A Selected History of Statistical Practice

Chris Spatz

Hendrix College

This chapter was originally published by the Society for Teaching Psychology on their website as part of the E-book, *Psychological specialties in historical context: Enriching the classroom experience for teachers and students*, edited by W. D. Woody, R. L. Miller, & W. J. Wozniak. The book's 27 chapters cover histories of many of psychology's subdivisions. In addition, some chapters append one-page vignettes that are focused on a specific incident, event, or effect.

Fair Use

Recommended reference form:

Spatz, C. (2016). A selected history of statistical practice. In W. D. Woody, R. L. Miller, & W. J. Wozniak (Eds.) *Psychological specialties in historical context: Enriching the classroom experience for teachers and students.* Retrieve from the Society for the Teaching of Psychology website: http://teachpsych.org/ebooks/psychspec

A Selected History of Statistical Practice

Chris Spatz

Hendrix College

Opinions vary about the value of statistics. Here's a selection from the last paragraph of a 1901 article in the *Atlantic Monthly*:

> ...statistics are the first resort of the ill-informed. They may be of use in the concrete but there is little beauty in them, and with due respect to the public, the World Almanac is not the highest achievement in American literature... It was delightful unmathematical, unstatistical blood which did most worth doing of that which has been done. The Greeks attained to passable prominence without the trail of Arabic notation smeared across their national life or sullying conversation in Academe. The Elizabethans did much without referring to the decimal system. And Genesis was written before Numbers (White, 1901, p. 844).

At the other end of the distribution, from 1889, we have:

> Some people hate the very name of statistics, but I find them full of beauty and interest... Their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man (Galton, 1889, p. 63).

As it turned out, 20th century investigators bought into Galton's idea that statistics is a powerful technique for dealing with complicated phenomena. Statistics developed and was embraced because its techniques produced better understanding, regardless of the investigator's field of interest. Many disciplines and business operations contributed.

This chapter recounts a progression of changes in statistics during the past two hundred years. The changes provided solutions to limitations of that day's statistical techniques. Today, about one-fifth of the way into the 21st century, more change is on the horizon. These changes are quite apparent to those who study statistics and who recommend how they should be applied.

Early 19th Century

Numbers (0 – 9), commonly known as *Arabic* or *Hindu-Arabic* numerals, are the foundational elements of statistics. Invented in India before the 7th century CE, they spread westward to Arabic countries in the Middle East and North Africa. From there they were adopted in Europe by the 12th century CE (West, Griesbach, Taylor, & Taylor, 1982).

By 1834 there was enough interest in statistics in England for a group to found the Royal Statistical Society (Charter of the Royal Statistical Society, 1887). Early issues of their journal recount statistics on education, crime, and other social topics. In America in 1839, a group in Boston founded the American Statistical Society. Just three months later they renamed themselves the American Statistical Association (ASA), which continues today (Mason, McKenzie, & Ruberg, 1990). The statistical reports from the early 19th century would today be classified as descriptive statistics. Descriptive statistics endeavor to convey characteristics about a sample or population without any use of probability. Probability was well established in mathematics, but useful applications to statistics would come later in the century.

Adolphe Quetelet (1796-1874) was a Belgian astronomer of considerable fame and influence. (A passable English pronunciation is Kate'-ul) Quetelet established and directed the Royal Observatory in Brussels. Familiar with the Gaussian Law of Error, which predicts astronomical observations (today's normal distribution), Quetelet was the first to recognize that human and social data (such as heights and crime reports) mirror an approximately normal distribution (Boring, 1929). Thus, proportions associated with the normal curve could be used to predict the incidence of human and social phenomena.

Central Limit Theorem

By the latter half of the 19th century, researchers were finding the probability that a sample mean come from a population with a known mean. The probability figure came from the normal distribution, which was appropriate because of the central limit theorem (CLT). The CLT says that for any population of scores, regardless of their form, the distribution of means of randomly drawn samples will approach a normal curve as the sample size increases (Miles, 2005).

The most remarkable feature of the CLT is that it works regardless of the form of the original distribution of scores. The most serious concern that comes with the CLT is the question of how large a sample size must be. Nineteenth century researchers in all fields addressed this concern by using very large samples, often involving hundreds or thousands of observations. Another requirement of the CLT, hidden in the phrase *randomly drawn samples*, is that the samples must be independent of each other (not correlated). Of course, for many interesting questions, data are scarce and sometimes correlated. In such situations, some statisticians recognized that the CLT did not apply and probability figures obtained from the normal distribution were not trustworthy.

Francis Galton (1822-1911)

For Sir Francis Galton, one label won't do. Wikipedia uses 13, which range from *anthropologist* to *tropical explorer* and include *psychologist* and *statistician*. This overview will cover just one of his statistical contributions in use by many disciplines today: correlation.

Galton was enchanted with counting. He felt strongly that quantifying was the best path to better understanding of a phenomenon. But once the measurements were in, what to do with them? For example, he lamented the near universal attention to the mean:

> It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once (Galton, 1889, p. 62).

To remedy these deficiencies, Galton set about devising new methods. In particular, he wanted better ways to measure relationships among data sets. He complained that data sets were "entangled." The statistic that emerged from his effort to describe the co-relation between two variables was the predecessor to today's correlation coefficient. To examine what he called co-relation, Galton studied English scientists using the questionnaire method, which he gets credit for inventing. Early on, he listed head size down the first column of a sheet of paper and weight across the top row, creating what today is called a scatterplot. Galton could see that as head size got larger, weight increased. But how to convert the data into a statistic that described the relationship? Galton was a fair mathematician so he converted the raw scores in each distribution to standardized scores but failed to derive today's product-moment correlation coefficient. That task fell to his protégé, Karl Pearson.

If we ask what Galton wanted to accomplish with all his quantitative endeavors, the answer was that he wanted to improve the human condition by improving humans themselves. The first task though was to understand the forces that shaped humans and for Galton, the primary force was evolution. He was particularly interested in investigating the emerging field of genetics (Galton, 1901). Around the turn of the century, the feeling of progress for humankind was in the air, especially among intellectuals.

For most of the rest of this chapter, you might think of Galton as standing at the head of a long line of people who developed statistical techniques. The connections between those standing side by side were personal as well as academic. They knew each other, mentored each other, and argued, insulted, and supported each other. The result was statistics, the dominant way to investigate phenomena in the 20th century and beyond.

Karl Pearson (1857-1936)

Karl Pearson was 35 years younger than Francis Galton, who had no children. If you think of their relationship as father-son, you won't be far wrong. For example, Pearson published at his own expense a 3-volume biography of Galton after his death (Pearson, 1914, 1924, 1930a, 1930b).

Based on Galton's start, Pearson derived a better correlation formula, the product-moment correlation coefficient, and published it in 1896. It is widely used. Probably every discipline that uses quantitative data has correlation coefficients scattered in its literature. A correlation coefficient tells readers not only whether or not there is a relationship between two variables but also whether the relationship is positive or negative AND the strength of the relationship.

In 1900 Pearson published the chi square statistic, a technique that is also widely used in many disciplines. Pearson first used chi square as a goodness-of-fit test to determine the degree that observed data fit predictions from a theory. Often, the theory was that data from a population were normally distributed. A chi square test gives the probability that the observed data are consistent with those predicted by a normal curve. If the probability is very small, it is easy to conclude that the theory should be rejected. If the test produces a large probability, it means the data are consistent with what the theory predicts. Unfortunately, Pearson often interpreted large probabilities as evidence that the theory was correct. Of course, it is a logical fallacy to declare a theory is true just because it has not been contradicted. Later, Pearson used the chi square test to determine whether two variables were related, an application that is today's chi square test of independence.

In 1901, Pearson and W. R. F. Weldon, a geneticist, founded the journal, *Biometrika* with financing from Galton. They established the journal because the Royal Society of London, the principal scientific society in Great Britain, would not publish biology papers that contained statistics. *Biometrika* (bio and metric), in contrast, promoted the marriage of the two fields (Galton, 1901). Early issues of Biometrika report the relationship in flowers of petal length and sepal length (r = about .25) and the goodness of fit of the cephalic index of 900 modern Bavarian peasants to a normal curve (good fit).

William S. Gosset (Student) (1876 -1937)

William Sealy Gosset studied chemistry and mathematics at Oxford University, graduating in 1899. He went to work for Arthur Guinness, Son, & Co., where he gathered data about the brewing process and made recommendations (Biography 12.1: William S. Gosset). Gosset's samples were small, and his measures often correlated with the next measure taken in a sequence. Thus, the central limit theorem could not be relied on for accurate probabilities (Salsburg, 2001).

Gosset met Karl Pearson in 1905 and then spent the academic year 1906-07 with him in London, working to find a statistic that would be accurate for small, correlated samples. The result was the *t* test, which he published in 1908 under the pseudonym, *Student*. The pseudonym was necessary because the Guinness Company had a rule that their employees could not publish their discoveries (Salsbury, 2001). (The rule at Guinness seemed to be, "Publish AND perish.") Gosset's biggest contribution was to derive a new curve of probabilities, today called the *t distribution*. Gosset originally labeled the distribution *z* (Student, 1908).) The *t* distribution is a family of distributions that differ according to their degrees of freedom.

Ronald A. Fisher (1890-1962)

Sir Ronald Aylmer Fisher (who was knighted in 1952) has been referred to as the father of modern statistics. He was a one-in-a-billion scientist whose brilliant work in genetics has been overshadowed by his revolutionary presence in statistics. In genetics, Fisher was responsible for showing that Mendelian inheritance was not in conflict with evolution by natural selection (Yates, 1981). (For several years after 1900, genetics and evolution were considered incompatible.) Fisher also showed that natural selection is primarily a within-species phenomenon and that sexual selection of mates is a powerful force in evolution (Edwards, 2005).

In statistics, Fisher revolutionized the way we analyze data from experiments with a technique called *analysis of variance* (ANOVA). Student's *t* test permitted comparisons of two levels of one independent variable for small or large samples. ANOVA provided a technique for assessing the effect of more than two levels of an independent variable and for more than one independent variable. Interactions among variables are common in nature, and ANOVA provided a quantitative way to measure them with a factorial design. ANOVA also delivered a way to test the significance of correlated levels of an independent variable. Finally, analysis of covariance (ANCOVA) provided a way for researchers to remove the effects of an unwanted variable from the data analysis. It is not too strong to say that Fisher's approach to statistics profoundly affected every discipline that analyzes quantitative data.

A complete understanding of a difference between means can never be based only on the absolute size of the difference; the variability of the scores that produce the means must be taken into account. This principle is particularly clear in Student's *t*-test formula, which is a difference between means divided by the standard error of the mean, which is a pooled measure of the variability of the scores. Fisher recognized that both numerator and denominator of the *t* test were expressions of variability. The formula for ANOVA's *F* test is a numerator (a measure of variability produced by different means) and a denominator (a pooled measure of variability of all the scores). Thus, an analysis of variance.

Statistics such as ANOVA and the *t* test end with a probability. The accuracy of the probability depends on whether or not assumptions hold that were used to derive the test. Two assumptions that are required for accurate probabilities when testing independent sample means are that the scores come from normally distributed populations and that the variances of the populations are equal (Kirk, 2013).

As for Ronald Fisher himself, he was a small, wiry man with red hair and poor eyesight. As a youth, he won a mathematics scholarship to Cambridge where he was an outstanding student. He helped found and was chair of the Cambridge University Eugenics Society. After graduation, he married Eileen Guinness, a cousin of the Dublin family who owned the Guinness brewery. The Fishers were practicing eugenicists; they had eight children. Rich, however, they were not. The Fishers lived frugally, practicing subsistence farming for several years early in their marriage. Leonard Darwin, a son of Charles Darwin, was a mentor and advisor. Darwin provided work for Fisher by commissioning book reviews for publications for the Eugenics Society of London. (Box, 1978).

In 1919, Fisher accepted a post as chief statistician at Rothamsted Experimental Station, where he and a staff set about analyzing decades of data on crops and animals. He also created new experimental designs that controlled for many of the extraneous variables in agricultural research. Perhaps most importantly, he compiled his new statistical methods into a book, *Statistical Methods for Research Workers* (1925). The book was quite successful, introducing many research workers to ANOVA and going into 14 editions. Fisher was an idealist, committed to establishing truth and the advancement of humankind (Box, 1978).

Significance and the .05 level

Where to draw a line on the continuum of probabilities that separates “data very rare if the null hypothesis is true” from “data consistent with the null hypothesis”? This is the question of how to separate statistical significance from NS (not significant). The widely used cut-off point is $p \leq .05$. If the probability of an observed difference is .05 or less, scholars typically reject the null hypothesis and say the result is *statistically significant*.

Specifically, what does $p \leq .05$ mean? In statistics, p is the probability of the differences observed, if it is the case that the null hypothesis is true. That is, $p \leq .05$ means that if the null hypothesis is true, data with a low probability occurred. Thus, p is a conditional probability about observed data and not a probability about the null hypothesis. There are many ways to misinterpret the meaning of $p \leq .05$. (See Spatz, 2016 for a summary.)

In the 19th century, the word significant meant *signified* or *something is there*, much as a sign indicates something to see. During the 20th century, the word significant evolved to mean important. In statistics, *significant* carries its 19th century meaning. A significant result means only that a very low probability was found; there is something worth noting. The question of whether the result was important must be answered with analyses other than statistical. Lamentably, some fail to recognize that *statistically significant* does not ensure importance (Stevens, 1995).

Perhaps the earliest recommendation for a significance level was a 1910 article in the *Journal of Agricultural Science*. Thomas B. Wood, the principal editor of the journal, advised researchers to take "30:1 as the lowest odds which can be considered as amounting to practical certainty ... that the difference.... is significant" (Wood & Stratton, 1919, p. 483). Odds of 30:1 translate into a probability of .03. Fisher's *Statistical Methods for Research Workers* (1925) included tables with probabilities of .05 and .01 and an offhand comment "We shall not often be astray if we draw a conventional line at .05" (Fisher, 1936, p. 83).

Jerzy Neyman (1894-1981) and Egon Pearson (1895-1980)

Jerzy Neyman was born to Polish parents while they were in Russia. Home-schooled until he was 10, he was fluent in five languages. After his father died, the family moved to Kharkov in the Ukraine, where he finished university and taught mathematics. In 1921 he moved to Poland as a statistical assistant at the Agricultural Institute. He soon moved to Warsaw where he received a PhD in 1924. In 1925 he won a fellowship to work with Karl Pearson in London. While there he developed a friendship and collaboration with Pearson's son, Egon, who had grown up under his father's influence and was a lecturer in his father's Department of Applied Statistics. In 1938 Neyman moved to the University of California, Berkeley and later founded the Department of Statistics. He remained in Berkeley the rest of his life (Howell, 2005, Zabell, 2001).

Neyman's collaboration with Egon Pearson in the 1930's produced what has come to be known as Neyman-Pearson Inference (Nickerson, 2005). Besides the null hypothesis (usually that the population means are equal), there is an implied alternative hypothesis: the population means are not equal. Either of these hypotheses could be true or false. Raising the alternative hypothesis out of the shadows leads to a now-familiar 2 x 2 table of outcomes that can result from a statistical analysis:

Decision made on basis of sample data	Truth about the Population	
	Null hypothesis true	Null hypothesis false
Reject null hypothesis	***Type 1 error***	***Correct decision***
Retain null hypothesis	***Correct decision***	***Type II error***

This approach to statistical analysis is now known as *null hypothesis significance testing* (NHST). With this approach, Type II errors receive attention. You make a Type II error when you fail to reject a false null hypothesis. If the probability of a Type II error is low, the test is said to be powerful. Statistical power is defined as

(1 minus the probability of a Type II error). Thus, a powerful test is one that is likely to detect a real difference. Now, when an experiment produces data with a probability of a Type I error of around .05 or greater, the next step could be to ask if a Type II error was likely. If the answer is “likely,” factors that contribute to a Type II error can be examined to help in interpretation and perhaps to redesign the experiment.

NHST became the method to use in all fields that employ a quantitative approach. This is certainly true for medicine, business, finance, manufacturing, biology, economics, sociology, and psychology. It is not too strong to say that NHST statistics and experimental design came to dominate epistemology in the 20th century.

Jerzy Neyman also worked out the mathematics of confidence intervals and presented them in a talk to the Royal Statistical Society in 1934 (Salsburg, 2001). A confidence interval is a range of values within which a population parameter is expected to be. The degree of expectation is quantified, often as 90, 95, or 99 percent confidence. In addition, confidence intervals also provide the information needed to reject or retain the null hypothesis.

Neyman gets credit for the development of confidence intervals as analytic tools, although Egon Pearson and William Gosset addressed the subject of interval estimates earlier. Increasingly popular today, confidence intervals were questioned to begin with and caught on slowly (Salsburg, 2001).

Non-parametric Statistics

Accurate probabilities from statistical tests such as *t* tests, ANOVA, and confidence intervals depend on whether the assumptions the tests are based on are satisfied. Two common assumptions are that the scores are from normally distributed populations and the variances of the populations are equal. These conditions can be met if the data are continuous, quantitative measures but they certainly don’t hold if the scores are ranks (1 to *N*) or nominal or categorical data. Nonparametric tests, which are also called *distribution-free tests*, were developed to provide accurate probabilities for situations where the assumptions required for the *t* test, ANOVA, and confidence intervals do not hold.

The number of nonparametric tests is considerable. Lunneborg, (2005) lists 57 basic, distribution-free tests and points out that the availability of fast, inexpensive computing has fostered many more. Many of these tests are tailor-made to match particular situations. The oldest of nonparametric tests still in use is Pearson’s chi square test.

Jacob Cohen (1923-1998)

In 1962 Jacob Cohen published an article in the *Journal of Abnormal and Social Psychology* that slowly gained influence in statistical practice. Cohen analyzed 70 studies from 1960 and noted that researchers did not determine the power of their statistical tests. (A powerful test has a high probability of rejecting a false null hypothesis.) When he calculated the power of the statistical tests in those 70 studies, he found that the tests were not powerful. Only 1/5 had enough power to detect a small effect; one-half could detect a medium effect and 5/6 could detect a large effect (Cohen, 1962). This article also showed ways to calculate effect size indexes and offered recommendations for values to consider small, medium, and large. Cohen helped promote awareness of power and effect size indexes with subsequent books and articles that were conversational and easy to understand. (e.g., Cohen, 1994). Today, calculation of effect size and awareness of power are common.

Meta-analysis

A big problem with the NHST procedure is that it is not cumulative. NHST tests start each analysis from the same point, which is to assume the null hypothesis is true. There is no way to incorporate the results of previous studies into the test. Before the advent of meta-analysis, obtaining an overall view consisted of a researcher reviewing all the studies on a topic, making judgments about their quality, looking at the outcomes, and declaring an overall view. It wasn't a very objective method.

In 1976, Gene Glass published an article that named and explained *meta-analysis*. This analysis of analyses is an objective way to amalgamate into one conclusion the results of many studies on a topic. At its simplest, meta-analysis produces a pooled effect size from the effect size indexes of many studies. Glass used this new technique to summarize the results of more than 300 studies and showed that psychotherapy is effective compared to a no-treatment control group. The difference between the two groups was about one standard deviation, which is a large effect size (d) equal to about 1.00 (Cohen, 1962). Meta-analysis was immediately popular and spread to other disciplines. In 1989 Stanley and Jarrell developed meta-regression analysis as an objective way to synthesize regression results from many studies.

Into the 21st Century

Over the years various statisticians and philosophers of science raised objections to null hypothesis significance testing (NHST) and how it is used in practice. As the 20th century came to a close and the 21st began, more and more objections were voiced (Nickerson, 2000). Some argued that NHST should be banned from use (Hunter, 1997). Others defended its use but acknowledged some of the objections. Besides occasional misinterpretation due to ignorance, three major objections are:

1. The best outcome of a NHST test is that the populations are different. But for probably all empirical populations, the idea that they are exactly equal seems easy to dismiss. Thus, the best NHST can do is to lend support to what is obvious to begin with: the populations are different.

2. The probabilities associated with the most common NHST tests are based on assumptions about the sampled populations such as normally distributed and equal variances. The assumptions are seldom verified but when checked, often found to be unjustified, leading to tabled probability figures that are wrong for the data being analyzed (Erceg-Hurn & Mirosevich, 2008).

3. A NHST analysis always starts with a null hypothesis of no difference, regardless of previous results on the same topic. Thus, NHST fails to incorporate the scientific ideal of cumulative knowledge.

As a result of these and other objections to NHST, some statisticians and researchers are not only asking the question, "How shall we analyze data?" but also addressing the broader question, "How can we improve our chances of publishing results that are dependable?" A summary of the situation can be found in an editorial on *p* values in the *American Statistician*, a publication of the venerable American Statistical Association (Wasserstein and Lazar, 2016). Statistics in the early years of the 21st century is again in transition, much like it was in the early years of the 20th century.

References

Biography, 12.1: William S. Gosset (1972). *Dictionary of Scientific Biography* (pp. 476-477). New York, NY: Scribners.

Boring, E. G. (1929). *A history of experimental psychology.* New York, NY: Century.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York, NY: Wiley.

Charter of the Royal Statistical Society. (1887). *Journal of the Royal Statistical Society*, 50, 1-4.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.

Cohen, J. (1994). The earth is round, $p < .05$. *American Psychologist*, 49, 997-1003.

Edwards, A. W. F. (2001). Fisher, Ronald A. (1890-1962). In N. J. Smelser & P. B. Baltes (Eds.). *International encyclopedia of the social and behavioral sciences* (Vol. 8, pp. 5698-5701). Oxford, UK: Elsevier.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63, 591-601.

Fisher, R. A. (1925). *Statistical methods for research workers.* London, UK: Oliver and Boyd.

Fisher, R. A. (1936). Statisti*cal methods for research workers* (6th ed.). London, UK: Oliver and Boyd.

Galton, F. (1889). *Natural inheritance.* London, UK: Macmillan.

Galton, F. (1901). Biometry. *Biometrika*, 1, 1-4.

Howell, D. C. (2005). Jerzy Neyman. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1401-1402). West Sussex, UK: Wiley.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8, 3-7.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Mason, R. L., McKenzie Jr., J. D., &, Ruberg, S. J. (1990). A brief history of the American Statistical Association. *American Statistician*, 44, 68-73.

Miles, J. (2005). Central limit theory. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 249-255). West Sussex, UK: Wiley.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.

Nickerson, R. S. (2005). Neyman-Pearson Inference. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1402-1408). West Sussex, UK: Wiley.

Pearson, K. (1914). *The life, letters, and labours of Francis Galton: Birth 1822 to marriage 1853* (Vol. 1). London, UK: Cambridge University Press.

Pearson, K. (1924). *The life, letters, and labours of Francis Galton: Researches of middle life* (Vol. 2). London, UK: Cambridge University Press.

Pearson, K. (1930a). *The life, letters, and labours of Francis Galton: Correlation, personal identification and eugenics* (Vol. 3A). London, UK: Cambridge University Press.

Pearson, K. (1930b). *The life, letters, and labours of Francis Galton: Characterization, especially by letters, & index* (Vol. 3B). London, UK: Cambridge University Press.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the Twentieth Century.* New York, NY: Holt.

Spatz, C. (2016). *Exploring statistics: Tales of distributions* (11th ed.). Conway, AR: Outcrop.

Stanley, T. D., & Jarrell, S. B. (1989). Meta-regression analysis: a quantitative method of literature surveys. *Journal of Economic Surveys*, 3, 161-170.

Stevens, R. E. (1995). *Views from the hills*. Retrieved at http://shopperscientist.com/archive/views/28may95.html

Student (1908). The probable error of the mean. *Biometrika*, 6, 1-25.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician*, 70, 129-133.

West, B. H., Griesbach, E. N., Taylor, J. D., & Taylor, L. T. (1982). Numerals. *The Prentice-Hall encyclopedia of mathematics*. Englewood Cliffs, NJ: Prentice-Hall.

White, E. R. (1901, December). The plague of statistics. *Atlantic Monthly*, 842-844.

Wood, T. B., & Stratton, F. J. M. (1910). The interpretation of experimental results. *Journal of Agricultural Science*, 3, 417-440.

Yates, F. (1981). Fisher, R. A. In E. T. Williams & C. S. Nichols (Eds.). *Dictionary of national biography: 1961-1970* (pp. 361-362). Oxford, UK: Oxford.

Zabell, S. L. (2001). Neyman, Jerzy (1894-1981). In N. J. Smelser & P. B. Baltes (Eds.). *International encyclopedia of the social and behavioral sciences* (pp. 10645-10648). Oxford, UK: Elsevier.

Recommended reference form:
Spatz, C. (2016). A selected history of statistical practice. In W. D. Woody, R. L. Miller, & W. J. Wozniak (Eds.) *Psychological specialties in historical context: Enriching the classroom experience for teachers and students.* Retrieve from the Society for the Teaching of Psychology website: http://teachpsych.org/ebooks/psychspec