UNDERSTANDING STATISTICAL SIGNIFICANCE:
A CONCEPTUAL HISTORY

JOSEPH LITTLE
University of California, Santa Barbara

ABSTRACT

Few concepts in the social sciences have wielded more discriminatory power over the status of knowledge claims than that of statistical significance. Currently operationalized as $a = 0.05$, statistical significance frequently separates publishable from nonpublishable research, renewable from nonrenewable grants, and, in the eyes of many, experimental success from failure. If literacy is envisioned as a sort of competence in a set of social and intellectual practices, then scientific literacy must encompass the realization that this cardinal arbiter of social scientific knowledge was not born out of an immanent logic of mathematics but socially constructed and reconstructed in response to sociohistoric conditions.

INTRODUCTION

The National Center for Education Statistics, the primary federal entity for collecting and interpreting statistics on the condition of American education, has mandated that for all conclusions supported by hypothesis testing “[t]he level of significance (alpha) shall be .05” [1]. Accordingly, NCES publications, including Education Statistics Quarterly, “only discuss differences that are significant at the 95 percent confidence level or higher” [2, p. 1]. Such editorial policies are not anomalous; rather, they reflect the standard social scientific practice of defining statistical significance as $a = 0.05$. To take just one example from the research literature, Pennebaker, Barger, and Tiebout conducted an insightful study of the association between disclosure of past traumatic experiences and long-term health
among Holocaust survivors; in *Psychosomatic Medicine*, they report their cardinal finding:

> The primary prediction of the study was that degree of disclosure during the interview as measured by the trauma-SCL [skin conductance level] correlation would be associated with improved health in the months following the interview. The prediction was confirmed in that the more negative the trauma-SCL relationship, the less likely participants were to visit a physician for illness, \( r(31) = 0.34, \ p = 0.05 \ldots \) No other simple correlations between physician visits and interview-relevant variables attained significance [3, p. 582].

Had their correlation achieved \( p = 0.06 \), for example, Pennebaker, Barger, and Tiebout would have had nothing to report, for in the rhetoric of social scientific research “predictions” can only be legitimately “confirmed” (and then only provisionally) by results stamped with the status of statistical significance. In other words, “findings” at \( p > 0.05 \) aren’t findings at all.

What, then, is so special about 0.05? What theoretical or mathematical rationale warrants its privileged status over other \( p \) values, such as 0.045 or 0.078? Remarkably, the answer is: nothing [4; 5, p. 212]. Originally used in R. A. Fisher’s 1925 *Statistical Methods for Research Workers*, the level of 0.05 developed into the modern standard for social scientific research in response to regulatory pressures concerning what Bodewitz, Buurma, and de Vries call the “social management of trust” [6]. Ironically, this process of standardization, often carried out under the flag of Fisherian statistics, unfolded despite Fisher’s own forewarning: “[T]he calculation is absurdly academic,” writes Fisher in response to a hypothetical researcher who habitually chooses \( a = 0.01 \), “for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas” [7, p. 42]. Little did Fisher realize how prophetic his anecdote would prove in the light of our 21st century methodological practices.

Ian Hacking argues that fundamental to the work of social construction is the dismantling of the assumption of inevitability. Social constructionists, according to Hacking, begin by demonstrating that the subject of their study “need not have existed, or need not be at all as it is” and frequently urge, ultimately, that its present state is actually “quite bad” and in need of a radical transformation if not complete eradication [8, p. 6]. My article accepts the modest challenge of demonstrating that the current standard for statistical significance “need not be at all as it is.” It is within this light that I sketch a conceptual history of statistical significance from its 1710 emergence in the work of John Arbuthnot to its early 20th century incarnation in the work of R. A. Fisher. If we provisionally accept James Boyd White’s description of literacy as “the ability to participate fully in a set of social and intellectual practices” [in 9, p. 271], then scientific literacy must entail at least
the conceptual realization that the cardinal arbiter of social scientific knowledge, statistical significance, was not born out of an immanent logic of mathematics but repeatedly reconstructed in the social sphere. Before engaging that history, however, a brief overview seems warranted.

**AN OVERVIEW OF TESTS OF SIGNIFICANCE**

Generally speaking, a test of significance is a largely mathematical procedure used to determine whether the difference between two values is sufficient to eliminate the suggestion that the difference is due to chance; in other words, the test is used to determine whether an experimental difference between values is statistically significant. In its modern form, the test has five conceptual steps [10, p. 318]:

1. State the null and alternate hypotheses (H₀ and H₁, respectively).
2. Choose the test statistic (T).
3. Choose the level of significance (α) and determine the corresponding rejection region (R).
4. Collect data and calculate the value of T.
5. Decide whether the difference is significant:
   a. If T is within R, then reject H₀ in favor of H₁ (significant).
   b. If T is not within R, then do not reject H₀ (not significant).

To bring this conceptual scheme to life, consider the following example: Suppose we conduct a journal writing study in which we draw a random sample of 50 sophomores from a population of sophomores. We then randomly assign half of these sophomores to an experimental “journal writing” group and half to a control group. Over the quarter, the sophomores in the experimental group spend the last 10 minutes of each class period writing about their thoughts and feelings pertaining to the course material while those in the control group spend this time in more traditional small group discussions. At the end of the quarter, we administer a common final examination and find that the mean score for those in the experimental group is 69 whereas the mean score for those in the control group is 65.

The pivotal question is whether the difference in mean final examination scores is attributable to the experimental treatment or whether it is simply the result of chance. Within the context of a test of significance, our null hypothesis (H₀) is that the two mean scores are equal; it is an algorithmic step: we categorically assume, until convinced otherwise, that the journal writing activity had no systematic effect on final examination scores. Experimentally, however, we know that the two mean scores (69 and 65) do differ. The pivotal question then becomes one of formal probability: Assuming the null hypothesis is true, what is the probability of obtaining a sample mean of 69 or greater from the population of sophomores? If the probability is sufficiently unlikely under the null hypothesis, then we formally
reject the null hypothesis in favor of our alternate hypothesis; otherwise, we have little basis for such a rejection because we have failed to displace chance as a likely explanation for our experimental results.

The role of judgment in the test of significance is necessitated by Step 3, for here the researcher must choose the level of significance \( (\alpha) \) upon which to judge the experimental results. Essentially, the chosen \( \alpha \)-level is the quantified definition of what the researcher means by “sufficiently unlikely.” Within the context of our journal writing experiment, for example, an \( \alpha \)-level of 0.05 would mean that we would reject the null hypothesis in favor of our alternate hypothesis if the test of significance determined that we would obtain a sample score of 69 under the null hypothesis no more than five times per 100 attempts. Such a definition, however, cannot be ascertained rigorously from mathematics: Although tests of significance can compute the probability of obtaining a particular sample mean, they cannot establish the \( \alpha \)-level against which that probability will be judged, for the choice of an \( \alpha \)-level requires the consideration of two risks: the risk of mistakenly rejecting a true null hypothesis (a Type I error) and the risk of mistakenly not rejecting a false one (a Type II error). As Richard Shavelson reminds us, that choice is a deeply consequential one:

The problem, then, is to define operationally what is meant by unlikely. By convention, an unlikely sample result . . . is defined as one whose probability of occurrence is less than or equal to a fixed small quantity \( [\alpha] \): .05 or .01. This convention grew out of experimental settings in which the error of rejecting a true \( H_0 \) was very serious. For example, in medical research, the null hypothesis might be that a particular drug produces undesirable effects. Deciding that the medicine is safe (i.e., rejecting \( H_0 \)) can have serious consequences. Hence, conservatism is desired. Often in behavioral research, however, the consequences are not so dire. For example, suppose an instructor is to select one of two texts of approximately equal cost. The instructor might conduct a study to determine which book leads to greater achievement, satisfaction, or both. In this case, the instructor might be willing to reject the null hypothesis on the basis of a probability \( [\alpha] \) less than or equal to .25. If, in fact, there is no true difference between the texts, the “error” of selecting one over the other is not particularly serious. . . . Some wisdom, then, should be exercised in setting the level of significance [11, p. 263].

**A CONCEPTUAL HISTORY OF STATISTICAL SIGNIFICANCE**

The origin of statistical significance can be traced back to the English physician and writer John Arbuthnot. In 1710, Arbuthnot presented 82 consecutive years of male-dominated birth statistics for London as evidence for the claim that divine providence rather than chance governed the sex ratio of newborn children [12, p. 168; 13, pp. 225-226]. The probability of chance accounting for more boys than girls being born each year during this period was \((1/2)^{82}\).
According to Arbuthnot, which he presented as a probability of 1 in 483,600,000,000,000,000,000; couched in near certainty, Arbuthnot comfortably rejected the null hypothesis and concluded that “it is Art, not Chance, that governs” [in 13, p. 226].

Not stating, or even considering, a formal level of significance was common practice in Arbuthnot’s time, for in such an overwhelming probability as 1 in 483,600,000,000,000,000,000,000, significance was thought to be directly observable. At the time, the known universe consisted of only about 3,000 stars [13, p. 77]. Who, then, could argue against odds that literally surpassed the astronomical by nearly two dozen orders of magnitude?

Statisticians would have to wait just over a century for Friedrich Wilhelm Bessel’s 1815 introduction of “Probable error,” which provided the first continuum for judging significance, and another eight decades for Francis Edgeworth’s introduction of the term “significant” in a statistical context. In the meantime, statistics proceeded primarily through implicit understandings of what was “real” and “apparent,” and in this earliest and most conservative epoch of statistical knowledge making, that was accomplished primarily through the reporting of sheer and overwhelming probabilities.

In the century and a half that followed, a variety of levels of significance were informally adopted on individual bases, and these began slowly to supplant the presentation of purely overwhelming probabilities. French luminaries such as Jacob Bernoulli and Pierre Simon Laplace applied to their work the level of “moral certainty,” which they operationalized as a probability of 1 in 1,000 [13, p. 78]. Accordingly, in his 1781 comparison of the number of male births in Paris to that in Naples, Laplace rejected the probability of 1 in 100 that the difference owed to chance, claiming that the probability “is not sufficiently extreme for an irrevocable pronouncement,” whereas Simeon Poisson accepted the probability of 1 in 213 as “quite close to certainty” in his 1837 study of criminal verdicts [13, pp. 134-135, 190]. The conservative Joseph Fourier, on the other hand, chose 1 in 20,000 as his personal level of significance [14, p. 98].

Although undeniably extreme relative to modern standards, the conservatism exhibited by these early statisticians is understandable given the professional climate of their day. Statistics had yet to achieve a professional ethos. With neither the backing of disciplinary status nor the luxury of socially agreed-upon methods, statisticians appealed to impenetrable probabilities in their attempt to gain assent. As members of a fledgling community, they simply could not afford to commit Type I errors; missing an important statistical finding was far preferred to publicizing a mistaken one.

Although intersubjective variation in levels of significance continued well into the 20th century, the latter half of the 19th century saw the first semblance of a standard scale for statistical significance in the concept of “probable error.” Introduced in 1815 by Friedrich Wilhelm Bessel, probable error is the distance from the mean to the quartile in a normal distribution: Normally distributed errors,
then, are equally likely to be within one probable error of the mean as not [13, p. 230]. In such a seemingly sensible unit of measure, visualizable on the bell curve, statisticians found an improved way to compare and present their results, and such key figures as Adolphe Quetelet, Francis Galton, Hermann Ebbinghaus, G. Udny Yule, and Major Greenwood adopted it. Soon, reports of differences between means were accompanied by estimates of their probable error in the form difference ± probable error. In his 1904 study of the “human viscera,” for example, Greenwood reports his findings on the correlation of heart weight with age as “.1363 ± .0250,” which allows the reader to assess the relative strength of the correlation, 0.1363, in light of its probable error, 0.0250 [15, p. 68].

Yet, the question of how large a difference must be in relation to its probable error still remained. Ebbinghaus, for example, declared that a difference between means of six times the probable error was “fully proved” while a difference of twice the probable error was at least noteworthy [13, pp. 260-261]. Likewise, Greenwood interpreted a difference of five times the probable error as “distinctly sensible” [15, p. 68], whereas John Venn, Student, and D. Caradog Jones all independently insisted that a difference of three times the probable error was significant for most purposes [16, p. 148; 17, p. 13; 18, p. 160].

Others chose more idiosyncratic means of judgment: Edgeworth adopted twice the “modulus” as his level of significance, corresponding to 4.2 times the probable error, and accepted differences of 1.5 times the modulus as noteworthy [13, p. 311; 14, p. 266]. Wilhelm Lexis, on the other hand, alternated between probable error and an esoteric combinatorial result corresponding to approximately 6.3 times the probable error when interpreting his results [13, p. 229; 14, p. 244]. Whatever the demarcation in terms of probable error, the “moral certainty” guaranteed by a probability of 1 in 1,000 as well as Fourier’s level of 1 in 20,000 receded from use, and the era of excessive conservatism passed.

Despite Edgeworth’s and Lexis’s notable departures, probable error had become the standard by 1902, and readers of the British journal *Biometrika* complained with sufficient force about the lack of probable error reporting that its editors responded with a primer, “On the Probable Errors of Frequency Constants”; here, they lament, “Half the blunders made in superficial statistical investigations arise from neglecting the values of the probable errors of the results obtained” [19, pp. 273-281]. Still, the question of how large a difference must be in relation to its probable error failed to attract serious attention. Karl Pearson’s 1906 “Note on the Significant or Non-significant character of a Sub-sample drawn from a Sample” exemplifies this ambivalence, for nowhere in his three-page paper on statistical significance does he concern himself with an actual level of significance; instead, Pearson represents the threshold abstractly through the variable $c$.

---

1The probable error of a normal distribution is equal to 0.6745 times the standard deviation.
Now let it be reasonable to suppose a quantity significant when it is $\beta$ times its standard deviation, or $\beta \cdot 0.67449$ times its probable error. . . [20, p. 183].

Clearly, the choice of a suitable threshold, ranging anywhere from three to six times the probable error, was to be left up to the judgment of the individual researcher and based upon the particular assets and liabilities of the research problem in question.²

Strangely, talk of probable error was short lived: By the late 1910s, the slow migration toward a term Pearson coined in 1893 had gained sufficient momentum to effect a transformation of sorts: Statisticians began to conceptualize a difference between experimental values not in terms of its probable error but in relation to its “standard deviation,” and published accounts began to reflect that reconception [13, p. 328]. In 1902, the editors of Biometrika had essentially pronounced standard deviation dead:

Unfortunately custom has not taken this standard deviation as the measure of the goodness sample, but the whole theory having ultimately developed from the normal curve, the probable error instead of the standard deviation has been chosen, i.e., $0.67449 \times$ standard deviation [19, p. 273].

Yet, by the late 1910s, standard deviation occupied the focal position in the analysis and reporting of statistical data. In their 1911 comparison of death rates in Birmingham to that in Liverpool, for example, Pearson and J. F. Tocher worked to “discover whether there is actual differentiation between these two systems of deaths” by dividing the difference between means by the standard deviation of the difference, yielding 12.8648; “We conclude,” write Pearson and Tocher, “that the chance of such a difference arising from random sampling is enormous, ³ or the two deathrates are most certainly and markedly different” [21, p. 168]. Likewise, in his 1911 textbook, G. Udny Yule describes an observed difference that “greatly exceeds” three times the standard deviation as significant [22, p. 262], while Student’s 1923 work on varieties of cereals also evaluates differences between means in relation to their standard deviations [23].

In the 1910s, standard deviation was not a new term; in fact, it was an intermediate value in the calculation of probable error. However, its acceptance as the new scale for significance inaugurated a transformation of sorts, rather than a trivial transition, because with the move to standard deviation came increased conservatism in interpreting results: Although statisticians moved from probable error to standard deviation, few modified their personally adopted levels of

²Venn’s 1888 lamentation provides the only exception to the relative ambivalence among statisticians to actual levels of significance. “When we are dealing with statistics,” writes Venn, “we ought to be able not merely to say vaguely that the difference does or does not seem significant to us, but we ought to have some test as to what difference would be significant” [16, p. 147].

³Given the context of the findings, Pearson and Tocher must have used “enormous” here to mean infinitesimally small, or perhaps they simply misspoke.
significance; those who had considered a difference of three or four times the probable error as significant now considered a difference of three or four times the standard deviation as significant. Because one standard deviation is 48 percent larger than one probable error, the migration to standard deviation increased the stringency of what formally constituted statistical significance. Consequently, the likelihood of committing Type II errors increased proportionally.

With the move to standard deviation also came the custom of reporting decimal probabilities in conjunction with results, perhaps owing to the published and widely used tables of Pearson and Student, and, although decimal probabilities were based on multiples of standard deviations, they eventually replaced the latter in the context of reporting results. Pearson’s 1914 study of Trypanosome strains illustrates the style of reporting that was becoming commonplace at the time:

I questioned first whether the strains found in the two Hartebeeste were the same; they give $\chi^2 = 108.69$, and therefore $P < .000,000,1$. In other words not once in 10,000,000 trials would two such divergent samples arise if the Hartebeeste strains were samples of the same population. I now compare the Waterbuck and the Oribi; these provide $\chi^2 = 109.25$ and $P < .000,000,1$, and again the extraordinary divergence, not the sameness, is the statistical feature [24, p. 93].

Likewise, R. A. Fisher’s immensely popular 1925 textbook, *Statistical Methods for Research Workers*, further advocated the use of decimal probabilities as the appropriate index of significance. Appealing to an expanded audience of biological researchers with no assumed training in statistical theory, Fisher explains:

Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials. . . . The value for which $P = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as the limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant [25, p. 47].

Thus, the historical origin of what would become the standard level of significance in social scientific research was born [4, 10, 26].

Yet, as I mentioned in the introduction, Fisher was much more sensitive to the necessity of judgment than this formal definition conveys. In his 1926 paper on the design of field experiments, for example, he unmistakably acknowledges the role of individual judgment in choosing a suitable $\alpha$-level:

If one in twenty [$\alpha=0.05$] does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance [27, p. 504].
Unfortunately, only the most mechanical of Fisher’s work and the now-coveted 0.05 seem to have survived the last 75 years of standardization.

CONCLUSION

If we envision literacy as a sort of competence in a set of social and intellectual practices, as White suggests [in 9, p. 271], then that competence must include a critical and reflective awareness of the social and intellectual practices at hand. Accordingly, scientific literacy must encompass the realization that the practice of legitimating statistical findings by way of statistical significance, currently standardized at \( \alpha = 0.05 \), is not the result of an immanent logic of mathematics but a socially constructed response to a variety of post-theoretical conditions. By sketching the sociohistoric depth of the concept of statistical significance from its 18th century origins in overwhelming probabilities to its 20th century incarnation as twice the standard deviation, I attempt here a move toward such a literacy. For to the extent that we defer to the tradition of statistics textbooks in the education of our students, that goal will remain largely unmet.

REFERENCES


**Other Articles On Communication By This Author**


Direct reprint requests to:

Joseph Little
133 Deforest Road
Toronto, Ontario
Canada M6S 1J7