J Clin Epidemiol Vol. 51, No. 4, pp. 355–360, 1998

0895-4356/98/$19.00
PII S0895-4356(97)00295-3

EDITORIAL RUMINATION

P-Values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin

Alvan R. Feinstein[*]
Yale University School of Medicine, New Haven, Connecticut

ABSTRACT. For both P-values and confidence intervals, an α level is chosen to set limits of acceptable probability for the role of chance in the observed distinctions. The level of α is used either for direct comparison with a single P-value, or for determining the extent of a confidence interval. "Statistical significance" is proclaimed if the calculations yield a P-value that is below α, or a $1 - \alpha$ confidence interval whose range excludes the null result of "no difference." Both the P-value and confidence-interval methods are essentially reciprocal, since they use the same principles of probabilistic calculation; and both can yield distorted or misleading results if the data do not adequately conform to the underlying mathematical requirements.

The major scientific disadvantage of both methods is that their "significance" is merely an inference derived from principles of mathematical probability, not an evaluation of substantive importance for the "big" or "small" magnitude of the observed distinction. The latter evaluation has not received adequate attention during the emphasis on probabilistic decisions; and careful principles have not been developed either for the substantive reasoning or for setting appropriate boundaries for "big" or "small." After a century of "significance" inferred exclusively from probabilities, a basic scientific challenge is to develop methods for deciding what is substantively impressive or trivial. J CLIN EPIDEMIOL 51;4:355–360, 1998. © 1998 Elsevier Science Inc.

INTRODUCTION

During the first quarter of the 20th century, the word "significance" was calmly leading its original lexicographic existence as an ordinary term of human communication, having such meanings as important, impressive, or consequential. In 1925, however, R. A. Fisher, in the first edition of an extraordinarily influential book [1], gave a new role to the word "significance." He was discussing issues in numerical stability that had plagued mathematicians and researchers for about a century ever since the "law of large numbers" was proposed by Simeon-Denis Poisson [2] in 1837.

THE "LIMIT OF OSCILLATION"

Poisson had pointed out that the numerical result of a small number of observations was often unreliable. Using ideas in the "calculus of probabilities" that had been developed during the 18th century by the Bernouilli family and others [3], Poisson proposed a limit of oscillation for regarding an observed result as numerically trustworthy or reliable. In 1840, Jules Gavarret [4] extended Poisson's concept, which pertained to a single group of data, and applied it to a comparison of results for two groups.

In modern notation and symbols, the Poisson-Gavarret formula states that the limit of oscillation for the difference in two proportions, $p_1 = r_1/n_1$ and $p_2 = r_2/n_2$, is

$$2.828 \sqrt{(p_1\, q_1/n_1) + (p_2\, q_2/n_2)},$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. For example, if $p_1 = .33$ and $p_2 = .50$, the difference is $p_2 - p_1 = .17$. If $n_1 = 90$ and $n_2 = 80$, the "limit of oscillation" is

$$2.828 \sqrt{[(.33)\ (.67)/90] + [(.50)\ (.50)/80]} = .21,$$

which is larger than the difference itself. If $n_1 = 900$ and $n_2 = 800$ for the same proportions, however, the corresponding limit of oscillation becomes .067, which is much smaller than the observed difference. The "true" value of the increment in the second case would lie between $.17 \pm .067$ and could readily be accepted as reliable. In the first instance, however, the range of "oscillation" would be the unacceptably large zone of $.17 \pm .21$.

Many readers will recognize that the Poisson-Gavarret procedure constructs the exact counterpart of what is today called a confidence interval. The term inside the square root sign is the standard error of the difference in two proportions and the value 2.828 determines the level of confidence. In modern approaches that use Gaussian values of Z_α to get confidence levels of $1 - \alpha$, the value of Z_α is 2.828 for $\alpha \cong .0047$. Thus, the Poisson-Gavarret formula produces essentially a .995 or a 99.5% confidence interval. In an ex-

[*]Address correspondence to: Alvan R. Feinstein, M.D., Yale University School of Medicine, 333 Cedar Street–Room 1456 SHM, New Haven, CT 06510.

Accepted for publication on 17 November 1997.

cellent modern explanation for the choice and role of the level of α, Warren Weaver [5] has called it an "unconfidence coefficient," because "it measures the percent of time (the statistician) expects to be wrong rather than right . . . (in the) confidence we dare have in statements" such as a $1 - \alpha$ confidence interval.

Later in the 19th century, as investigators discovered the major difficulty of getting groups large enough to satisfy the Poisson-Gavarret demand for $\alpha = .0047$, other mathematicians [3] proposed "looser" standards, lowering the level of α in some instances to .17, thus making the limit of oscillation an 83% confidence interval.

HYPOTHESIS TESTING AND P-VALUES

The limits-of-oscillation approach was essentially abandoned, however, after R.A. Fisher proposed the strategy that is now called hypothesis testing. Fisher's approach begins by establishing a statistical null hypothesis that the two compared groups are essentially equivalent. The observed difference, d_0, in the two groups, and the calculated SED, which is the standard error of the difference, are then used to construct a "critical ratio," $Z_0 = d_0/SED$.

When referred to the appropriate associated statistical distribution, the value of the critical ratio corresponds to a probability value, called P. It denotes the chance, under the null hypothesis, of finding a difference at least as large as the observed value of d_0. If the P-value is below the boundary of α, the null hypothesis is rejected, and the observed result is called "significant." Fisher deemed the procedure a "test of significance" and he set the level of α at .05 because it was a "convenient . . . limit in judging whether a deviation is to be considered significant or not" [6].

Fisher's "convenient" boundary of .05 for α became well known thereafter when it was adopted by regulatory agencies (as well as journal editors) seeking a criterion to judge "significance." After completing a research project, and hoping to get approval by regulators, reviewers, or editors, the investigators would be delighted if the calculations showed P = .04, and distressed if P = .06. In many medical journals and other enclaves of statistical approbation, new drugs would be licensed, grants approved, and manuscripts accepted for publication, all according to whether the *p*-values were below or above the crucial boundary of .05.

Advantages and Disadvantages of $\alpha = .05$

One main virtue of P-values was that an exact probability for the observed result could avoid previous problems when the "limits of oscillation" would vary according to the choice of a level for α (or its 19th century counterpart). Until the modern advent of ubiquitous electronic computation, however, P-values were seldom exact, and had to be approximated from special tables that showed corresponding values of Z for such α levels as .2, .1, .05, .01, etc. Consequently, intermediate values of Z_0 were usually converted to imprecise expressions such as $P > .2$ or $P < .05$. Today, however, the computer printout for the appropriate test will usually produce exact values, such as P = .073.

The main problem, however, was using $\alpha = .05$ as a rigid boundary to separate "significance" from "non-significance." Although facilitated by P-values, this rigidity is not their fault, being produced by a steadfast adherence to Fisher's original proposal, despite his own later advocacy of changes and flexible usage. Fisher later declared [7] that "no scientific worker has a fixed level of significance at which from year to year, in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in light of his evidence and his ideas."

Renascence of Confidence Intervals

During the past decade, the "limits of oscillation" have returned to medical research as the confidence intervals that have been advocated for estimation rather than hypothesis testing. The idea of estimation is best illustrated by political poll-taking, in which the "parameter" of a parent population (e.g., opinions of all voters) is estimated from the results found in a smaller sample (e.g., the group used in a poll).

With the estimation strategy, "statistical significance" can be declared at the level of α, if the null hypothesis value of 0 (or 1, for a ratio) is excluded from the $1 - \alpha$ confidence interval. Confidence intervals, however, have a major additional advantage beyond their role in this single decision. They can be used to estimate "the size of difference of a measured outcome between groups, rather than a simple indication of whether or not it is statistically significant" [8]. If the P-value is not "statistically significant," the estimate of how large the difference might have been is particularly useful in helping avoid possibly erroneous conclusions that the two groups have similar results.

In the lively dispute between prominent statisticians on both sides of the controversy, the proponents of confidence intervals want to supplement or replace the single value of P, which does not give the scope of information offered by a confidence interval. The defenders want the P-value preserved because it indicates the important distance from the stochastic boundary of α. Confidence intervals seem to be winning the battle in prominent medical journals, where the editors, after presenting "tutorial" discussions to educate readers, proudly announce their enlightened new policies demanding confidence intervals instead of (or in addition to) P-values. The enlightened new policies are presumably expected to bring medical literature out of the statistical "dark ages" induced by previous editorial policies that set the demand for P-values.

Statistical Similarities of the Two Methods

An even-handed reaction to this controversy would be to agree with both sides, since both sets of contentions are

correct and both sets of proposals have merit. In fact, an excellent policy would ask for both types of information: confidence intervals to show the potential range of results, and P-values to keep authors and readers from being misled by an arbitrarily chosen interval.

Regardless of the opposing viewpoints, however, the argument itself resembles a debate about whether to give an unsatisfactory medication orally or parenterally. The argument ignores the reciprocal similarity of P-values and confidence intervals in decisions about "statistical significance" while also overlooking several prominent problems in the statistical procedures themselves and in the distraction, by both P-values and confidence intervals, from the scientific challenge of setting quantitative standards for descriptive boundaries. The rest of this section is devoted to the similarity of the two procedures. The next two sections are concerned with the problems and challenges. To separate the two types of "statistical significance," I shall use quantitative for the descriptive boundaries of magnitude, and stochastic for the mathematical issues in inferential probabilities.

When α is used as a stochastic boundary, the reciprocal similarity of P-values and confidence intervals can easily be shown from the construction and "location" of the concomitant stochastic hypothesis. In ordinary stochastic contrasts of two groups, the "location" of the tested hypothesis can be cited as Δ. For the conventional null hypothesis, symbolized as H_0, the value assigned to Δ is 0. Thus, in an incremental contrast of two observed means, $\overline{X}_A$ and $\overline{X}_B$, with parametric means, μ_A and μ_B, the hypothesis is cited as H_0: $\Delta = \mu_A - \mu_B = 0$.

In the conventional two-group Z test, after a boundary is chosen for α, with a corresponding value of Z_α, the data analyst constructs what might be called a "decision interval" around the null hypothesis value of 0. If d_o is the observed difference for $\overline{X}_A - \overline{X}_B$, and if SED is its standard error, the critical ratio is calculated as $Z_0 = d_0/(\text{SED})$. The null hypothesis is conceded if Z_0 falls inside the interval of $0 \pm Z_\alpha$, and is rejected if Z_0 is outside. Thus, a two-tailed P-value will achieve stochastic (or "statistical") significance if $|Z_0| > Z_\alpha$.

For example, suppose group 1 has 41 members, with mean 17.3 and standard deviation 5.8. The corresponding values in group 2 are 40, 12.9, and 6.3. The value of d_0 is $17.3 - 12.9 = 4.4$. With suitable calculations not shown here, the pooled standard deviation in the two groups will be 6.05; and the standard error of the difference in means (SED) will be 1.34. The value of d_0/SED will be $4.4/1.34 = 3.28$, for which $2P = .0011$.

A two-tailed confidence interval for comparing two groups is constructed not around $\Delta = 0$, but around the observed value of d_0. After α and an appropriate Z_α are selected, the interval is calculated as $d_0 \pm Z_\alpha(\text{SED})$. The null hypothesis is conceded if its value of 0 falls inside this interval and is rejected if 0 is outside. Thus, the stochastic requirement for "significance" is $|d_0| - Z_\alpha(\text{SED}) > 0$. When both sides of the latter expression are divided by SED to express the result in standard error units, the symbolic formulation becomes $|d_0/\text{SED}| - Z_\alpha > 0$, which is $|Z_0| - Z_\alpha > 0$, or $|Z_0| > Z_\alpha$.

In the foregoing example, a 95% confidence interval would be calculated, with $Z_\alpha = 1.96$, as $4.4 \pm [(1.96)(1.34)] = 4.4 \pm 2.63$. The result would exclude 0, and would be stochastically significant at $P < .05$.

Because the requirement for stochastic significance is exactly the same whether Z_α is converted to a confidence interval or to a P-value, confidence intervals are essentially a type of "reciprocal" for P-values. For the P-value, $|d_0|/\text{SED} = Z_0$ is calculated and compared directly with Z_α. For the confidence interval, $|d_0| - Z_\alpha(\text{SED})$ is calculated and compared against 0. If the goal is a stochastic determination of "statistical significance," the two procedures give identical results.

STATISTICAL PROBLEMS

Despite the advantage of demonstrating how large $|d_0|$ might really be, the confidence-interval approach can produce a new problem while preserving two old ones, as discussed in the sections that follow.

Arbitrary or Misleading Results

When the boundary of .05 for α is eliminated as rigid standard for "significance," a data analyst can make arbitrary or idiosyncratic choices of whatever standard may be desired. If the confidence interval is cited for a specific level (such as 95%), the rigidity of $\alpha = .05$ is retained, but if no level is mentioned for α, the standard becomes a "dealer's choice," enabling investigators to do whatever they would like, and to focus on whatever end of the confidence interval is attractive.

To advocate a "big" difference, an investigator can focus on the upper end of the confidence interval, ignoring the lower end, which might extend below the null hypothesis level. Although the observed distinction would not be stochastically significant, the investigator may nevertheless claim support for the "big" difference. This type of abuse evoked Fleiss's recommentation [9] that P-values always be reported to prevent the problem. If both ends of the confidence interval are always cited, however, readers of the published results can readily discern whether the lower end goes beneath the null-hypothesis boundary.

A more substantial problem, against which the reader has no overt protection, occurs if a fixed level of α is no longer demanded for a $1 - \alpha$ confidence interval, and if non-specific "confidence intervals" are formed arbitrarily. A zealous analyst can choose whatever level of α is needed to make the result as large or small as desired. To illustrate this point, suppose the observed difference is $d_0 = 8.0$, with a standard error of SED = 4.0. To make the confidence interval exclude 0, a relatively large α, such as .1, will have $Z_{.1} = 1.645$, and the confidence interval component will be $(1.645 \times$

4.0) = 6.58. The two-tailed 90% interval, constructed as 8.0 ± 6.58, will extend from 1.42 to 14.58, thereby excluding 0. To make the interval include 0, the choice of a smaller α, such as .02 with $Z_{.02} = 2.32$, will produce the component of $2.32 \times 4 = 9.28$. Constructed as 8.0 ± 9.28, the two-tailed 98% confidence interval will include 0 by extending from −1.28 to 17.28. To be given an unimpressive upper end, the 50% confidence interval constructed with $\alpha = .5$ and $Z_{.5} = .674$, will extend only to 8.0 ± 2.7 = 10.7. To receive a much higher upper end, the 99.9% confidence interval, with $\alpha = .001$ and $Z_{.001} = 3.29$, will extend to 8.0 + 13.16, and will exceed 21.

As a customary standard for stochastic decisions, $\alpha = .05$ may have all the invidious features of any arbitrary choice, but it has the laudable virtue of being a generally accepted criterion that cannot be altered whenever, and in whatever direction, the data analyst would like. Unless the proponents of confidence intervals agree on using a fixed (although suitably flexible) boundary for α, stochastic decisions can become a subjective process, based on personal *ad hoc* goals for each occasion.

Until a consensus is established on this matter, readers should be wary of anything reported merely as a "confidence interval." If not accompanied by an indication of $1 - \alpha$ (such as 95%, 90%, 50%, etc.), the unspecified result may be misleading. If the reported $1 - \alpha$ level departs from the customary 95%, a suitably convincing justification should be provided.

Uninformed Readers

The two main elements used for calculating most P-values or confidence intervals are d_0 for the difference in two groups, and SED for the standard error of the difference. The value of d_0 is usually evident in the published results, but SED may be omitted or obscured when stochastic calculations are cited only for a P-value or for a confidence interval. The problem of the unknown SED is particularly pertinent when an approximate rather than exact value is listed for P (such as <.04 rather than .036) or when the $1 - \alpha$ level is not cited for a confidence interval listed merely as "1.07 to 8.63." In both of the latter two situations, the reader has no way to discern the crucial value of SED. For example, the ratio of SED/d_0 for the two sets of data would be a counterpart of the coefficient of variation calculated for a single set of data as $s/\overline{X}$, where s is the standard deviation and $\overline{X}$ is the mean. A suitably small value for the SED/d_0 ratio would indicate the relative stability of a "tight" range of possible variation for d_0, and would not depend on the choice of Z_α. Since the inverse value, d_0/SED, is regularly converted to P-values and is usually"statistically significant" when ≥2, a value of ≤.5 for SED/d_0 could readily denote a "stable" zone of variation for d_0.

If given the information about SED, readers can make their own decisions or do additional calculations, without having to depend on arbitrary choices of stochastic boundaries. Editors who require confidence intervals or P-values (or both) in statistical reports can offer readers a useful service if the values of SED were always demanded in addition or instead.

Inadequate Calculations

A separate problem in both P-values and confidence intervals is that they may not always be calculated correctly. In the usual formulas offered for the calculations, the confidence interval, expressed with "±" in the formula $d_0 \pm Z_\alpha$ (SED), is placed symmetrically around the value of d_0.

This symmetry may not occur, however, in dimensional data that have an eccentric (non-Gaussian) distribution or in binary data where the summary proportion (such as .12) departs substantially above or below the "meridian" value of .50. The symmetry is also unlikely for groups with relatively small numbers of members. Consequently, symmetric calculations with Z_α coefficients from Gaussian distributions may be inappropriate for various individual situations, although often correct on average.

These problems can readily be avoided with modern computer-intensive statistics, where a re-sampling procedure or a random-permutation arrangement (such as the Fisher exact test) can produce the exactly correct values for P. Re-sampling procedures can also yield confidence intervals [10] more accurate than those obtained with the traditional formulas.

Scientific Defects

The greatest problem of all, however, is that the "coin" of P-values and confidence intervals is scientifically unsatisfactory, regardless of how the two sides of the coin are constructed. They focus only on the "statistical significance" and potential ranges determined by stochastic issues in probability, and they give no attention to the significance of the quantitative magnitude of an observed distinction.

The scientific implication of the observed distinction is regularly cited with terms such as impressive difference, biologic importance, clinical significance, or medical pertinence. The idea has seldom been given the dignity of its own name, however, and boundaries for this type of substantive "significance" are almost never mentioned when stochastic boundaries are emphasized for decisions based on P-values and confidence intervals.

The statistical index of contrast used for comparing two groups of data has also not received a standard name or distinctive identity, such as "index of contrast." Throughout this discussion, the index has been the difference in results, but they can also be compared as diverse ratios, proportionate increments, numbers needed to be treated [11], or the standardized increment called "effect size" [12]. Regardless of which index of contrast is used, however, it will

have distinctive boundaries of "big" and "small" that must be considered for at least three important statistical decisions:

1. The advance calculation of sample size for a randomized trial or other comparative research requires a previous designation of δ, which usually represents the magnitude of a "big," "important," or "impressive" difference in the compared groups.
2. The customary test of "statistical significance" is usually done when an observed difference, such as d_0, seems large enough to exceed a "big" value such as δ, which may or may not have been stipulated before the research began.
3. If the test yields a nonsignificant result, i.e., $P > \alpha$, the upper border of the confidence interval will show how large d_0 might have been, but a decision that the upper value is really large, requires comparison against the magnitude of the δ set for "big."

(An additional demand for demarcating "big" occurs when "loss functions" or "utility values" are established in statistical procedures that use Bayesian inference [13] or quantitative decision analysis [14]).

Although the choice of δ is required to decide whether results have quantitative significance, this decision has hardly been considered during all the probabilistic deliberations about the stochastic components of "statistical" significance. Another descriptive boundary value that has received even less attention than δ is ζ, which is the upper magnitude of a tiny, trivial, or "insignificant" value of d_0. The value of ζ has also been neglected despite increased attention, in the past two decades, to questions of equivalence for therapeutic agents.

Neglecting the boundaries of "big" and "small," which are part of descriptive rather than inferential statistics, has led to major scientific and statistical difficulties. A trivial and unimportant difference may be called "significant" because it was brought across the α boundary by a huge sample size. Conversely, a difference of major magnitude and importance may be dismissed because the sample size was too small to let the results pass the α test. Furthermore, in the absence of a separate boundary for ζ, decisions about equivalence may depend on the fallacious idea that "small" is represented by anything smaller than the "big" value of δ.

The fundamental flaw of P-values and confidence intervals, therefore, is not what they do, but what they omit. They offer no guidance for the basic quantitative scientific appraisals that depend on purely descriptive rather than inferential boundaries. A descriptive scientific question cannot be suitably answered when an inferential choice of Z_α multiples an inferential estimate of SED to form an arbitrary mathematical entity, $Z_\alpha \times (SED)$. Aside from obvious problems when the Gaussian Z_α and SED do not adequately represent "eccentric" or asymmetric data, the doubly inferential product gives no attention to the investigator's descriptive concepts and goals.

As crucial scientific boundaries for quantitatively significant and insignificant distinctions respectively, δ and ζ must be chosen with criteria of magnitudes in scientific comparison, not in mathematical inference. The stochastic intervals constructed with Z_α and SED calculations do not address the fundamental substantive issues, and may disguise or distort the absence of truly scientific standards for the quantitative boundaries.

The methods that can be used for choosing and establishing boundaries are beyond the scope of this essay, but will involve judgments about both an appropriate index of contrast and appropriate demarcations of "large" and "small" for that index. In one approach to the problem, Burnand *et al.* [15] tabulated the boundaries that seemed to be used for quantitative significance when results in general medical publications had stochastic significance. The boundaries were $\geq$1.2 for a ratio of two means, $\geq$0.28 for a standardized increment, $\geq$.32 for a correlation coefficient, and $\geq$2.2 for an odds ratio of two proportions. The standardized increment, proposed as an index of "effect size" in psychosocial research, has been demarcated into zones of graded magnitude by Guilford [16] and by Cohen [12]; but the index is seldom used in biomedical research, and does not seem suitable for comparing the rates and proportions studied in clinical epidemiology. The odds ratio, although often applied for the latter comparisons, is much more difficult to understand than the simpler index called the number needed to treat [11].

Regardless of whatever indexes are chosen, however, the quantitative boundaries cannot be rigidly fixed because they will vary in diverse situations according to the type of measuring system, the biologic implications of the measurement, and sometimes the populational impact of the results. Nevertheless, such standards can readily be established in each research situation if thoughtful investigators recognize the fundamental need.

Developing suitable expressions for indexes of contrast, and choosing appropriate boundaries for "quantitative significance" and "insignificance," are prime descriptive challenges, ignored during stochastic controversies about P-values and confidence intervals, in the basic scientific aspects of biostatistics today.

I thank John Concato and Peter Peduzzi for helpful comments.

References

1. Fisher RA. **Statistical Methods for Research Workers**. Edinburgh: Oliver and Boyd; 1925.
2. Poisson SD. **Recherches sur la Probabilité des Jugements en Matiere Criminelle et en Matiere Civile**. Paris: Bachelier; 1837.
3. Matthews JR. **Quantification and the Quest for Medical Certainty**. Princeton, NJ: Princeton University Press; 1995.
4. Gavarret J. **Principes Généraux de Statistique Medicale**. Paris: Libraires de la Faculte de Medecine de Paris; 1840: 286–288.

5. Weaver W. **Lady Luck. The Theory of Probability**. New York: Dover Publications; 1982: 320–321.
6. Fisher RA. **Statistical Methods for Research Workers. 14th Edition**. Edinburgh: Oliver and Boyd; 1970: 44.
7. Fisher RA. **Statistical Methods and Scientific Inference**. Edinburgh: Oliver and Boyd; 1959: 42.
8. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. **Br Med J** 1986; 292:746–750.
9. Fleiss JL. Confidence intervals vs. significance tests: quantitative interpretation. **Am J Public Health** 1986; 76: 587. (Letter to editor).
10. Simon JL. Resampling: The "new statistics." Belmont, CA: Wadsworth Publishing; 1993.
11. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. **N Engl J Med** 1988; 318: 1728–1733.
12. Cohen J. Some statistical issues in psychological research. In: Wolman BB, Ed. **Handbook of Clinical Psychology**. New York: McGraw-Hill; 1965: 95–121.
13. Berry DA, Ed. **Bayesian Biostatistics**. New York, NY: Marcel Dekker; 1996.
14. Weinstein MC, Fineberg HV. **Clinical Decision Analysis**. Philadelphia: Saunders; 1980.
15. Burnand B, Kernan WN, Feinstein AR. Indexes and boundaries for "quantitative significance" in statistical decisions. **J Clin Epidemiol** 1990; 43: 1273–1284.
16. Guilford JP. **Fundamental Statistics in Psychology and Education**. New York: McGraw-Hill; 1956.