Jessica Hagelstein
March 15, 2005
BE.104

In describing the evolution of statistical analysis in his book The Lady Tasting Tea, David Salsburg illuminates the ways in which the development of statistical analysis was shaped by the personal and political circumstances surrounding its pioneers. Each chapter focuses on the unique story of a particular statistical method and its discovery, providing the reader with a vivid sense of its temporal relationship and importance to the growing body of statistical knowledge. Though these discoveries are not always presented chronologically, a clear path emerges from abstract mathematics in the late nineteenth century to current applied statistics. Salsburg's personal knowledge of many of the founding fathers of statistical analysis draws the reader into this tight knit community and adds to an understanding of the genius that brought the science out of the air and into the hands of millions around the world. He continually brings to light the pervasive nature of statistical analysis and its crucial role in so many, often unexpected, realms of contemporary society.

Today, many statisticians work in epidemiology, the biomedical sciences or in the business sector, but the first statisticians were, in fact, mathematicians interested in solving abstract problems. While earlier discoveries in statistics were made by eighteenth century mathematicians like Gauss and Bernoulli (Salsburg, 16), it was not until the late nineteenth century that Karl Pearson and his rival, R.A. Fisher, began to elucidate the fundamentals. Pearson began by collecting massive amounts of data and analyzing them without a physical problem to solve, developing the first statistical model of a skew distribution and its four key parameters. Pearson believed, correctly, that his observations were a random sampling from the actual distribution that existed in nature, and this distribution, not the data themselves, was the important result (17). However, his error lay in the assumption that his four parameters could actually be determined, that they were the same for the distributions of both the sample and the population. As Fisher later pointed out, these parameters could only be estimated for the population distribution, never known (35). Yet, this did not diminish the importance of Pearson's

assertion that the key aspect of an experiment was not the data obtained, but their distribution, a basic tenet of statistical analysis. Pearson is also responsible for giving statisticians the chi-square measurement and the basics of hypothesis testing, both of which are heavily used today (96). Salsburg charts the animosity of these two men toward each other throughout their careers. Through his early work with crop variation data, Fisher developed sound formats for experiments, including randomization, as well as the role of variance and co-variance in statistical analyses (48). However, because of his earlier efforts to correct Pearson's theories, Fisher's work was consistently rejected from *Biometrika*,[1] one of the earliest and most well known journals of statistical literature (35). Pearson's refusal to publish Fisher's work in his prestigious journal inhibited the spread and notoriety of the latter's accomplishments and their usefulness to statisticians. Pearson's early dominance over the field ensured his methodology was initially enforced, but his son, Egon Pearson, and R.A. Fisher continued to disprove many of his theories and saw the beginnings of applied statistics.

William Sealy Gosset, however, found himself in Pearson's good graces and was able to publish his theories in *Biometrika* under the name "Student." Though his major contribution to statistics is based upon a correction to Pearson's four initial parameters, *Biometrika* published Gosset's paper in 1908 in which he introduced what is now known as "Student's t-test" (Salsburg, 27) He corrected Pearson's method by saying it was enough to estimate the parameters of a distribution and analyze their ratios, no longer requiring knowledge of the origin of the data or its true standard deviation (σ) (Salsburg, 30). Instead, the standard deviation of the population is estimated by the standard deviation of the sample (s) (Schork, 81). Gosset created a family of distributions, the t-distributions, which measure how good an approximation this is (Salsburg, 30). In order to prevent iterations in the estimates of the four parameters, Gosset assumed the original data obeyed a normal distribution (Salsburg, 30), thus saving future statisticians hours of agonizing calculations (especially those operating without computers). The t-test left a

[1] Fisher would later attack Jerzy Neyman's work on hypothesis testing and confidence intervals at a 1934 meeting of the Royal Statistical Society (120) because it presumed a correction to one of Fisher's own models. Despite receiving the same treatment from Pearson for the same reasons, Fisher refused to acknowledge the significance of Neyman's work and its potential to further his own theories.

lasting impact on the statistical community, and it remains a pervasive tool that allows analysts to determine whether data sets are significantly different.

The 1930's was a decade of great advancement in statistical methods, but it was also a time of great adversity for European scientists working under the oppressive, fascist regimes of Hitler and Stalin. As Salsburg makes clear throughout the book, the political tastes of these two heads greatly hindered the growth of statistical analysis. As in Allied countries, the Axis powers turned many of their scientific minds toward military design, creating the infamous, yet technologically superior, *Luftwaffe* and U-boats. However, Hitler and Stalin would only permit those scientists who fit their ethnic preferences and displayed total loyalty to their governments to conduct research and live within the territories they controlled. Soviet scientists were forced to dissociate themselves from the study of statistics because it did not conform to Marxist ideology, which said nothing occurred by accident (Salsburg, 148). In 1938, the Soviet scientist Andrei Kolmogorov simplified a critical theoretical problem posed by Fisher in his "Studies of Crop Variation III" by developing time series, or "stochastic processes" (144). While Kolmogorov was forced by the Marxist Soviet government to stop all research in statistics and pursue only theoretical mathematics, Norbert Weiner of MIT found military applications for these stochastic processes which greatly aided the Allies. Other scientists, such as Chester Bliss, inventor of the fifty percent lethal dosage measurement (LD 50) (76), were unable to fall from the Soviet government's radar simply by changing their title. Bliss was an American scientist working with insects in the Soviet Union during Stalin's rise to power. He eventually came under suspect by the government and was forced to flee the country, despite appearing in court to defend his research (80). Jewish intelligentsia throughout German controlled territories found themselves in danger, and many, like Richard von Mises and Emil Gumbel, were fortunate enough to flee the regime (58). Gumbel was hired at Columbia University and is one of many refugee scientists whose efforts furthered the Allied cause rather than that of their native countries.

During the decades preceding and following World War II, political agendas dictated the types of statistical problems analysts researched. Governments forced statistics out of Fisher's crop fields and Pearson's jungles, away from random data

collection and esoteric theories, and pushed scientists to apply statistics to war efforts. Some statisticians, like F.N. David, were able to remain in the midst of their abstract mathematical models. A female British statistician working with the Allies, David developed combinatorics (Salsburg, 158), a complex counting method allowed the British government to estimate casualties and utilities damage from German bombs (157). Other Allied scientists of various specialties, many of whom had escaped from fascist countries, were grouped together to simply think logically about problems facing the armed forces. These groups solved logistical problems such as the most efficient ways to supply troops with food or the most detrimental way to use certain weapons against their targets (92). This approach became known as "operations research," and proved extremely useful to the Allies who valued their intelligentsia far more than their opponents. However, one of the most infamous examples of the success of operations research is the development of the atomic bomb; even the multitudes of scientists in the Manhattan project did not anticipate its ramifications. In the decades after World War II, W. Edwards Deming pioneered the incorporation of operations research into the business sector, using the idea of continuous quality control to revitalize the Japanese economy in 1947 (252). While the pressure of war time agendas led to many valuable technological advancements, it also produced those whose consequences were unknown to society until it was too late.

As war research occupied many of the established scientists in the United States during the 1930's and 40's, younger graduates grappled with the unemployment brought on by the Great Depression. Many graduates found themselves entering unexpected fields; statisticians ventured into the business world and businessmen learned to be statisticians. With unemployment rampant throughout the country, the government needed a way to measure the size and extent of its presence. To a modern reader familiar with the completion of census forms, the solution may seem obvious. At the time, however, applied statistics was still a very new concept, and it was not a statistician but a young economist with a scant training in mathematics who developed modern census methods. Morris Hansen, with advice from Jerzy Neyman, led the US Census Bureau away from lengthy surveys, which attempted to gather data on too many demographics,

and toward random surveys (Salsburg, 175). Hansen's ideas, later included in his[2] textbook Sample Survey Methods and Theory, emphasized simplistic random sampling and laid the foundations for sample survey methodology (176). In the late 1940's, Deming, a statistician, was brought to Japan by General MacArthur to help remodel Japanese society after American ideals (248). Through his teachings of continuous quality control, the Japanese economy revived in less than the expected five years (252) and has since surpassed that of the US in many respects. Deming's message was readily accepted by Japanese companies, unlike their US counterparts, and his work represents a major contribution of statistics to the business world.

Applied statistics thrived in the decades before and after World War II, producing solutions to a myriad of practical problems, but abstract mathematical proofs of these new statistical methods were initially neglected. During the 1950's and 60's, hastily developed applications of statistics such as estimation theory and Kolmogorov's stochastic processes were revisited and justified via complex mathematical proofs (Salsburg, 214). The esoteric nature of publications in the field of statistics[3] began to undo much of the progress toward socially beneficial statistics and created fractures in a community once tightly banded behind a common political cause. By the 1980's, this schism between statistics and society had aroused sufficient alarm from within the field that efforts were made to re-integrate statistics into the surrounding world. Out of this initiative came medical applications including biostatistics and epidemiology, which are now one of the central focuses of statistical analysis. Because these applications involve the use of human subjects in many cases, ethics and experimental procedure have become an integral consideration in every study performed.

Not only has its focus changed dramatically since its invention in the late eighteen hundreds, but the face of modern statistics has been greatly altered in the wake of several technological advances. After years of resistance, statisticians finally embraced the computer in the 1980's and developed many techniques, such as the Glivenko-Cantelli

[2] The book was actually a collaborative effort between Hansen and co-author William Hurwitz (Salsburg, 176).

[3] Sam Wilks led an initiative to bring prestige to statistical mathematicians and earn their inclusion in the mathematics community. In the process, as editor of the *Annals of Mathematical Statistics*, he set a trend in statistics publications, rendering them more and more abstract in the image of popular mathematical journals (214).

lemma, that are not possible without a machine that can perform iterative processes (Salsburg, 288). There has always been some degree of collaboration in statistics (such as Box-Cox transformations and Pearson's calculators), but modern statistics have become a joint effort among many, often international, scientists. For example, the fewest possible criteria necessary for a set of numbers to be classified as a normal distribution, known as the martingale, was initially proposed by Frenchman Paul Levy. However, the mathematical theory was developed by scientists from six different countries, including the Norwegian statistician Odd Aalen, Erik Anderson of Denmark and Richard Gill of the Netherlands who modified the theory and brought it to clinical trials (272). With the increased ease of information transfer via the internet and international travel, many more collaborations are seen in recent times than when Pearson and Fisher began their studeis with such individual approaches.

The nature of the problems studied by statisticians has changed in the century since the development of statistical methods, but many of the founding techniques are still applicable to modern analyses. A unique study[4] recently published in *Cancer Letters* examines the cytogenic effects on children of methylphenidate, commonly known as Ritalin. The authors of the study make use of common analytical methods such as confidence intervals, p-values and paired t-tests to determine whether there is a significant adverse health effect associated with therapeutic doses of this widely drug (El-Zein, 3). Experimenters collected data on three characteristics of the children's cells from blood samples taken before treatment with Ritalin and three months into the treatment process (El-Zein, 2). The amount of three different types of genetic mutations were measured from these blood samples: chromosomal aberrations (CAs), sister chromatid exchanges (SCEs) and nucleoplasmic bridges (NPB), determined from micronucleus (MN) assays (El-Zein, 3). In each case, the experimenters found a significant increase in the number of these genetic mutations after patients were treated with Ritalin. The significance of the difference between the levels of genetic mutations before and after is measured using a p-value, which was calculated to be zero to three

[4] Randa A. El-Zein, Sherif Z. Abdel-Rahman, Matthew J. Hay, Mirtha S. Lopez, Melissa L. Bondy, Debra L. Morris, Marvin S. Legator, Cytogenic effects in children treated with methylphenidate, Cancer Letters XX (2005) 1-8.

decimal places for each characteristic (El-Zein, 3,4). However, while this general trend holds for the twelve children who completed the study, the researchers admit larger sample sizes are necessary to confirm their findings.

The issue of sample size has been intimately linked to statistical analysis since its inception, and William Sealy Gosset was the first to see that analytical methods could still apply to small sample sizes (Salsburg, 29). By realizing that a sample was only representative of a population, Gosset developed the t-test which is now used as a form of hypothesis testing. The t-test is a form of hypothesis testing that examines the probability that accepting one hypothetical outcome (usually the null hypothesis) over another (the alternative hypothesis) will be an error. In Neyman's model, the null hypothesis is tested by the study, and nay other possibilities, or alternative hypotheses, are elucidated by default (Salsburg, 109). Hypothesis testing cannot prove one hypothesis; it can only potentially disprove the other (Schork, 173). According to Neyman, the power of a hypothesis test is the probability that it will detect the alternative hypothesis through rejection of the null (Salsburg, 109). In studies of pharmaceuticals such as Ritalin, the null hypothesis is that the drug has no effect on the subject, while the alternatives could include either harmful or beneficial effects. El-Zein et. al. performed paired t-tests on each set of before and after data to measure the significance, indicated by the p-value, of the difference between them. A paired t-test is advantageous because it reduces the variance in the data due to interpersonal differences, and so it was properly employed in this study to evaluate the significance of the difference before and after treatment in the same individual. However, there is a disadvantage to using a paired t-test versus a single t-test to compare two sets of data; in a paired t-test, the two data sets are reduced to one because the actual data used in the test is the difference between the two initial sets. This decrease in sample size leads to a decrease in the number of degrees of freedom of the test, which in turn reduces the power of the test.

The significance of the difference between the before and after data sets in the t-test is measured by a p-value, and both this value and the t-test are intimately linked to Neyman's confidence intervals (CI). In a normal distribution, ninety-five percent of the probability distribution lays two standard deviations away from the origin in both directions. Two standard deviations in a normal distribution (for a two-sided test) are

equivalent to 1.96, which is the t-value for a 95% CI with an infinite sample size (and so infinite degrees of freedom). If the sample size is finite (taken to be two hundred samples or less in t-tables), then the t-value corresponding to 95% probability will be greater than 1.96 (Schork, 419). The p-value can be determined by finding the range of tabulated t-values between which the calculated t-value lies and subtracting the probability of these t-values from one. Typically, a p-value of less than or equal to 0.05 is defined as the threshold for significance in a 95% CI. Above 0.05, the t-value lies within the interval, and the deviation of the data from the null hypothesis is not significant and the null hypothesis is retained. In a paired t-test, the formula for calculating the t-value depends directly on the mean of the differences between the two data sets and the square root of the sample size, and indirectly on the estimated standard deviation (s_d) of the difference set (Schork, 188). Using a t-table for a two sided test (as the number of genetic mutations could potentially increase or decrease) (Schork, 165), the researchers calculated a p-value of zero to three decimal places for each type of mutation (El-Zein, 4). Since this value is less than 0.05, it implies that the genetic mutations of the patients before and after treatment are significantly different, and thus the null hypothesis can be rejected. However, the problem with a p-value of 0.000 is that it implies a t-value for 100% probability, which is impossible to quantify because the normal distribution tails off to infinity. While the researchers were undoubtedly attempting to emphasize the significance of the damaging effects of Ritalin, their p-value presentation does not leave the reader confident in their understanding of the theory behind Gosset's t-distributions. This represents a problem in statistical analyses that Salsburg points out throughout his book, which is that researchers often confuse the application of a statistical method with its actual mathematical explanation. For example, a common misconception surrounding Neyman's 95% CI is that they allow the experimenter to say she is "95% confident" that her assumption of cause and effect is correct. In fact, Neyman's actual definition of a 95% CI is an interval in which the mean of a measurement, one of Pearson's parameters, would be found 95% of the time (Salsburg, 123). In fact, this study computes the 95% CI for the change in mean between the before and after data sets for each of the three blood analyses, in keeping with Neyman's original definition (El-Zein, 4).

In a related study[5] on the genetic toxicity of methamphetamines (METH) in humans, researchers also used a single t-test to analyze data sets on similar genetic mutations in humans, rats and bacteria. Only SCE and micronucleus (MN) assays were performed on human cells. Cells from self-identified METH addicts were compared to those of a control group matched for gender, age, alcohol consumption and smoking habits (Li, 235). This experimental design differs from that used in the Ritalin study where patients served as their own controls and blood samples taken from the same patient before and after treatment were analyzed (El-Zein, 2). In order to see how well the controls correlated to the exposed group, Li et. al. performed chi-square tests, or as Pearson called his invention, the "chi square goodness of fit test" (Salsburg, 96). They defined a p-value greater than 0.05 as indicating a significant difference between the two groups (Li, 235), and for each variable tested, no significant difference between the controls and those exposed to METH was found (Li, 239).

Fisher first proposed the idea of controlling certain factors in an experiment when testing the effects of fertilizers. He designed studies that would isolate just one potential variable and analyze the extent to which it was responsible for the observed effect (Salsburg, 46). He also introduced the scientific community to the concept of randomization in experiments, demonstrating how common confounding coincidences in fertilizer tests could be avoided. For example, when every other plot in a row or those plots along a diagonal are treated with fertilizer, there is a risk that these plots will lay along a water gradient such that the fertilizers will spread and their effects be obscured (Salsburg, 47). However, if the plots treated with fertilizer are randomized, the probability that each sample will be perturbed decreases. In each of the two pharmaceutical studies discussed, the sample pools were not entirely randomized because of the self-selecting nature of the exposed groups. In the Ritalin study, children who sought medical advice from the clinic at the University of Texas Medical Branch became eligible for participation (El-Zein, 2), while the methamphetamine study used METH addicts who were already seeking treatment at the Taipei City Psychiatric Center, as well as some users incarcerated at the Tu-Chen Detention Center for METH abuse (Li, 235).

[5] Jih-Heng Li, Heng-Cheng Hu, Wei-Bang Chen, Shih-Ku Lin, Genetic Toxicity of Methamphetamine In Vitro and in Human Abusers, Environ. Mol. Mutagen. 42 (2003) 233-242.

Clearly, each of these studies was localized to a specific area, also limiting the sample pool and decreasing the study's randomization.

Despite the changing nature of applied statistics, fundamental methods have remained a part of statistical analysis. Pearson's realization that the distribution of data sets provided the basis from which to draw conclusions regarding the outcome of an experiment was the key to the development of early hypothesis testing as well as the characterization of several different types of distributions, including the normal. Fisher's experimental designs including randomization are still followed, as much as possible given personnel and financial restrictions. Gosset's t-test and Neyman's 95% confidence intervals together form a powerful tool which allows experimentalists to determine the significance of their findings and make decisions regarding necessary action based on these findings. Salsburg's chronicle of the development of this science creates cohesion between many seemingly separate, genius innovations in both the mathematical theory behind statistical methods and their application to practical problems. As the two pharmaceutical studies discussed demonstrate, these fundamentals are widely used for biomedical applications, but Salsburg makes clear the integral role of statisticians in other sectors of society, including environmental studies and the business world. While many of the theories behind statistical techniques are masked behind a veil of complex formulae and mathematical reasoning, Salsburg is extremely effective in making the use and importance of these tools accessible to a much wider audience.

Bibliography

El-Zein RA, Abdel-Rahman S, Hay M, Lopez M, Bondy ML, Morris DL, Legator MS, Cytogenic effects in children treated with methylphenidate, Cancer Letters XX (2005) 1-8.

Li JH, Hu HC, Chen WB, Lin SK, Genetic Toxicity of Methamphetamine In Vitro and in Human Abusers, Environ. Mol. Mutagen. 42 (2003) 233-242.

Salsburg, David. The Lady Tasting Tea. Henry Holt and Company, LLC, New York, NY, 2001.

Schork, M. Anthony and Richard D. Remington, Statistics with Applications to the Biological and Health Sciences, Prentic-Hall, Inc., Upper Saddle River, NJ, 2000.