correlations and unions of analyses

Recall that the ψ -measure of the union of two analyses ω_i and ω_j is $\psi\left(\omega_i \cup \omega_j\right) = \psi\left(\omega_i\right) + \psi\left(\omega_j\right) - \psi\left(\omega_i \cap \omega_j\right)$. In order to formulate the role of correlation within the measure of the union of analyses, we must consider how it is embedded in the measure of their intersection.

Our consideration of correlation should reflect not just the correlation of variables used in the same analysis, but also the more "distant" correlation of variables that are different between the two analyses yet nevertheless correlated with each other. This latter circumstance should permit these "inter-analysis" correlations to contribute to the measure of the overlap between the analyses.

We can proceed by asserting that, if $\psi\left(\omega_i\right) = n_i v_i$, and $\psi\left(\omega_j\right) = n_j v_j$, then we can assemble the measure of the intersection

$$\psi_{ijk\ldots m} = \psi\left(\omega_i \cap \omega_j \cap \omega_k \cap \ldots \cap \omega_m\right) = n_{ij} v_{ij}\left(d + \left|\mathbf{R}_u\right|\right).$$

Here d is the number of variables that are in common between the two analyses and u is the number of variables that are not in common but may be correlated. This permits correlated variables that are not common across the analyses to contribute to the measure of the intersection through their correlations.

The following are examples of how these formulations operate.

Example 1 – Regression analysis families

Consider the circumstance in which a collection of straight line regression analyses are carried out in a clinical research effort. The dependent variable is the same for each analysis; the change in left ventricular ejection fraction from baseline to follow-up. This change is the response variable that is assessed against a baseline value of 1 of *m* phenotypes.

Thus, in this collection of analyses, there are *m* different regression analyses, each a function of three variables (two variables determine the change value and 1 variable is the phenotype. Each analysis has the same number of participants *n*. The goal is to compute the measure of $\psi\left(\bigcup_{i=1}^{m}\omega_i\right)$. As is our process, define $\{B_i\}$ as the collection of disjoint sets, such that $\bigcup_{i=1}^{m}B_i = \bigcup_{i=1}^{m}\omega_i$, and $\psi\left(\bigcup_{i=1}^{m}\omega_i\right) = \sum_{i=1}^{m}\psi\left(B_i\right)$.

Let's first presume that the phenotype variables are independent of each other. In this circumstance, calculation for $\psi\left(B_j\right)$, $j = 1...m$ is straightforward because $\psi\left(\bigcap_{i=1}^{m}\omega_i\right) = 2n$ (each analysis contains the same number of participants and has two variables in common). Thus,

$$\psi\left(B_1\right)=3n.$$
$$\psi\left(B_2\right)=3n-2n=n.$$
$$\psi\left(B_3\right)=3n-2n-2n+2n=n.$$
$$\psi\left(B_4\right)=4n-2n-2n-2n+2n+2n+2n-2n=n.$$
...

In general, $\psi\left(B_j\right)=3n\mathbf{1}_{j=1}+n\mathbf{1}_{1<j\leq m}$ and

$$\psi\left(\bigcup_{i=1}^{m}\omega_i\right)=\sum_{i=1}^{m}\psi\left(B_i\right)=3n+\left(m-1\right)\left(n\right)=n\left(m+2\right).$$

However, in reality, the *m* phenotype variables are correlated. The implication of this correlation is that the measure of any intersection among a collection of the $\omega_i, i=1...m$ analyses written now as $\psi\left(\bigcap_{i=1}^{m}\omega_i\right)=2n$ will be a misrepresentation of the intersection's measure. Following the development of the methods section for dependence, write

$$\begin{aligned}\psi\left(B_2\right)&=\psi\left(\omega_2\right)-\psi\left(\omega_1\cap\omega_2\right)\\&=3n-n\left(d+\left(1-\left|\mathbf{R}_{12}\right|\right)\right)=3n-n\left(2+\left(1-\left|\mathbf{R}_{12}\right|\right)\right).\end{aligned}$$

The measure of the intersection of the first two analyses ω_1 and ω_2 $\psi\left(\omega_1\cap\omega_2\right)=n\left(d+\left(1-\left|\mathbf{R}_{12}\right|\right)\right)$ reflects the observation that 1) the same number of *n* participants are included in each regression analysis and 2) the number of variables in common across the two analyses is $d+\left(1-\left|\mathbf{R}_{12}\right|\right)$, where $d=2$ reflects that in each analysis, the

same two variables determine the responder analysis. $\left|\mathbf{R}_{12}\right|$ is the determinant of the 2 x 2 correlation matrix for phenotypes 1 and 2. Since $\left|\mathbf{R}_{12}\right| = \begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix} = 1 - r_{12}^2$, find

$$\psi\left(\omega_1 \cap \omega_2\right) = n\left(d + \left(1 - \left(1 - r_{12}^2\right)\right)\right) = n\left(2 + r_{12}^2\right).$$

When the correlation is zero, then the number of common variables is 2, and the solution defaults to the earlier case in this example. However, as the correlation increases, the greater is the propensity of two phenotype variables to measure the same feature, and the greater their contribution is to the measure of the two analyses' intersection. Thus,

$$\psi\left(B_2\right) = \psi\left(\omega_2\right) - \psi\left(\omega_1 \cap \omega_2\right) = 3n - n\left(2 + r_{12}^2\right) = n\left(1 - r_{12}^2\right).$$

Note that if $r_{12} = 1,$ then $\psi\left(B_2\right) = 0.$ This follows from the observation that if the correlation is 1, then the two phenotypes essentially measure the same relationship, and the incremental value of the regression on the second phenotype is zero after assessing the first regression.

Continuing on, find

$$\begin{aligned}
\psi\left(B_3\right) &= \psi\left(\omega_3\right) - \psi\left(\omega_1 \cap \omega_3\right) - \psi\left(\omega_2 \cap \omega_3\right) + \psi\left(\omega_1 \cap \omega_2 \cap \omega_3\right) \\
&= 3n - n\left(2 + r_{13}^2\right) - n\left(2 + r_{23}^2\right) \\
&+ n\left(2 + \left(1 - \left|\mathbf{R}_{12}\right|\right) + \left(1 - \left|\mathbf{R}_{13}\right|\right) + \left(1 - \left|\mathbf{R}_{23}\right|\right)\right) \\
&= 3n - n\left(2 + r_{13}^2\right) - n\left(2 + r_{23}^2\right) + n\left(2 + r_{12}^2 + r_{13}^2 + r_{23}^2\right) \\
&= n\left(1 + r_{12}^2\right).
\end{aligned}$$

The calculation proceeds for $i > 3$.

Summary

The procedure laid out in this chapter is certainly not the only way to incorporate correlation into quanta analysis. However, it is consistent with what ψ -measure is attempting to assess, relatively easy to compute, and is smoothly integrated into the calculation of the quanta measures.

Incorporating Exploratory Analyses

The flexibility of the duality/quanta analysis lies principally in the quanta component. The implication of ψ - measure means that any collection of analyses that can be assembled from our set theory tools of unions, intersections and complements can be measured.

This is the principal advantage of requiring ψ to be a measure. Thus, it can measure combinations of safety and efficacy endpoints, measure the various paths through efficacy outcomes, and permit us to disconnect from the notion of declaring a prospectively described outcome as primary one simply from statistical considerations.[*]

However, ψ -measure also provides freedom that we may not know how to handle. This brings us to exploratory analyses.

Exactly what are exploratory analyses?
Exploratory analyses are evaluations conducted in health care research that were not prospectively declared.[†]

[*] That is, for the assignment of type I error prospectively.
[†] There is very little written about the theory of exploratory analyses beyond statements about the problems such analyses can cause, e.g.,

Commonly described as retrospective or *post hoc* appraisals, these assessments follow no prospective plan and have no protocol.

The motivation for the analyses might be findings identified in other studies or a *de novo* observation from the investigators' own ongoing research. Because these *post hoc* evaluations appear to answer questions that the investigator did not think to ask prospectively, exploratory analyses can be engaging and even exciting.

There are many reasons to conduct unplanned, hypothesis-generating analyses in clinical trials. In some circumstances, during the course of a study, a new outcome measure is determined to be of value in a second trial.

With this new outcome in hand, the research community is particularly interested in how that new outcome measure performs in the current trial. The investigators may in fact be compelled to report this outcome, even though it was not part of the prospective plan.

In other circumstances, a journal reviewer or the editor may desire to see a particular analysis. The reason this is typically well motivated, and the clinical investigators, anxious to satisfy these arbiters of publication, will provide the analysis, which may or may not be published.

In fact, it is not uncommon for the United States Federal Food and Drug Administration (FDA), when

---

Nicenboim B, Vasishth S, Engelmann F, Suckow K.Exploratory and Confirmatory Analyses in Sentence Processing: A Case Study of Number Interference in German. Cogn Sci. 2018 Jun;42 Suppl 4:1075-1100. doi: 10.1111/cogs.12589. Epub 2018 Feb 7.

Much of this section if from my article Moyé L**.** What Can We Do About Exploratory Analyses in Clinical Trials?Contemp Clin Trials. 2015 Sep 18. pii: S1551-7144(15)30088-4. doi: 10.1016/j.cct.2015.09.012. [Epub ahead of print] PMID: 26390962

reviewing the voluminous filings of pharmaceutical companies that are supplied in support of a product, to ask for additional analyses from pivotal Phase III studies that were not prospectively declared by the investigators. Of course, these requests are granted. The National Institutes of Health also engages in these analyses. [1].

Also, there is the investigational motivation to examine all of the data at hand in new combinations to see if a new facet of the disease can be examined.

The use of exploratory analyses is not surprising. Unanticipated findings play an undeniable role in science. Radiation was not found because it was sought, but because the researchers stumbled across it. Minoxidil and sildenafil are examples of medications that were designed for one purpose, but perspicacious investigators identified unanticipated new effects that opened the door to new indications*.

It is therefore no surprise that, exploratory analyses are critical in first-in-human studies. In these cases, there is very little information to determine the universe of effects of a biologic agent or small molecule. Thus, these early protocols tend to be especially restrictive. The observed effects generated by a biologic or small molecule – by the very nature of the poor state of *a priori* knowledge – are likely to be a surprise.

An example is determination of optimum cell preparations for cell therapy clinical trials.[2]. There is particular interest in publishing exploratory analyses in

* Both minoxidil and sidenifil were developed as antihypertensives. Minoxidil was discovered to be one of the first effective medical treatments for alopecia, and sidenifil evolved into a treatment for erectile dysfunction.

oncology, with the exploratory components clearly marked. [3]

Therefore, Phase I studies have and should have a heavy exploratory component since the earlier the study, the less is known, and the more is identified *post hoc*. There is little question that such findings must be confirmed in a standard confirmatory evaluation. However, we must also acknowledge that if the question was not first raised by the exploratory finding, the confirmation would not have been forthcoming.

The problem with exploratory analyses

However, there are difficulties with exploratory analyses. The principal difficulty is not with the data evaluation itself, but with the interpretation of the data.

Early clinical trial experience did not differentiate prospectively declared evaluations from non-prospectively declared assessments. If the *p*-value was less than 0.05, the result was considered not just statistically significant, but valid and reliable.

In Chapter 2, we saw the problems with this approach. The MRFIT , program, as well as INVEST, ELITE, PRAISE, and the US Carvedilol programs are just some of the examples that demonstrated that there were hazards with non-prospectively declared outcomes. The source of these hazards is practical, and theoretical.

Logistical concerns

The principal justification given for the unreliability of an exploratory analysis is the effect of the absence of prospective planning on the precision of the exploratory estimators of effect size. We will call this the *logistical rationale*.[4]

For example, if investigators wish to conduct a clinical trial on the effect of an intervention on heart muscle perfusion, they are obligated to ensure superior quality and high precision images for the endpoint measures, e.g., identifying a core laboratory. These trial design controls reduce endpoint variability and *ceteris paribus* increase power.

However, should these same investigators observe at the study's conclusion a treatment attributable benefit for coronary artery disease death, they will be hard pressed to defend the reliability of this unanticipated finding. The absence of its prospective declaration meant that there was no opportunity to organize resources for its reliable estimation.

For example, without prior definition of coronary artery disease death, there could be no *a priori* structure in place for the formal collection of death records and no endpoint committee of specialists to adjudicate findings. In addition, the analysis suffered from an absence of prospective statistical planning that, had it been present would have produced both informative power computations and the minimum number of deaths required to draw a conclusion with some statistical regularity.

Theoretical concerns

A second concern is more theoretical. It is the random selection of the analyses. Prospective outcomes are chosen based on knowledge of mechanism of action, reliable data, and the availability of precise outcome estimators, before the data are collected.

Exploratory analyses, on the other hand, are essentially chosen for publication randomly. They are selected due to their unanticipated small *p*-value. When chosen in this

fashion, we have learned that these analyses are not reliable (i.e., they are unable to be reproduced).

Does that mean they should not be published?
Exploratory evaluations are commonly the first databased view of the future. Today's exploratory research can be tomorrow's new confirmatory outcome. Thus, exploratory evaluations should have a place in the literature.

They just have to be clearly labeled as exploratory and not permitted to replace the findings of the prospectively declared outcomes.

A problem is that it is difficult to have the results of exploratory analyses published in some areas of research whether they are clearly labeled as such or not. It is as though, to the editors, the very moniker "exploratory" is too controversial. The opisthotonic reaction of journals to exploratory analyses that are clearly labeled as requiring replication does a disservice to the medical research community.[*]

The inclusion of exploratory analyses clearly labeled as such raises type I error concerns as well. In addition to concerns about the interpretation of *p*-values in statistically underpowered environments, one cannot prospectively apportion type I error in non-prespecified environments.[†]

---

[*] There are counter examples to this. For example, there is interest in publishing exploratory analyses in oncology as previously mentioned in this chapter. In addition, there is substantial interest in the behavioral sciences in understanding exploratory factor analysis.

[†] While one could allocate, for example, 15% of the available type I error to exploratory analyses, the impact on the sample size is not inconsiderable and would be judged to big a price to pay to assess analyses unknown to the clinical investigators at the start of the study.

Duality and quanta analyses in exploratory evaluations

The advantage that quanta analysis holds is that exploratory analyses can be incorporated into the omnibus benefit-risk ratio while discounting their contributions for the reasons outlined above.

To be clear, we will only consider exploratory analyses that are well conducted as worthy of inclusion. Sloppy evaluations, low quality data, and imprecise definitions have no place in any study, whether it be based on statistical hypothesis testing or duality/quanta analyses.

With these exclusions, there are exploratory outcomes that are measured precisely. For example, the variables that are produced from cardiac MR imaging provide volumes of data and variables about which little are known. All of the MR variables, be they prospectively declared or not, are measured with the highest possible precision.

Consider a clinical trial with two treatment groups and three outcomes; one primary and two secondary. They are the difference in the change in left ventricular ejection fraction (primary), the difference in the change in left ventricular end systolic volume (ESV) (secondary), and difference in the change in left ventricular end diastolic volume (EDV) (secondary).

The data demonstrate no clinically important change in these outcomes. However, there is one MR based exploratory outcomes measured at high precision; MR1.(Table 1)

(Table 1 here)

Note that the duality analysis demonstrates evidence for more (but not much more) benefit than harm for each of EF, ESV, and EDV. However, it does note substantially

more evidence for benefit from the exploratory endpoint. However, for the reasons that have been provided in this chapter, considering the MR exploratory analysis on par with the prospectively declared non-MR analyses is problematic. However, the exploratory evaluation can occur last in the path analysis (Table 2).

(Table 2 here)

Only two rotations are required since the primary outcome always appears first, the exploratory outcome is last, and there were only two secondary outcomes. The result reveals that there was a small increase in the benefit/hazard ratio due to the exploratory outcome. However, the uptick was modest given its position as last in the analysis path, where it retained over 22% of the total $\psi -$ measure.

Clearly, the more primary and secondary endpoints there are *ceteris paribus*, the less information there is for the exploratory outcome.

The purpose of this chapter was to demonstrate the flexibility of the duality/quanta approach through its ability to readily absorb exploratory analyses. However, this facility should not dominate concerns over exploratory analyses. Many such evaluations are not worthy of further consideration due to poor or missing data and incompletely conceived analysis plans. Mathematics cannot adumbrate these grave matters.

References

1 . Kaltenthaler E, Carroll C, Hill-McManus D, Scope A, Holmes M, Rice S, Rose M, Tappenden P, Woolacott N.The use of exploratory analyses within the National Institute for Health and Care Excellence single technology appraisal process: an evaluation and qualitative analysis. Health Technol Assess. 2016 Apr;20(26):1-48. doi: 10.3310/hta20260.

2 . Landin AM, Hare JM. The quest for a successful cell-based therapeutic approach for heart failure.Eur Heart J. 2017 Jan 10. pii: ehw626. doi: 10.1093/ eurheartj/ ehw626. [Epub ahead of print]

3 Tahara M, Schlumberger M, Elisei R, Habra MA, Kiyota N, Paschke R, Dutcus CE, Hihara T, McGrath S, Matijevic M, Kadowaki T, Funahashi Y, Sherman S.Exploratory analysis of biomarkers associated with clinical outcomes from the study of lenvatinib in differentiated cancer of the thyroid. Eur J Cancer. 2017 Apr;75:213-221. doi: 10.1016/j.ejca.2017.01.013. Epub 2017 Feb 24.

4 Moyé L. What Can We Do About Exploratory Analyses in Clinical Trials?Contemp Clin Trials. 2015 Sep 18. pii: S1551-7144(15)30088-4. doi: 10.1016/j.cct.2015.09.012. [Epub ahead of print] PMID: 26390962

Contributions of Other Measurable Functions

The mathematical concept of measurability plays an central role in this book's development. For example, we have measurable functions $\mathbf{Y}$ and $\mathbf{L}$ through which the parsed plausible intervals for benefit and harm are channeled. We have also established a formal measure ψ to help us manage precisely the redundancy in analyses, permitting to accumulate these functions over regions of analyses e.g., $\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi$.

However, there are additional possible uses for measurable functions in our application of duality theory to health care research.

Recall that a measurable function must meet three criteria 1) it must be real-valued, 2) it must be nonnegative, and 3) it must generate its numerical value based on an inspection of the properties of ω_i.

Recall that any particular analyses ω_i from our Ω has many different properties (Chapter 11). Thus far we have utilized only the plausible interval, but other properties of the analysis should and can influence the impact of an

analysis. Since the use of these can be the basis of a measurable function on (Ω, Σ), we can create helpful measurable functions that permit us to modulate or amplify the influence of a particular analysis on the assessment of benefit and harm.

One such influence is that of the characteristics of an analysis. Setting statistical hypothesis testing aside, there are other features of analyses that we quite correctly consider when assessing the impact of an analysis.

A critical feature is whether the analysis is prospective or retrospective (exploratory). Other features have to do with the presence of a contemporary control group, the presence of randomization, and a degree of blinding.* When these features are present, we provide more weight to the analysis.

How would this work mathematically?

Recall that all of our work have revolved around the concept of $\int_{\omega_i \subset A_q} f(\omega_i) d\psi$ which is an accumulation of the function $f(\omega_i)$ (which for us has been a benefit function or a harm function) over all analyses responses to the research q, expressed as $\omega_i \subset A_q = \{\omega_i / q_i \triangleq q\}$.

If we wish to modify or amplify the function $f(\omega_i)$ without fundamentally changing the function f we can simply develop another measureable function $\mathbf{M}(\omega_i)$ ($\mathbf{M}$ for "methodology"). This function will place a highest value on analyses that have the strongest methodology.

* These last three features are hallmarks of clinical trials.

We can implement this function quite simply. One way is to permit the set function $m_j(\omega_i)$ to be the j^{th} methodologic property of analysis ω_i and is assigned a value based on the presence of that property. These properties are quite easily enunciated (Table 1).

Table 1. Defining methodologic features for $\mathbf{M}(\omega_i)$.

j	**Methodology Characteristic**	**Value**
1	prospective analysis	3
2	use of a core lab	2
3	pilot study control group	1
4	pilot study randomization	1
5	pilot study- blinded	1
6	pivital study control group	2
7	pivoital study randomization	2
8	pivital study- blinded	2
9	adequate sample size	2

Table 1 provides some possible values for the analysis characteristics. For example if the analysis is prospectively designed from a pilot study which had a control group but was neither randomized nor blinded, then $\sum_{j=1}^{9} m_j(\omega_i) = 4$.

We can then define

$$\mathbf{M}\left(\omega_i\right)=\frac{2e^{\sum_{j=1}^{m} m_j(\omega_i)}}{1+e^{\sum_{j=1}^{m} m_j(\omega_i)}}-1.$$

This is function that is trapped between 0 and 1. For analyses which have relatively weak methodology, $\mathbf{M}\left(\omega_i\right)$ is close to zero. Analyses with the strongest methodologies produce values $\mathbf{M}\left(\omega_i\right)$ close to one. Our benefit and hazard integrals would then be written as

$$\mathbf{B}_q = \int_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)}\right)\mathbf{M}\left(\omega_i\right)d\psi$$

$$\mathbf{H}_q = \int_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(h)}\right)\mathbf{M}\left(\omega_i\right)d\psi$$

The multiplication within the integral permits us to modulate the effect of the benefit or hazard function based on the methodology that is utilized by the analysis. This feature operates independently of the path analysis. An analysis that occurs early on the analysis path, but is crippled by its weak methodology (for example, the absence of blinding) will have reduced influence on the benefit and harm interval.

This feature provides the clinical investigator the second of two control features that manage the impact of an analysis.

Limitations

Is Duality and quanta analyses ready for prime time?
No.

It is a fine idea, but it has limitations and contains arbitrary decisions that must be more closely examined with the goal of improvement. No doubt you have identified your own such set of concerns. Here is mine.

The quanta measure

As iterated in this text, the quanta measure is the key to this tool's flexibility. However, the decision that $\psi(\omega_i) = n_i v_i$ was arbitrary. It was selected because it a simple quantity directed related to the data and when investigated, proved to be a measure and is easily computed.

However, this is not the only such measure – in fact there are uncountable many measures available. The use of $\psi(\omega_i)$ is useful, but it must be seen as only a starting point in this new field of exploration.

The most important new ingredient is not the specific formulation of $\psi(\omega_i)$, but the use of a formal measure that permits wide latitude in computing the value of a union of analyses. This is the key new ingredient. What the best

measure might be is wholly up to debate, discussion, and improvement.

The topside function is not optimal

Parsing the plausible interval into one of benefit and one of harm is simple. How one incorporates these fragments into a function is complicated, and I confess that this part of the project took far more time than I anticipated.

Of one thing I am certain. This topside function can be improved. From my perspective, the most important features that it must have are

1- It provide a unitless measure of benefit and one of harm.
2- The wider the plausible interval, the less emphasis the actual estimator of benefit and harm receive.
3- The further the lower bounds of the plausible intervals for benefit and harm are from the null value (i.e., that value which indicates no effect), the greater the strength of the finding.

In addition, the assumption that the region of plausible values must include the possibility of benefit and of harm in research (that is, they must cover the null value of the estimate), is valid I believe, but it requires reexamination.

My experience informs me that in studies that demonstrate even overwhelming benefit, there are individuals who receive the exposure who are harmed by it, either by an outcome (e.g., diastolic blood pressure) moving in the wrong direction or the occurrence of a safety event reliably attributed to the exposure. This was my motivation for assuming the plausible interval must include the null value.

The investigators have full freedom in choosing the plausible intervals, but it is recommended that they be wide, not narrow. Its goal is not to include only probable effects, but those that are unlikely but possible, since the improbable commonly occurs in health care.

The methodology function is not unique

Chapter 24 introduced a measureable function $\mathbf{M}(\omega_i)$ on our usual (Ω, Σ) regions of analysis. This function was designed as a conduit for the impact of the methodologic rigor of the research effort on the functions of benefit and harm produced by that effort's estimators. $\mathbf{M}(\omega_i)$ was defined as quantitative metric based on the characteristics of the research method. This is clearly not the only definition of such a function. There are other measures of the quality of a research effort. In addition, there are other functional forms besides $\mathbf{M}(\omega_i) = \frac{2e^{\sum_{j=1}^{m} m_j(\omega_i)}}{1+e^{\sum_{j=1}^{m} m_j(\omega_i)}} - 1$. These should be developed and examined.

There is no sample size formula

This is an important concern. The determination of the number of subjects that there should be in a research effort is a critical practical consideration for researchers because it is an important driver of the logistics (how many recruiting centers and co-investigators are required) and the likelihood of funding. Its absence is a crucial impediment to the implementation of duality/quanta analyses as the only evaluation of the contribution of the research effort.

However, this is just a technical issue. Our mathematical development demonstrates that the quanta contribution to our measurement of, for example benefit from the collection of questions $\omega_i \subset A_q = \left\{ \omega_i / q_i \triangleq q \right\}$ is

$$\mathbf{B}_q = \int_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)} \right) \mathbf{M}\left(\omega_i \right) d\psi$$

$$= n_s^{-1} \sum_{s=1}^{n_s} \sum_{\omega_i \subset A_q} \mathbf{Y}\left(\chi_i^{(b)} \right) \mathbf{M}\left(\omega_i \right) \left[\frac{\psi\left(B_{[i],s} \right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s} \right)} \right]$$

The quanta component, $\dfrac{\psi\left(B_{[i],s} \right)}{\sum_{\omega_i \subset A} \psi\left(B_{[i],s} \right)}$, since it represents a collection of percentages is sample size independent, as is $\mathbf{M}\left(\omega_i \right)$ which is a measure of methodologic rigor. The impact of the sample size is therefore on the contribution of the benefit function, which is itself a function of the plausible interval for benefit. Recall that this plausible interval is a function of accuracy, precision, and bias of the statistical estimator. Thus, we simply need a sample size that controls these three features across each of the estimators in the set $\omega_i \subset A_q = \left\{ \omega_i / q_i \triangleq q \right\}$. This must be developed, but it also should be tractable.

Lack of Independent Confirmation

A disadvantage of this development is that it was conducted by one and only one person – the author. Assessments by independent researchers either in recapitulating my own efforts, or through their own

derivations, regenerating my conclusions is essential. The *raison d'etre* for this book is to call for that process while at the same time producing my work for this required review.

A real work test is lacking

Duality and quanta measure must be put to the actual test, i.e., actual clinical research data should be run through it. This will improve the robustness of the software, and also provide some calibration of the unitless results. I have no access to clinical trial data at this point so could not provide this essential calibration myself. However, this research effort must be put to test with real data.

Conclusions – Queen Anne's Decree

Are we better off with a new process for mathematically assessing the impact of health care research? My answer is "Yes". I have provided one in this book.

Duality/quanta analysis is not the only alternative. Perhaps it is not the best alternative. But it does not have the well-known panoply of weaknesses that afflicts statistical hypothesis testing that were reviewed in chapter two's germane discussion.

Quanta analysis

Duality/quanta analysis addresses the straightforward, central, almost Reaganesque question* "Are my patients better off being exposed to the intervention? Through a process of parsing, channeling and accumulating, it gathers and weighs the evidence for and against benefit.

This evidence can be accumulated across many analyses in clinical research that bear on the research question.

* In a 1980 US presidential debate, candidate Ronald Reagan reduced the complex cultural and economic questions facing voters by asking "Are you better off now than you were four years ago?"

Duality/quanta analysis respects the role of prospective evaluations, while also providing a way to incorporate safety analysis (at the beginning) and exploratory analyses with precise outcomes (at the end) of the analysis paths.

Finally, it provides the basis for additional mathematical research to sharpen its contribution to health care research.

The limitations of this approach have been provided; however, those limitations are simply methodology and will be removed through continued mathematical developed and with experience that comes from practical use of these tools with real data.

The need for a solid research foundation endures

The duality/quanta approach recommended in this text is an alternative to statistical hypothesis testing, but it is not the only alternative. Over the generations, Bayes procedures and artificial intelligence algorithms consistently show promise, but the community energy is not behind these approaches (or any *p*-value alternative).

However, any new approach requires a solid epidemiologic and logistical foundation. The investigation must be well designed. Logical contemporary control groups, precise endpoints, and sensible outcomes must be selected. Effect sizes must make clinical sense. Quantitative metric of successes should be preannounced.

With this solid runway in place, several different methodologies besides can be used to bring the research plane in for a successful landing.

Cultural conflict of interest

Reviewing quanta/duality analysis as well as other admissible substitutes as possible *p*-values replacements is an imperative. However, we biostatisticians and to some

degree clinical investigators must begin this review with an acknowledgment of our own conflicts of interests.

Our conflict of interest is not necessarily financial, but intellectual. It is a cultural conflict of interest.

Statistical hypothesis testing has been in wide use since the mid 1950's. The decades from then to now span the working careers of most all of us.* We understand how to conduct statistical hypothesis, and have some comfort level with their degrees of complexity.

For many biostatisticians, work consists of designing research endeavors that will produce *p*-values, discussing which *p*-value is most suitable to the circumstance, and generating tables with *p*-values. Our careers have become *p*-value centric.

A move away from statistical hypothesis testing holds important implications for the suitability of our knowledge base, our productivity, and our careers. If would be a profound midcourse correction.

Clinical investigators, while in a somewhat different boat, are caught up in the same current. These researchers are forced to develop some facility with *p*-values, since these computations are required for grant applications, the consideration of result-laden abstracts at influential meetings, and ultimately manuscript publication.

Thus, clinical investigators have become used to *p*-values even if they do not like them. However, despite their mistrust, and (commonly) miscomprehension of them, when a clinical investigators' results are accompanied by small *p*-values, the researchers can't help but rejoice, regardless of the effect size.

P-values may be rancid butter, but they are on the right side of the investigator's bread when less than 0.05,

* Mine began in my third year of medical school, 1977.

We therefore must acknowledge that taking a step away from statistical hypothesis testing while introducing opportunity also injects new uncertainty into research programs.

Some will resist any change, simply because it is change. This later group has a vested interest in seeing statistical hypothesis testing remain in place and will fight to keep them in their place of dominance. We as a community must manage our conflict.

Taking matters into our hands

Clinical investigators can no longer wait to be rescued from statistical hypothesis testing. As we have seen, this patience has so far been rewarded by the straightjacketing of complex clinical research interpretation and the attempt by some quantitative workers to "double down" on the *p*-value, driving it from 0.05 to 0.005. Since these are not a helpful solution as pointed out in Chapter 2, investigators must therefore take matters into their own hands.

They have the power they need to affect change. It is clinical investigators – not biostatisticians – who develop the ideas for new interventions. It is the clinical investigators who understand the disease and the population of patients in whom the intervention will be used. They also understand the necessary duration of follow-up, and the serious adverse events when they occur.

While these critical abilities do not permit investigators to choose the metric that will objectively assess their work, they have the power to resist the current statistical hypothesis testing metric, call for its replacement, play a role in the developing the criteria for its replacement, contribute to the choice of a new metric, and monitor its performance.

Clinical investigators are no longer innocent bystanders being hit by statistical hypothesis testing crossfire. We have to work our way to a place at the table.

Researchers should have at their disposal. a wide range of biostatistical support procedures including, but not requiring statistical hypothesis testing. Such an approach is wholly consistent with the desire of researchers 70 years ago when all recognized the sad state of affairs of research protocols (Chapter 2). It provides the rigor of a well-conceived protocol but is not *p*-value centric. And statistics should be contributory and supportive, but not dominant in research design. That is a position reserved for clinical investigators and epidemiologists.

The role of NIH and the FDA is to support these new innovative efforts, not enforce an obsolete administrative and statistical metric of success simply because it is the system to which they have become accustomed.

These federal administrators have a need for an administrative metric for assessing the results of a research program. This should be developed for them. But they should no more enforce a single metric of statistical data analyses on investigators than university researchers should dictate to NIH the funding pay line.

Longitude

Perhaps what is required in a new version of the emergency decree of Queen Anne.

In the late 17^{th} century, European maritime commerce was failing. Although the development of the New World was well underway, it could not be reliably reached.

While late arrivals were commonly explained by the ship's captain as "losing the weather gauge"* enroute, the principal reason for delay was simply that ships were all too frequently lost at sea. This occurred because while seamen could easily determine their latitude† they could not find their longitude. The absence of a real time longitude assessment had grave consequences for trade development as well as important naval ramifications.

In response to this, at the end of Queen Anne's reign, the English Parliament passed the Longitude Act of 1714. It established a Board of Longitude and offered monetary rewards to anyone who could establish a simple and practical method for the determination of a ship's longitude. This generated excitement in the maritime community, many different ideas were suggested, and ultimately a solution was found. ‡

While investigators anguish over the use of statistical hypothesis testing, and many of us statisticians are apparently "lost at sea" when asked to define a *p*-value,§ there is sadly little "wind in the sails" of the statistical community to find a *p*-value replacement. However, an Act

---

*Losing the weather gauge met no longer having a favorable wind

† From the date and duration of the day.

‡ Even though this Board was rife with conflicts of interest as members demonstrated favoritism for different candidates, in the end they settled on the best solution. The winning candidate produced an accurate, portable clock that would work reliably as sea. With two such clocks on board, one reading London time, the other the time at high noon on ship, one could convert the difference in the time to a specific change in longitude.

§ As stated in the preface, confusion among statisticians became so bad that the American Statistical Association, for the first time in its 177 year history, felt compelled to issue a statement clarifying for its own membership what *p*-values mean and how they should be used. This statement let to further clarifications.

of Congress in accordance with the NSF, superseding the FDA and NIH (deeply invested in *p*-values) could provide sufficient inducement.

Biographies
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.[1]

He probably died for it.

In the 19th century, mathematics had not yet escaped the grasp of religion. One such captive content area was infinity. While all mathematicians knew that the counting numbers were infinite, very little was understood about the concept of infinity. The only intuitive concept was that of eternal life, and since God created that, then that must be where God had a special place for himself. And if the natural numbers were infinite, then he had a special place there as well.

The oldest of six children, Georg Cantor was known at an early age for his abilities not as a mathematician but as a violinist.

Born in the western merchant colony of St. Petersburg, Russia, his family moved to Germany in part to escape the brutal Russian winters. Receiving a substantial inheritance following his father's death in 1863, Cantor shifted his studies to the University of Berlin where he completed his dissertation on number theory there in 1867.[1]

After a brief period where he taught at a Berlin girls' school, Cantor accepted a position at the University of Halle, where he spent his entire career. Within ten years he married Vally Guttmann and with her had six children.

During this time, Cantor entered into correspondence with Richard Dedekin and Gösta Mittag-Leffler. In responding to one of Cantor's submissions to his journal, Mittag-Leffler's stated that Cantor's writing was "about one hundred years too soon."

This was 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.

Cantor changed that in the space of ten years.

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 seemed full of contradictions. [1]

For example, it was well known 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 the interval [0,1]. This infinite set of numbers, when added to the whole numbers (themselves rational), meant to the mathematicians and philosophers at the time that there were more rationale numbers than whole numbers. Yet, but sets were infinite. Wasn't this a contradiction?

---

* Whole numbers are the counting numbers 0,1,2,3….

† A rationale number is any number that can be expressed as a ratio of whole numbers (including those multiplied by -1)

Cantor began here. He defined first finite and infinite sets, then divided the infinite sets into "denumerable" or countable versus non-denumerable or uncountable sets. He introduced fundamental constructions in set theory, such as the power set of a set A.*

He then provided that when the set *A* is infinite, the number of elements in the power set of A is strictly larger than the size of A. His work demonstrated that infinity was far more complex than anyone imagine. This result soon became known as Cantor's theorem.

Despite growing criticism, Cantor continued his breakthrough work. He developed the one-to-one concept which is a cornerstone of set theory. He showed that sets could be quite complicated (e.g., his famous Cantor set), and thereby demonstrated the utility of different types of "infinity". There was one infinity for the rationale numbers, and another, larger concept of infinity for the irrational numbers. [1] He also defined irrational numbers to be the limit of a sequence of rational numbers.[2] These distinctions caused havoc with the 19^{th} century understanding of the real number line.

This worked rocked the religious community. It was a bombshell. 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 pushed further, saying that Cantor denied the existence of one God, and that the multiple infinity concept − since it must imply multiple Gods − meant Cantor was a pantheist. [1]

* The power set of a set *A* is the set of all subsets of the set *A*.

Cantor, weary from his continued work in a complex and controversial field, and unprepared for the *ad hominem* attacks began to suffer emotionally.

Cantor suffered his first known bout of depression in 1884 after a damaging series of attacks on his work by Kronecker, who criticized Cantor as a charlatan, renegade, and a corrupter of youth. [1] He doubted whether he would ever be able to return to mathematics. He was place in a sanatorium in 1899, and soon after that, his youngest son died, and event which sapped much of his intellectual strength.

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 was profoundly affected, and began a bout of chronic depression that lasted for the rest of his life. [1]

He retired in 1913 and lived the rest of his life in poverty until he died in a sanatorium in 1919.

---

1. https://en.wikipedia.org/wiki/Georg_Cantor last accessed 1-14-2020.
2. https://www.britannica.com/biography/Georg-Ferdinand-Ludwig-Philipp-Cantor last accessed1-14-2020.

Bernhard Riemann

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 intererst in mathematics and history. Encouraged by his family, he entered preparatory school in Hannover, later moving to Lüneburg [1].

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. [1]

However he attended some mathematics lectures and in the process, always close to his family, asked his father if he could transfer to the faculty of philosophy so that he could study mathematics. [2]

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. [1]

Riemann studied the work of Cauchy, who had created of the $\varepsilon - \delta$ method of calculus, and his work on integration through the development of the Riemann integral is still taught today.

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 – one of Riemann's teachers, Frederick Gauss). Riemann developed the powerful tool of studying limits using the $\varepsilon - \delta$ method of examining a functions 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.*

Riemann married in July 1862, and later that year developed tuberculosis, a disease that at the time had no known cure. In order to recuperate, he travelled to Italy

---

* This involves the Riemann zeta function, which is a function $\zeta(s)$ of a complex variable s defined as follows. If the real part of s is greater than 1, define $\zeta(s)$ to be the sum of the convergent series $\sum_{n \&ge 1} n^{-s}$; then extend $\zeta(s)$ to the whole complex plane by analytic continuation. The Riemann hypothesis states: *if $\zeta(s) = 0$ 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. [1]

several times, befriending among the most important mathematicians, Betti and Beltranni [1]. He died in the Italian village of Selasca where he spent his last weeks with his wife and three-year-old daughter.

---

1 https://www.usna.edu/Users/math/meh/riemann.html last accessed 1-14-2020.

2. http://mathshistory.st-andrews.ac.uk/Biographies/Riemann.html last accessed 1-14-2020.

Henri Lebesgue

At the end of the 19th century, the evaluation of functions was considered to be essentially complete. Continuous functions were well understood, while discontinuous functions, remaining somewhat outside the mainstream as curiosities were given relatively little attention. However, discontinuous functions were of increasing attention given the demonstration that the integral of Bernhard Riemann did not apply to them in general.[11] It was Henri Lebesgue who, while a graduate student, formulated the Lebesgue integral, which covered both continuous and discontinuous functions, greatly expanding the power of integration theory.

Henri Lebesgue (pronounced La-BÁK)was born on July 28th, 1875 in Beauvais, France. His father, a typesetter, died of tuberculosis when Lebesgue was very young, forcing his mother, a teacher to support him by herself.

However, after observing Lebesgue's early talent for mathematics, one of his instructors arranged for community support to continue his education. This was a remarkable initiative of charity and a benevolent community response. It would pay handsome dividends.[2]

Lebesgue entered the École Normale Supérieure in Paris in 1894 and was awarded his teaching diploma in mathematics in 1897. It was at this time this he learned of Émile Borel's work on the rudiments of measure theory, and Camille Jordan's work on the 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.[2]

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.

Between March 1899 and April 1901 Lebesgue published six notes in Comptes Rendus. The first of these, unrelated to his development of Lebesgue integration, dealt with the extension of Baire's theorem to functions of two variables.

Building on the work of others, including that of Émile Borel and Camille Jordan, Lebesgue formulated the theory of measure in 1901, in which he gave the definition of the Lebesgue integral. This generalized the notion of the Riemann by extending the concept of the area below a curve to include many discontinuous functions.

This generalization of the Riemann integral revolutionized the integral calculus. In 1902 he earned his Ph.D. from the Sorbonne with the seminal thesis on "Integral, Length, Area", submitted with Borel, four years older, as advisor. His contribution is one of the major achievements of modern analysis. [2] His concept brought the notion of measure, (then incompletely formulated) to integration, opening the door to the use of the integral as an application of measure theory. [1]

Having graduated with his doctorate, Lebesgue obtained his first university appointment when in 1902 he became mâitre de conférences in mathematics at the Faculty of Science in Rennes. In 1903 he married Louise-Marguerite Vallet and they had two children. However they divorced thirteen years later.

It is interesting that Lebesgue did not concentrate throughout his career on the field which he had himself started. This was because his work was a striking generalization, yet Lebesgue himself was fearful of generalizations. Instead, he chose to make 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. He spent the rest of his life working on elementary geometry, teaching materials, and historical work. [3]

---

1 https://www.britannica.com/biography/Henri-Leon-Lebesgue last accessed 1-14-2020

2. https://en.wikipedia.org/wiki/Henri_Lebesgue last accessed 1-14-2020.

3. http://mathshistory.st-andrews.ac.uk/Biographies/Lebesgue. html last accessed 1/14/2020.

Thomas Joannes Stieltjes

Thomas Joannes Stieltjes was born in Zwolle, Holland in 1856. 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, Thomas, spending his time reading the mathematics of Gauss and Jacobi, rather than focusing on the requisite civil engineering tracts, repeatedly failed his exams.[1]

However, he was able to secure, which his father's help, a job at the Leiden Observatory where he began a lifelong correspondence with Charles Hermite in celestial mechanics and mathematics, devoting his spare time to mathematical research. He made many contributions to number theory and harmoic analysis [2]

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. Supported by his wife, he moved completely into 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 provided a generalization of

Riemann's work, and is widely used for applications in physics.

Commonly theoreticians affix the name of this mathematician to integration. Riemann interation is sometimes referred to as Riemann-Stieltjes integration, and as in this treatise, Lebesgue integration is referred to as Lebesque-Stieltjes integrals.

After many years, and the intervention of Hermite, Stieltjes received an honorary doctorate from Leiden University, enabling him to become a professor.

References

1. http://www.britannica.com/biography/Thomas-Jan-Stieltjes last accessed 1/14/2020.
2. https://en.wikipedia.org/wiki/Riemann%E2%80%93Stieltjes_integral last accessed 1/14/2020.

Andrey Kolmogorov

Modern probability theory begins with Kolmogorov. He laid the mathematical foundations of probability theory and the algorithmic theory of randomness, making crucial contributions to the foundations of statistical mechanics, stochastic processes, information theory, fluid mechanics, epidemiologic modeling, and nonlinear dynamics.

Andrey Kolmogorov was born in 1903 in Tambov, Russia. There is not much 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] His mother, named Kolmogorova, died in childbirth.

Kolmogorov was raised by two aunts at his grandfather's estate. He attended the village school and demonstrated genuine curiosity about mathematics, having his mathematical works (as well as his early literary writings) printed in the school newspaper.

As a teenager, he developed "perpetual motion machines", hiding their defects so adroitly that his teachers could not find the flaws. In 1910, his aunt adopted him and then they moved to Moscow, where he went to high school, graduating from in 1920.

For a time, Kolmogorov had an eclectic existence. After he left school, he first worked for a while as a conductor on

the railway. [2] During this time, he wrote a treatise on Newton's law of mechanics.

He later entered Moscow State University, but, uncommitted to mathematics, studied a number of fields, including metallurgy and Russian history, about which he had a strong passion.

Well before he graduated, he lit his star in the international arena by writing a paper on set operations in 1922, a major generalization of Suslin's. By June 1922 he had constructed a summable function which 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 published eight papers, all while an undergraduate. [1]

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 the stochastic calculus. By this time he had 18 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 (Ph.D.) degree, from the Moscow State University, and became a professor at Moscow State University in 1931, devoting himself to a rigorous examination of the underlying tenets of probability, reformulating probability in a 1933 paper in which he assembled its development from a fundamental collection of axioms, much like Euclid developed geometry.

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 1933, Kolmogorov published his book, the *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.[3] It was in this work that he developed the concept of 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 measure).

In 1935, Kolmogorov became the first chairman of the department of probability theory at the Moscow State University.

In 1939, he was elected a full member (academician) of the USSR Academy of Sciences. 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.

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 developed 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.

References

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. http://arbor.revistas.csic.es/index.php/arbor/article/viewFile/551/552. Last accessed 1/14/2020

Index