Statistical Reasoning in Medicine
The Intuitive *P*-Value Primer

Second Edition

Lemuel A. Moyé

Statistical Reasoning in Medicine

The Intuitive *P*-Value Primer

Second Edition

Springer

Lemuel A. Moyé
University of Texas
Health Science Center at Houston
School of Public Health
Houston, TX 77030
USA
moyelaptop@msn.com

Library of Congress Control Number: 2006930252

ISBN-10: 0-387-32913-7
ISBN-13: 978-0387-32913-0

Printed on acid-free paper.

Printed in the United States of America. (EB)

9 8 7 6 5 4 3 2 1

springer.com

To Dixie and the DELTs

Preface

You and I have some important and interesting conversations coming up shortly. However, I propose that we postpone those for a moment while I share with you my motivations for writing this book.

The frustrations of doctors, nurses, judges, legislators, and administrators that arise as they interpret healthcare research efforts are the unfortunate and predictable products of their meager research backgrounds. It is only human for them to grab for whatever supporting grips are available; one such handhold is the ubiquitous *p*-value.

This reduction of a research effort to a single number is regrettable, but quite understandable. The complexity of a modern healthcare research endeavor requires a clear understanding of the circumstances in which one can generalize results from relatively small samples to large populations. Even though the concept of generalization is nonmathematical, many researchers are not its master. Recognizing their disadvantage, they latch onto the *p*-value, believing that it neatly binds these complicated features into one tidy package.

However, like continually substituting desserts for nutritious meals, the habitual replacement of *p*-values for clarity of vision is unfulfilling and dangerous. This book reaches out to these principle-starved people. Specifically I want to use the ubiquity of the *p*-value as an overture to the discussion of statistical reasoning in medicine.

Statistical reasoning in medicine is the process by which one determines whether sample-based results can be extended or generalized to the population at large. The concepts are straightforward, intuitive, and quite precise. However, their application requires thoughtful consideration.

For many years the tendency in the research community has been to replace this deliberation with a quick and simple assessment of the *p*-value's magnitude. The research community, in its quest for significant results, has created a polluted sea of *p*-values in which we all restlessly swim. Although *p*-values were designed to make a simple statement about sampling error, for many they have become the final arbiter of research efforts.

Investigators often gnash their teeth over this entity's value at their study's conclusion: is it less than 0.05 or ≥ 0.05? To these workers, *p*-values are the switching signal for the research train. If the *p*-value is less than 0.05, the research-train moves down the main track of manuscript publication, grant awards, regulatory approval, and academic promotion. However, if the *p*-value is greater than 0.05, the switch moves the other way, directing the research train off to the elephant's graveyard of discarded and useless studies. Replacing the careful consideration of a research effort's (1) methodology, (2) sample size, (3) magnitude of the effect of interest, and (4) variability of that effect size with a simple, hasty look at the *p*-value is a scientific thought-crime.

P-values continue to be the focus of research discussions in academic centers, remaining a staple of the medical community's research effort. The approval of a new medical intervention commonly includes consideration of the *p*-value, and arguments in courts of law for the scientific basis of an assertion frequently concentrate on the size of the *p*-value. Clearly, many researchers, journal editors, regulators, and judges cling doggedly to its use. It is therefore all the more curious that so few of these specialists understand either what the *p*-value is or precisely what information it is designed to convey. Although they understand the message that the *p*-value "had better be less than oh five," there is little understanding of either the source or justification of this ubiquitous mantra.

I don't think we statisticians have been as helpful as possible. A biostatistics professor at a school of public health once asked a statistics student sitting for his qualifying exam (that must be passed to enter the Ph.D. candidacy phase), "Explain what a *p*-value means." The professor never received a satisfactory response.[*] When biostatisticians do respond to this question, we often give the following response, "the *p*-value is the conditional probability of rejecting the null hypothesis in favor of the alternative hypothesis when the null hypothesis is true." I fear that to the non-statistical world, this answer smacks of Orwellian double-speak.

This text emphasizes an intuitive understanding of the role of the *p*-value in sample-based research, deemphasizing the underlying mathematics. This nonmathematical approach is available when the foundation principles of statistical reasoning in medicine are clearly articulated. Our purpose here is to clearly state and develop the principles that govern when and how one takes results from a small sample and applies them to a larger population in healthcare research. The enunciation of these principles brings the roles and limitations of *p*-values into sharp focus.

Lemuel A. Moyé
University of Texas
School of Public Health
April, 2006

[*] Related by Dr. Sharon Cooper, Chair of Epidemiology and Biostatistics, Texas A&M Rural School of Public Health.

Acknowledgments

Approximately 8,000 North Carolina Cherokee are descended from those who did not take the *Nunadautsun't* to the Oklahoma reservations; one who remained was my grandmother. An important lesson she taught her grandchildren was never to begin the day until you have given thanks to God for the people He has placed in your life to guide you.

I owe a special debt of thanks to the leaders and members of the Cardio-Renal Advisory Committee of the Food and Drug Administration, with whom I served as a statistician from 1995 to 1999. They are Ray Lipicki, JoAnn Lindenfeld, Alastair Wood, Alexander Shepherd, Barry Massie, Cynthia L. Raehl, Michael Weber, Cindy Grimes, Udho Thadani, Robert Califf, John DiMarco, Marvin Konstam, Milton Packer, and Dan Roden. The public sessions held three times a year are among the best training grounds for cardiovascular research epidemiology and biostatistics. I am also indebted to Lloyd Fisher, with whom I have often disagreed and from whom I have always learned.

I am indebted to my alma mater, The University of Texas School of Public Health, and to public school education in general, which has provided much of my education. To all of you, and especially Asha Kapadia, Palmer Beasley, Robert Hardy, Charles Ford, Barry Davis, Darwin Labarthe, Richard Shekelle, Fred Annegers, and C. Morton Hawkins, consider this book an early payment on a debt too large to ever repay. I would also like to thank Craig Pratt and John Mahmarian, who have been patiently supportive in the development of my ideas. Furthermore, I would like to express my gratitude to Marc Pfeffer, Frank Sacks, Eugene Braunwald – you have been more influential than you know.

To the fellow classmates and friends known as the "DELTS"— Derrick Taylor, MD, Ernest Vanterpool, and Tyrone Perkins. Ever since high school in Queens, New York, we have pushed and prodded each other to excellence. You are a part of all my worthwhile endeavors.

Additionally, I have been blessed with friends and colleagues who were always willing to share from their wellspring of good sense with me. Four of them are James Powell of Cincinnati, Ohio, John McKnight of Washington D.C., Joseph Mayfield of New York City; and Gilbert Ramirez, now in Iowa.

John Kimmel, the reviewing editor, and many chapter reviewers have provided good questions and additional references, improving the book's structure. In addition, Dr. Sarah Baraniuk, Dr. Yolanda Muñoz, and Dr. Claudia Pedroza each suggested important improvements in the last chapters of this book. Dr. Steve Kelder provided some especially helpful insights for the chapter on epidemiology.Lisa Inchani, a graduate student of mine, selflessly volunteered her time to patiently review and copyedit this text. Its readability is largely due to her steadfast efforts.

Finally, my dearest thanks go to Dixie, my wife, on whose personality, character, love, and common sense I have come to rely, and to my daughters Flora and Bella, whose continued emotional and spiritual growth reveals anew to me each day that, through God, all things are possible.

Lemuel A. Moyé
University of Texas
School of Public Health
April, 2006

Contents

Introduction

I sat quietly, awaiting the one question that would utterly destroy my career. It exploded in a gasp of exasperation from the great man…

Like every new fourth-year medical student at Indiana University School of Medicine in 1977, I thought hard and carefully about life as an M.D. With an undergraduate degree in applied mathematics, I had decided to begin graduate studies in statistics after I graduated from medical school. However, several practicing doctors convinced me to postpone my formal education in statistics for a year to first complete a one-year medical internship upon graduation from medical school.

My personal life required that I stay in Indianapolis for any post-medical school work. Thus, while my classmates applied to many hospitals across the country and around the world, I applied to only two, and was interested in only one of those — Methodist Hospital Graduate Medical Center. Studying their response, I was stunned to learn that one of my questioners would be Dr. William Kassell, Chairman of Obstetrics and Gynecology.

This was terrible news! My performance in obstetrics and gynecology as a third year medical student the year before was not auspicious. As my first clinical rotation after psychiatry, Ob-Gyn was a rude awakening to the ceaseless and pressing responsibilities of surgeons. I awkwardly tried to juggle 6:30 AM rounds, patient responsibilities, demanding surgeons, steep learning curves, and long overnight hospital hours. Now, I would have to meet with the tough-minded chairman of that department, a man known for his quick appraisals and blunt critiques.

Self-inflicted brutality characterized the night before my interview with Dr. Kassell. Reluctantly pulling out my old Ob-Gyn notes, I again reviewed colposcopy findings and cervical cancer treatment procedures. Re-memorizing the workup of pre-eclampsia, I rubbed my brain raw with the sequence of gynecologic examination procedures.

By the next morning, I was jammed full of gynecologic and obstetrics information. However, while I bullied myself into believing that my head was ready, my heart dreaded the coming confrontation scheduled to begin shortly. Driving to the hospital, I paid no attention to the fall foliage, distracted by a new, persistent smell of defeat that now spoiled the crisp, clean air.

Arriving five minutes early, I stood alone in his huge office, the bright light from the open curtains illuminating my anxiety. Heart racing, shirt sticky with nervous perspirations, I waited his arrival with growing dread.

Suddenly, Dr. Kassell burst into the room, gruffly throwing a greeting in my general direction and waving me to a seat facing him. He was tall and clean-shaven, with large expressive eyes and a thick mane of silver hair. He carefully scrutinized me over his expansive desk, heaped high with textbooks, papers, and hospital charts. I sat, quietly awaiting the one question for which I had not prepared the night before — the one question that would send my career plans crashing into ruin. It exploded in a gasp of exasperation from the great man:

"Dr. Moyé, will you please tell me where a *p*-value comes from?"

"What!" Well,…It….I mean…." I stammered, trying to catch my intellectual breath.

"I asked you," he repeated impatiently, raising his voice for emphasis, "to tell me what a *p*-value is. Can't you do that? What's all of this 0.05 business about? What's so special about that number?" the doctor continued, his frustration conveyed by the boom in his voice.

Just prior to my interview, the chairman had reviewed a manuscript for his journal club (the author and topic I have long since forgotten) in which *p*-values served as the yardstick against which the results were measured. It seemed to Dr. Kassell that these *p*-values were like some unfeeling arbiter, dispassionately determining if study results were positive or negative. Having reviewed my record with its annotation of my undergraduate statistics background, he had decided to spend our interview time discussing this research issue, and not the details of Ob-Gyn.

This book is for everyone in healthcare who requires a nonmathematical answer to Dr. Kassell's question. *Statistical Reasoning in Medicine: The Intuitive P-value Primer* focuses on both the underlying principles of statistical thought in medicine and the ethical interpretation of *p*-values in healthcare research.

A physician confronted with a new finding in her field, a director of a pharmaceutical company analyzing a series of experiments, an expert sitting on an advisory panel for the government, or a judge assessing the scientific aspect of a lawsuit must have an understanding of *p*-values and the underlying statistical thinking that guides their interpretation. If you are a decision maker without in-depth training in statistics, but now find that you must grapple with the thorny theory of statistical reasoning and the nettlesome issues of *p*-values, then this book is for you.

With an emphasis on patient and community protection, *The P-value Primer* develops and emphasizes the *p*-value concept while deemphasizing the mathematics. It also provides examples of the *p*-value's correct implementation and interpretation in a manner consistent with the preeminent principle of clinical research programs: "First, do no harm".

The *P-value Primer's* prologue, describes the controversies that have engulfed statistical reasoning for 400 years, providing a brief history of the development of the concept of data analysis. In Chapter One, the concept of sampling error, and the reasons that physicians have such difficulty understanding

the population perspective that is so prevalent in research is discussed. The requirement of sample-based research is developed from basic principles, building up the reader's nonmathematical intuition of the notion of sampling error. The natural, intuitive, and nonmathematical notion of study concordance (in which an experiment analysis plan is immune to incoming data) versus study discordance (in which an analysis plan itself is severely perturbed by its own data) is introduced in Chapter Two.

Chapter Three reviews the statistical hypothesis-testing paradigm and introduces the concept of *p*-values and power from the sampling error perspective. Stressing the concept rather than the mathematics permits the development of a useful definition for the *p*-values in laymen's terms. Chapter Four discusses the principles of epidemiologic research and the role of *p*-values. Progressing from there, the longstanding debate over the propriety of *p*-values is discussed in Chapter Five. This chapter reveals that the concern about *p*-value use is not simply whether they are interpreted correctly, but about the proper role of mathematics in healthcare research.

The *P-value Primer* moves from there to discuss relevant issues in the applications of hypothesis testing for the investigator. Chapter Six provides a modern discussion of the issues of power and sample size. Discussions of how the courts view scientific evidence in general and statistical inference in particular is offered in Chapter Seven. Chapter Eight focuses on one-tailed versus two-tailed hypothesis testing. I then describe the basics of alpha allocation in the research effort that has multiple clinical measures of interest and combined endpoints (Chapter Nine). Subgroup analyses are data dredging are developed from first principles in Chapter Ten. Chapter Eleven discusses the interpretation and utility of regression analysis. Finally, an introduction to Bayes analyses is presented (Chapter Twelve). The book's conclusion provides concrete advice to the reader for experimental design and *p*-value construction, while offering specifics on when the *p*-values of others should be ignored.

The unique feature of *The P-value Primer* is its nonmathematical concentration on the underlying statistical reasoning process in clinical research. I have come to believe that this focus is sorely lacking yet critically needed in standard statistical textbooks that emphasize the details of test statistic construction and computational tools. I quite consciously deemphasize computational devices (e.g., paired *t*-testing, the analysis of variance, and Cox regression analysis), focusing instead on the nonmathematical features of experimental design that either clarify or blur *p*-value interpretation.

There is inevitable tension between the mathematics of significance testing and the ethical requirements in medical research; this text concentrates on the resolution of these issues in *p*-value interpretation. Furthermore, the omnipresent concern for ethics is a consistent tone of this book. In this age of complicated clinical experiments, in which new medications can inflict debilitating side effects on patients and their families, and where experiments have multiple clinical measures of success, this text provides concrete, clear advice for assembling a useful type I error structure, using easily understood computations, e.g., the asymmetric apportionment of alpha and the intelligent allocation of alpha among a number of primary and secondary endpoints in clinical experiments.

The *P-value Primer* is written at a level requiring only one introductory course in applied statistics as a prerequisite; the level of discussion is well

within reach of any healthcare worker who has had only a brief, introductory statistics background. It will be valuable to physicians, research nurses, healthcare researchers, program directors in the pharmaceutical industry, and government workers in a regulatory environment who must critique research results. It is also useful as an additional text for graduate students in public health programs, medical and dental students, and students in the biological sciences.

So, how did I answer Dr. Kassell back in 1977? Fortunately, I answered him accurately, but unfortunately, I gave the knee-jerk response many statisticians give, "A *p*-value is the conditional probability of rejecting the null hypothesis when the null hypothesis is true," I most certainly confused him. Even though my response hit the technical nail on the head, I failed in providing the clear, direct answer that would have more usefully answered his query. Over the years, I could never shake the feeling that, after listening to this terse, reflexive reply, a nonstatistical listener remains befuddled about what these *p*-values really are. Like the newcomer to a foreign language who gets a verbose reply to his short and hesitant question, the inquisitor is frustrated and overwhelmed. This book's goal is to dispel much of that confusion.

Prologue

Aliis exterendum

It is difficult to appreciate the bitter contentions probability and statistics engendered when introduced to Western society. Considered unnecessary in a world where all events were predetermined by higher powers, the study of the relative frequency of events was discouraged for centuries. The nascent field of statistics (not known by that name when first introduced) was all but torn apart by the political and religious controversies its initial use sparked. While some resistance to these areas can be found in jealousies that plague the human heart, an important source of this active resistance was the inability of an unprepared society to first grasp, and then be transformed by the illumination these fields provided.

Thus, an understanding and appreciation of the role of statistics—past, present, or future—can be found in an examination of the culture in which it operates. The Persian practitioner Avicenna in the eleventh century provided seven rules for medical experimentation involving human subjects [1]. Among these precepts was a recommendation for the use of control groups, advice on replication of results, and a monitory against the effects of confounding.* These observations represented a great intellectual step forward; however, this step was taken in relative isolation. While probability, statistics, and the principles of reasoning from data first require a set of data to evaluate, data was available for centuries before these fields developed. An additional 500 years passed before the line of reasoning that led to the concepts of modern probability in applied healthcare emerged; and it was another 300 years before statistical hypothesis testing and *p*-values were produced. In order to understand the initial twists and turns of the development of this curious discipline, we need to take a quick diversion to life in Europe 500 years ago.

Europe's Emergence from the Middle Ages

Ensuring society's survival before developing society's statistics was the necessary order of progress. The continent struggled, unevenly emerging from the provincialism and ignorance of the Middle Ages in the sisteenth century. Although the majority of Europeans subsisted in the abject rural poverty, groups of Europeans were coming together in numbers. Naples, Lisbon, Moscow, St. Petersburg, Vienna, Amsterdam, Berlin, Rome, and Madrid each contained more than 100,000 people in the

* Confounding is the observation that the effect of one variable may confuse the effect of a second variable. For example if only women are exposed to an active therapy, and only men are exposed to the control therapy, then the effect of the therapy is confounded, or bound up, with the effect of gender.

1500s, with London and Paris being the largest of these new urban centers [2]. This movement to urbanization accelerated, albeit slowly, creating new links of interdependence among the new city-dwellers. However, with little knowledge about themselves, the residents of these new cities remained blind to their own corporate needs, and could therefore not direct their social progress.

Although rural inhabitants vastly outnumbered urban-dwellers, the contrasts between the large city with its incipient education system and exciting culture on the one hand, and the surrounding, poverty-stricken countryside, on the other, were striking. The one-sided economic relationships between the two environments reflected the undesirability of rural life. Although towns required the resources of the countryside, these agrarian products were not purchased, but instead, were extracted through tithes, rents, and dues. For example, the residents of Palermo, Sicily, consumed 33% of the island's food production while paying only one-tenth of the taxes [2]. While peasants often resented the prosperity of towns and the ensuing exploitation, the absence of rural political power blocked attempts to narrow the widening disparity of wealth.

However, the attraction of cities was only relative; they had their own share of maladies. The unstoppable influx of unemployed rural immigrants looking for work generated a great job demand. Since cities proved no professional paradise for these unskilled workers, poverty emerged as a serious problem in the eighteenth century [2]. Additionally, this rapid arrival of destitute immigrants produced overcrowding that sparked a new round of disease. Despite an end to the most devastating ravages of plague, cities continued to experience high death rates (especially among children) because of unsanitary living conditions, polluted water, and a lack of sewage facilities. One observer compared the stench of Hamburg (which could be smelled from miles away) to an open sewer.

Why was this intolerable situation tolerated? One explanation was the lack of opportunity for most people to reflect on the quality of urban life. Consumed with work, sleep, church, or illness, citizens had little time for considered thought on how life could be improved. Additionally, there was no quantitative measurement of the problems of poverty and illness on a societal level. While each citizen had his or her own poignant anecdotal experience, these personal stories and examples provided conflicting views of the state of urban affairs. No group or person was able to assemble a corporate sense of the quality of life. Thus, there was no way to determine exactly how much poverty existed in a city, and therefore no procedure to track its level over time. The only widely accepted standard for urban life was rural life, and this, everyone agreed was worse than the current city conditions. Finally, the prevailing view was that living conditions improved exceedingly slowly, with progress measured over centuries rather than within a lifetime. This progress rate was far too slow to either track or influence.

Poverty was ubiquitous in the new urban centers, with as much as 10% of the population dependent on charity or begging for food in England and France. Earlier in Europe the poor had been viewed as blessed children of God, and the duty of Christians was to assist them. However, this point of view was replaced with a newer, darker suggestion that the poor were slovenly and unwilling to work themselves out of their lot in life. These opposing points of view produced a contentious search for the cure to poverty. From this emerging cauldron of social con-

flict came a fervor for change and the need to understand environmental and social influences on human culture. Since the cities were made up of individuals, were there not some features of the whole urban unit that could be influenced?

The first, natural place for people to turn for answers was not data, but the ruling class, seated at the pinnacle of European power.

Absolutism

The religious and social tragedies of the sixteenth and seventeenth centuries sparked the rise of the absolute king. The reformation in the early sixteenth century had been relatively and remarkably free of bloodshed. However, the growing division between the Christian churches in Europe, driven primarily by Protestant dissatisfaction with Catholic kings, unleashed a series of armed conflicts that would rage across Europe for more than 100 years. These vicious international and civil wars produced the deaths of tens of thousands of civilians in the name of religion. The institution of absolute monarchies was originally proposed as a solution to these violent religious disorders, and many in Europe were pleased to exchange local autonomy for peace and safety [3].*

With the exception of England, which experienced the replacement of its omnipotent monarchy by first a republic and then a weakened king, the rest of Europe supported the institution of supreme monarch. The icon of these monarchs was the "Sun King", Louis XIV of France, who under the claim of Divine Right, centralized the government, the civil bureaucracy, the legislation, and the judiciary [3].

Following his example, continental Europe moved in mass to the concept of an absolute monarch. Brandenburg, Prussia† would become one of the most powerfully centralized states in Europe under Frederick the Great. The Hapsburg emperors worked (ultimately, in vain) to consolidate the Czech-speaking territories into what would become the Austrian–Hungary empire. Tsar‡ Peter the Great of ruled Russia until 1725, brutally centralizing and westernizing its unique culture. Each of these empires converted from loosely governed autonomies to centralized states.§

---

* Europe with its history and memory of Roman rule intact, understood the problems that came with acquiescing to the power of an absolute monarchs. However, people were desperate for respite from the current civil slaughter underway. In March, 1562, an army led by the Duke of Guise attacked a Vassy Protestant church service Champagne province of France, slaughtering men women and children—all of whom were unarmed. Thus began the French Wars of Religion which were to last for almost 40 years and destroy thousands of noncombatants [3]. Ten years later, on August 24, 1572, the day before St. Bartholomew's Day, royal forces hunted down and executed over 3,000 Huguenots, in Paris itself. Within three days, soldiers under the direct command of disciplined officers systematically executed over 20,000 Huguenots in the single most bloody and systematic extermination of European civilians until World War II [3]. The war would last another 20 years.

† This would become modern-day Germany.

‡ The words "Tsar" or "Czar" are taken from the Latin "Caesar" [3].

§ An interesting historical irony is that the most absolutist states in history (Third Reich, Soviet Union) would not be created until the twentieth-century.

The monarchs of these new kingdoms wielded absolute authority in their nation-states, diminishing local rule of law. It was only natural that the people would turn to these super-rulers for wisdom on how to improve their lot. However, by and large the response was subjugation to the will of the king, and the requirement to pay national taxes to support new standing state armies.

A natural tax system was required to support the growing centralized superstructure. However, since a tax system would be unenforceable without at least the façade of equity, a census was required to raise revenue to support the new large standing armies. Thus, the first modern issues in statistics were issues not of statistical hypothesis testing, but of simple counting – the basis of demography.

Refusing to be Counted

The idea of counting individuals had its roots in antiquity and was described in the New Testament [4]. However, the notion of counting citizens fell into disfavor during the Dark Ages. The earliest modern attempt has been traced to a fourteenth century parish in Florence, Italy where births and deaths were recorded by beans (black for boys and white for girls) to determine the sex ratio [5].*

Early demographers in seventeenth-century England faced a daunting challenge, since anything approximating modern census machinery was nonexistent. Fearing that their involvement would lead directly or indirectly to higher taxes, many in the population actively refused to be counted. Additionally, the national government, recognizing the potential military value of a count of men available for service in the new national army, labored to keep whatever demographic data it had a secret. Thus, the first demographers lacked both a counting methodology and reliable data on which to base previous population size claims. Facing an unwilling population that actively resisted enumeration, the demographers' ingenuity was tested as they labored to create indirect estimates of the population's size, age and sex distribution. Multiplying the number of chimneys by an assumed average family size, or inferring age distribution from registered information concerning time of death were typical enumeration procedures [6].

On these basic estimates, taxes were collected, guaranteeing the existence of armies. These, in turn, guaranteed wars, as another lethal era of armed conflict began in the eighteenth century [2]. The Seven Year's War, involving the five great European powers, which spread from the Far East, across Europe and into the New World can legitimately be viewed as the first world war. As the new scale of conflict and terror rapidly drained resources, the demand for new taxes increased. These strident calls for new levies in the absence of the monarchs' interest in learning of the needs of its citizens, fueled cries for change. New, virulent diatribes against the privileged orders caught these ruling monarchs unprepared. This social ferment, created by sustained population growth, abject poverty, and more rapid communication through trade, created disruptive tensions that undermined the foundations of the old order [2].

* No one knows the name of the priest who attempted to use church records for counting.

Thus, the restricted use of early data tabulations indirectly added to the burdens of the increasingly impoverished and desperate populations. New calls for tax relief generated more extreme cries for restructured social order. The monarchs' inability to deal meaningfully with these stipulations led to a revolutionary outburst at the end of the eighteenth century, heralding the beginning of the end of absolute royal law. Attention to technology and profit replaced fealty to kings. And with the appearance of leisure time, people began to play games of chance.

No Need for Probability

Games of chance have been recorded throughout history. They had a prominent place in ancient Greek literature and society [7].[*] The idea of casting lots is mentioned in the Old Testament of the Bible. This concept spread across the Western world, and was converted into games that were played throughout the Middle Ages. The ubiquity and popularity of these diversions would have been a natural proving ground for the laws of probability. However, despite the long-standing existence of these games, people dismissed the idea of developing rules to govern their outcomes. There was a manifest absence of concerted effort to use mathematics to predict the results of games of chance for hundreds of years.

The explanation lies not only in the immaturity of mathematics, but also in the culture in which the mathematics would be viewed. For thousands of years, up through the fourteenth-century, not only was there no clear idea of a random event, but the need for such a concept did not exist. The prevailing perspective viewed all outcomes as determined by either man or deities (benevolent or malevolent). Thus, for many, a "game of chance" was played not to watch random events, but to observe supernatural forces at work. When biblical characters cast lots, they commonly did so not to gamble, but to engage in a process that removed man from the outcome, directly invoking the action of the supernatural. Many thought-leaders of the time believed that all events were pre-determined, further banishing the idea of random events.

Thus, for many, winning a game of chance in an otherwise brutal and unforgiving world equated with receiving, if only for a moment, the undeserved favor of God. The idea of predicting the outcome of the game, thereby diminishing the perceived role of the supernatural, was both anathema and anti-cultural. Such prediction activities flirted with lewd conduct at best[†] and witchcraft at worst.

---

[*] Aristotle said, in a justification of gambling, "Amusement is for the sake of relaxation and relaxation, must necessarily be pleasant, since it is a kind of cure for the ills that we suffer in working hard." Aristole. *Politics* VIII5, 1339b;15–17, trans. T.A. Sinclair.

[†] The conversion of harmless games of chance to gambling by the injection of money tarnished the spirituality of the pasttime, and the practice was seen as less benign. As the recognition grew that gambling attracted the seamier side of cultural elements, attempts were finally made to limit its practice. For example, nobles who chose to fight in the Crusades were permitted to gamble, but the games they could play and the number of attempts they could make were strictly regulated.

Such prevailing opinions pushed the idea of the random events and their predictions beyond the reach of mathematics.* Many generations would pass before culture could openly embrace the reasoning of workers who argued that some results appeared to be governed by chance. This acceptance, which first appeared in the 17th century, permitted society to be illuminated by the development of new natural laws.†

Intellectual Triumph: The Industrial Revolution

The instigating activity of the industrial revolution substituted inanimate energy forms for organic (human and animal muscle) ones. This replacement transformed society as never before. Unlike the prevalent ethereal forces directing peoples' loyalties to the old order of monarchies, this energy conversion required direct, information-based, cerebral activity. The Industrial Revolution's knowledge-based approach to productivity required quantitative data in ever-increasing amounts. A triumph of the intellect, the Industrial Revolution represented not just a one-time jump in productivity and wealth but a process of ever-accelerating change.

England of all European countries was the best poised for this thrust forward. Its low interest rates, stable government, available lending sources, (relatively) low taxes and the a weakened guild structure sparked enterprise. Once created, this environment catalyzed a cascade of innovation as one invention sparked another.

The development of the flying shuttle, the spinning jenny, the water frame, the power loom, and the spinning mule in the eighteenth century were just a few of the technical innovations that increased productivity. As iron and steel production became a reality, large-scale mechanization was possible for the first time.

The product of this innovation was either sold at home or easily shipped abroad. Demand from larger markets led to improved transportation systems. New agricultural techniques decreased the vulnerability of food crops to bad weather. There were improvements in fodder crops, with a subsequent rise in meat production. Coal was used as fuel, and the implementation of fire-resistant materials (brick and tile) produced by coal heat led to a drop in the frequency of disastrous city-based fires. Consequently, resources needed for rebuilding were conserved. A new belief in the principle of resource conservation paralleled the development of both insurance and government-sponsored food surplus stockpiles.

Quarantine measures helped to eliminate the plague after 1720. The population of London increased from 20,000 in the year 1500 to 500,000 by 1700. A relatively wellfed workforce, now using these technologies for achieving unanticipated new levels of productivity, began to alter its perspective. People became healthier, stronger, better-rested, and more comfortable. Looking anew at their surroundings, the citizenry wondered about the true limits to growth. Although meas-

---

* During the Middle Ages, trying to use mathematics to predict an outcome made as much sense as it would now to use modern probability to predict the winner of an election 100 years from now.

† Even Albert Einstein, criticizing the statistical approach to particle physics, said, "God does not play dice with the universe."

uring quality of life was generations away, a collective sense suggested that it could and must be evaluated.

The productive climate and the improved standard of living excited intellectual initiative. No longer seen as heretical, the enterprising spirit was now respected and encouraged. As opposed to the closely guarded estimates of population size, technical know-how in the marketplace and the incipient halls of science were not restricted to a few geniuses, but shared by many. These new industries excited artists e.g., Turner and the Impressionists. This was a time when old facts long accepted without proof were unceremoniously discarded. Now, new thinkers would endeavor to answer questions by querying nature directly, bypassing the traditional appeals to monarchs. However, even the most basic information on the citizenry of England itself was absent.

Reasoning from a Sample

The rise of capitalism with its need for market-size estimates required new knowledge of the population's demography. However, in the mid-seventeenth century, even the most basic of facts remained out of reach For example, no one knew the size of London's population; some believed that the city had over two million residents (an exorbitant estimate). The monarchy had a particular interest in this question because of their need to tax the citizenry at a rate the public could bear.

John Graunt's (1620-1674), *Natural History and Political Observations on the London Bills of Mortality* in 1662 was the first modern work in demography. Prior to its appearance, data on the number of deaths in London had been available in the *London Bills of Mortality* [8]. However, no one had actually undertaken a study of this data. Graunt's reviews of these records, and his subsequent careful deductions based on his analyses, revealed new observations and generated novel hypotheses about London death rates. Graunt's singular contribution was to establish the value of careful observation of imperfectly collected human data. [*]

He produced several unique computations, e.g., the process of counting burials to estimate the proportion of deaths. From this preliminary work, Graunt showed that the widely circulated but unsubstantiated speculation that millions of people lived in London was a profound overestimate. His effort established a universal registration of births and marriages, not for religious purposes, but for the purposes of accurate reports on population size to the government and citizenry. Graunt initiated work on the first lifetable, and was honored by nomination to the Royal Society [9].[†]

William Petty received impetus from Graunt's early tabulations, and together they labored to develop lifetable methodology, a procedure that permitted a crude estimate of death rates in London. Under their auspices, information was collected on both the number and causes of deaths, producing the first scientifically

---

[*] The data for the bills was collected by women who were commonly elderly, inebriated, open to bribes, and ignorant of medicine. See Sutherland, referenced at this prologue's end.

[†] It is alleged that King Charles II himself nominated Graunt for fellowship in this august group. See Sprat T. (1722) *History of the Royal Society*, London, p 67, or, more recently, Kargon, R. (1963) *John Graunt, Francis Bacon, and the Royal Society: The Reception of Statistics. Journal of the History of Medical Allied Sciences*. **18**: 337–348.

based cause-specific death rate estimates. Finally, the number of deaths from bubonic plague, consumption and "phthisis" (tuberculosis) could be quantified and followed over time [5].

This was a seminal time in statistics. Prior to the determinations of Graunt, the purpose of counting was simply to take an inventory, with no interest in, nor methodology for, inference. The work of Graunt and Perry held out the idea that there were circumstances in which one could extend results from samples to populations. This notion, so critical to the application of statistics to medicine generated rapid development in the new field. Huddes book *Annuitties* appeared in 1671. Petty's *Political Arithmetic* appeared in 1699, and Greogeory King's *Nature and Political Observations* in 1696. Charles Davenaut's *Discourses on the Public Revenues* (1698) [*] was followed by the first census in modern times, which took place in Ireland in 1703 [10]. Thus a period of slow development produced a critical mass of new thought, producing an eruption of new concepts and products. This cycle of slow development followed by rapid, indeed, sometimes chaotic and unchecked growth can be seen most recently in the development of air travel and the evolution of the modern computer.

Political Arithmetic

However illuminating these first demographic investigations were, the innovative workers behind them were not known in their contemporary world as statisticians. That term, derived from the Italian *statistica* for "statesman" was reserved for constitutional history and political science [5]. The contemporary term for the incipient demographic work of Graunt and Perry in the early 1700s was *political arithmetic* [5], defined as "the art of reasoning by figures, upon things related to government.[†]

It was John Sinclair who argued that the term *statistics* should be usurped to describe the process by which one inferred new meaning about the state of human affairs and interrelationships:.

> the idea I annex to the term is an inquiry into the state of a country, for the purpose of ascertaining the quantum of happiness enjoyed by its inhabitants, by the means of its future improvement…

However, the political arithmeticians at the time soon found themselves embroiled in an intellectual controversy that endures to the present.

The Role of Religion in Political Arithmetic

The collection of this first vital statistics data by Graunt and Perry instigated not merely a new collection of queries but controversy as well. The very nature of their work shattered the old order of looking to the monarchy or diviners for insight into the social order of culture. However the inquiries of these "political arithmeticians"

---

[*] It remains a point of contention as to whether this political arithmetician was a grandson of William Shakespeare.

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

declared that social issues could be addressed through the examination of data. This new approach was fraught with major political consequences that became clear when some suggested that a new list of questions could be addressed by the demographers.

Although the initial sampling work started as simple tabulations of the total number of people in London by gender, other more interesting questions rapidly followed: "Why are there more burials than christenings? How many men are married? How many are fighting men? How long does it take to replenish housing after a wave of the plague? How many men ignore their civic duties? In what proportion do men neglect the orders of the church?" The techniques used by the demographers went through a series of refinements in attempts to answer these questions. However, the consequences of sample-based answers to these politically volatile questions produced a series of hotly debated answers. These debates generated the question of who was best able to provide the interpretation of this controversial data, the political arithmeticians, or the cultural thought leaders who had vested interests in the answers.

It is impossible to understand the world of seventeenth to eighteenth century England without paying explicit attention to the overwhelming issue of religion [6]. Religion was not seen as a private matter at the time but as the vital, sustaining bond that held society together. The foundation of all political organization, it permeated everyday discourse, education, social interaction, organization, and all matters of public commerce. Therefore new data-based queries rapidly escalated into controversies involving religion and the spiritual state of England, which in turn had the potential of disrupting the function of the state and established cultural relationships.

The clerics themselves stood disunited on the important religious issues of the day as they found themselves mired in bitter internecine disputes. In the sixteenth century, the general struggle between the Roman Catholics and the "Reformed" or Protestant churches had been resolved on the basis of a compromise under the Tudors.[*] The accession of Queen Elizabeth I to the throne of England in 1558 shattered this arrangement, marking the decisive victory for Protestantism [11].

However, the fabric of Protestant leadership threatened to become unraveled from new Puritan pressure. This new sect held that the work of Protestant reformation was incomplete and pushed for more changes that were unacceptable to new Protestant dogma. By the middle of the seventeenth century the Puritans were a large and broad-ranging group in English society, wielding profound influence within their local communities. With its strong patriotism and fierce anti-Catholic creed, Puritanism became a formidable force to be reckoned with in trade and commerce. Additionally, Puritans reached into the stratus of aristocracy, influencing the larger landowners and lawyers who populated the Houses of Parliament. However, at this point, the Puritans managed to divide themselves on matters of church organization.

---

[*] It was hoped that the compromise of declaring the monarch as the heard of national Anglican church, itself part Catholic (High Church) and part Protestant (Low Church) in structure, doctrine, and dogma, would provide a lasting solution.

The strong religious–culture link, in concert with inter-sectarian conflict, meant that changes in influence among the religious sects would be transmitted through the fabric of English culture. Thus, all intellectual work was interpreted in this religiously polarized environment, and the competing religious philosophies spilled over into contentious interpretations of the early demographers' work. Even John Graunt's reputation was besmirched by his conversion to Catholicism late in life [5].* It is easy to appreciate the irony in calling these early demographers not statisticians, but political arithmeticians.

However, the development of sample-based data collection continued. Throughout this period, the demographers' technical problem of estimation was taken up by the mathematicians Neumann, Halley, DeMoivre, Bernoulli, Euler, and other mathematicians throughout Europe. This work, further developed by Poisson and Laplace became the foundation of the laws of probability and the inception of the mathematical science of statistics.

Probability and the Return to Order

Probability, discounted as an alien effort by the deity-centric cultures of the Dark Ages developed its first real blooms in the 1600s. By the seventeenth century, as Graunt and others developed the concept of vital statistics, and games of chance continued to be the rave in England and France, gifted observers began to use the data from each of these endeavors. For the first time, this information was collected into datasets as these analysts explored the possibility of producing reliable predictions. The parallel development of mathematical notation sufficient to capture the reasoning process of these experts permitted important progress. 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. His work, completed toward the end of the seventeenth century by Laplace, led to the conclusion that the mean of a small sample of data will approach a recognizable population mean in a predictable fashion.† This was the genesis of modern probability theory [5].

The cultural philosophy that had thus made no room for the role of randomness in the occurrences of life, and had previously dismissed attempts to predict the results of games of chance, had itself evolved. However, it continued to see the predictive, data-based calculations of gambling as disruptive. The DeMoivre-–aplace theorem allayed this concern by demonstrating that random events followed their own laws; outcomes beyond the control of man were not unfathomable, but instead demonstrated an overarching order. This long-term view of random events revealed a stability that found a natural home in the religious–centric world.‡

---

* Graunt's work was criticized; he himself was subjected to the outrageous accusation that he was responsible for the great fire of London

† This was known as the DeMoivre–Laplace theorem, and is now recognized as the weak law of large numbers.

‡ De Moivre, although impoverished his entire life and forced to make a living helping gamblers, was a very likable man and a good friend of Isaac Newton. De Moivre is believed to have died of somnolence, sleeping longer and longer each day in his old age until he finally did not wake up.

The drive to use mathematics to provide a clearer view of the world continued through Newton's work to the present day. However, the acceptability of the inclusion of probability in this quest was provided during the 17th century by DeMoivre and Laplace. Their results gave the process of studying random events an order, and therefore offered the world a lens though which it might gain an elevated perspective of the laws of the universe.

"Let Others Thrash It Out!"

As probability offered the world order by identifying how random events could be predicted, the issue of who would be the best interpreter of data continued to befuddle scholars of the time. The arguments took an interesting turn in the 1830 when a proposal was made to form a statistics section of the British Association for the Advancement of Science. Under the august leadership of Thomas Malthus, a subcommittee was created to answer the question, "Is statistics a branch of science?"

The distinguished committee readily agreed that the process by which the field of statistics collected, organized, and tabulated data was indeed a science. However the question, "Is the statistical interpretation of the results scientifically respectable?" produced vibrant polemics. The anti-inference sect won this debate. The decision to imbue the notion of inference with the respectability of science would have burdened statisticians with the responsibility for interpreting politically sensitive data accurately; this was a task beyond their abilities, since the science of inference had not yet been developed.

They repeated their victory a few years later in 1834 when the Statistical Society of London (later to become the Royal Statistical Society) was formed. Their victory was symbolized in the emblem chosen by the society — 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" [12].

Although this action appears out of step with current thinking, the decision provides insight into the prevalent perspective 200 years ago. At the time, statistics was not widely accepted as a science. Those working to correct this did not want to overwhelm the new discipline with politics. Permitting data interpretation to be incorporated as part of statistical science, with its social, economic, religious, and political undertones would provide the tendentious, nonscientific perspective the society hoped to avoid. Thus it was excluded. However, by 1840, the society began to push hard against this limitation.

Early Experimental Design

While the political arithmeticians developed and defended their work in the intellectual cauldrons of the urban environment, much of the development of eighteenth and nineteenth century experimental science took place in agricultural field studies. The parallel progress in experimental science did not rely on the advances of the probabilists and demographers/data analysts of the time. The agricultural science work was not a matter of the mere tabulation of data with associated inference, but of controlling the application of an intervention (e.g., a new seed). The process of designing the experiment to minimize any ambiguity of its conclusions did not

draw on mathematics so much as it did on the powers of observation and deductive reasoning.

In 1627, Francis Bacon published an account of the effects of steeping wheat seeds in nine different "nutrient mixtures" on germination speed and the heartiness of growth [13].* One hundred fifty years later, a body of useful contributions to experimental design was constructed by a relatively unknown experimentalist.

Agricultural Articulations

In 1763, a young man, Arthur Young, inherited a farm in England. Within eight years, this agronomist had executed a large number of field experiments, publishing his conclusions in a three-volume book, *A Course of Experimental Agriculture* (1771). With clear insight, he articulated ideas that are the basis of current experimental methodology.

Young expressed the importance of surveying the available data, and each of his volumes began with a literature review. He paid particular attention to biases that were accepted as truth because they were expatiated by "authorities", frequently providing examples of authors who slanted the presented data to support their favored conclusion.

Additionally, Young stressed the importance of comparative experiments, insisting that, when comparing a new method and a standard method, both must be present in the experiment [14]. However, he recognized that, even in comparative experiments, many factors other then the experiment's tested intervention influence the final outcome. Soil fertility, drainage, and insects were just a few of the factors contributing to the yields of experimental plots. Because the overall impacts of these extraneous factors had a variable effect, increasing yields in some years while decreasing them in others, the results of a single experiment in one year could not be completely trusted. Young therefore concluded that experimental replication was critical in agricultural work, often replicating his experiments over each growing season for five years [14].

Additionally, Young was careful to measure the end result of the experiment accurately. When it was time to determine the experiment's outcome, all expenses that could be traced to the intervention being tested were recorded in pounds, shillings, pence, halfpennies, and farthings[14]. At harvest time, one sample of wheat from each of the control field and the treatment field was sent to market on the same day to determine the selling price.

Finally, he recognized the dangers of experimental result extrapolation, warning that his own conclusions about the influences of crop development and growth may not apply to a different farm with different soil and land management practices. By carefully noting that his results would not apply as a guide to long-term agricultural policy, he stressed the pitfalls of unjustified inference from a sample to a population [14].

These important principles of experimental design (review, control, reproducibility, and inference) focus more on the logical infrastructure of the experiment

* Bacon concluded that urine was the most effective "nutrient" mixture.