REVISTA MEXICANA DE ANÁLISIS DE LA CONDUCTA
MEXICAN JOURNAL OF BEHAVIOR ANALYSIS
1999, **25**, 39-68
NÚMERO 1 (JUNIO)
NUMBER 1 (JUNE)

# PROBLEMS OF STATISTICAL INFERENCE

## PROBLEMAS DE LA INFERENCIA ESTADÍSTICA

**HAYNE W. REESE**[1]
WEST VIRGINIA UNIVERSITY

## ABSTRACT

The standard method of statistical inference involves testing a null hypothesis that the researcher usually hopes to reject in order to accept a specific alternative hypothesis. The method is problematic in some ways; for example, consistency with a stringent underlying mathematical model and random sampling are needed, in principle, and decisions need to be based on not only objective outcomes of tests but also subjective evaluation of effect sizes. Some other problems have been wrongly attributed to it, for example, that it often involves null hypotheses that are preposterous or obviously false, permits only a decision to accept or to reject the null hypothesis, is misdirected because the real issue is not accept/reject but degree of belief, is an unfortunate hybridization, involves faulty syllogistic reasoning, and is inferior to the Bayesian approach. These criticisms are answered in this article; the conclusion is that the standard method is sound and, unless misused, it is valuable.

Keywords: Bayesian approach, Fisherian approach, inferential statistics, meta-analysis, methodology, null-hypothesis testing, research design, statistical inference, Type I and II errors

## RESUMEN

El método estandard de la estadística inferencial involucra probar una hipótesis nula, que el investigador usualmente desea rechazar para aceptar una hipótesis alterna específica. El método es problemático en algunos respectos; por ejemplo, es necesario, en principio, tanto la consistencia estricta con el modelo matemático subyacente, como contar con un muestreo aleatorio y las decisiones necesitan basarse, no sólo en pruebas

[1] I am indebted to the reviewers for identifying parts of the original paper that were incorrect, infelicitous, or otherwise inappropriate. Responding to them has greatly improved the paper, I believe. Address reprint requests to Hayne W. Reese, Department of Psychology, P.O. Box 6040, West Virginia University, Morgantown, WV 26506-6040, USA.

con resultados objetivos, sino también en evaluaciones subjetivas de los efectos debidos al tamaño de la muestra. Algunos otros problemas se han atribuído erróneamente a dicho método. Por ejemplo, que el método generalmente involucra hipótesis nulas que son totalmente absurdas o obviamente falsas, o que sólo permite una decisión respecto a la aceptación o rechazo de la hipótesis nula; éstas son críticas erróneas puesto que el problema real no es aceptar/rechazar, si no el grado de creencia. Este razonamiento está mal dirigido porque es un híbrido desafortunado, que involucra razonamiento silogístico falso y que es inferior a la aproximación Bayesiana. En este trabajo se responde a estas críticas; la conclusión es que el método estandard es sólido y que, a menos que sea mal utilizado, tiene valor.

Palabras clave: Aproximación Bayesiana, aproximación Fisheriana, estadística inferencial, meta-análisis, metodología, prueba de la hipótesis nula, diseño experimental, inferencia estadística, erores tipo I y tipo II

---

In this article, I summarize the standard method of statistical inference and discuss criticisms of this method. My goal is to allay doubts about the logic of statistical inference that the criticisms might have engendered in group researchers and, especially, behavior analysts. I have argued in a companion article (Reese, 1998) that group research methodology is legitimate for behavior analysis. Statistical inference is traditionally an essential component of this methodology, and unless doubts about statistical inference are allayed, behavior analysts who otherwise might lean toward using group research methodology might for the wrong reasons decide against using it. Actually, group research methodology can be used without statistical inference, as some critics have noted (the point is discussed later), and statistical inference can be used in single- subject research, as others have argued (references cited by Hopkins, Cole, & Mason, 1998). I agree with the latter argument, and thus I disagree with the conclusion of Hopkins et al. (1998) that at best, statistical inference is not useful to behavior analysts. Empirical demonstrations that visual inspection of graphs is only moderately reliable (DeProspero & Cohen, 1979; Fisch, 1998) are highly relevant to this disagreement; I address some other relevant issues elsewhere (Reese, 1998, in press a).

## The Standard Method of Statistical Inference

### The Role of Probability

The standard method of statistical inference is not arcane, not occult, not inherently dangerous, and in principle not difficult, even though many behavior analysts, some group researchers, and some statisticians seem to believe that it is. The mistaken belief is bolstered by an embarrassingly large

number of errors made by many proponents of the method in statistics textbooks and articles and in research reports (documented by, e.g., Cohen, 1994; Hagen, 1997, 1998; Schmidt, 1996; Tryon, 1998; and references they cited). However, the topics are not complex and most of the errors seem to reflect carelessness rather than misunderstanding (Schlesinger, 1991, p. 16). An analogy is a very large number of errors in descriptions of the Watson and Rayner (1920) "Little Albert" study in psychology textbooks and articles (documented by, e.g., Cornwell & Hobbs, 1976; B. Harris, 1979; Prytula, Oster, & Davis, 1977; Reese, in press b). Watson and Rayner's report was much more straightforward and simple than the error rate might imply (Reese, in press b).

***Probability as relative frequency.*** The standard method of statistical inference is based on the *relative-frequency*, or "frequentist" (e.g., Savage, 1961/1964), concept of probability. Probability in this sense is defined as the number of events of a specified kind in a population of events, relative to the number of all events in this population (e.g., Hays, 1963, chap. 2). Relative frequency can be determined empirically a posteriori or hypothetically a priori. For example, if the target events are ones and fives on a die thrown 80 times, the population of events is the specific 80 throws and the *empirical* probability of obtaining ones and fives is the actually obtained number of ones and fives, divided by 80. The *hypothetical* probability of ones and fives refers to an hypothetical population consisting of an infinite number of throws; it is one third if the die is unbiased.

Analogously, a researcher could determine the *empirical* probability of obtaining a particular class of treatment effects by replicating the study, as Cohen (1994) and Hubbard (1995) recommended and Fisher (1956, e.g., pp. 77-78) opposed. The class could be defined qualitatively as a particular direction of effects or quantitatively as a particular numerical range of effects. In either case, a fairly large number of replications would be needed to get a useful measure of the probability of the class. The measure would be meaningless unless the replications constituted a single population, which would require using the same procedures and random samples of research participants in all of the replications. More simply, a researcher could determine the *hypothetical* probability of the outcome of a single study, using the procedure outlined later.

***Probability as degree of belief.*** An alternative to the relative-frequency concept is "nonfrequency" or "personal" probability, defined as a person's degree of belief in the truth of a statement (e.g., Bakan, 1967, p. 60; Hagen, 1997; Kyburg & Smokler, 1961/1964). It is the basis of the Bayesian approach (discussed later), which is favored by many critics of the standard method. Degree of belief is sometimes called "level of confidence," but the latter phrase

is used in the standard method as a synonym of "level of significance," which is a relative-frequency probability (level of significance is discussed later, under *Step 2*). Degree of belief is also called "subjective probability" (e.g., Kyburg & Smokler, 1961/1964) and "opinion" (Savage, 1961/1964), though it is usually understood to refer to beliefs based on evidence (Kyburg & Smokler, 1961/1964).

For example, if a die is thrown six times and a five is obtained on four of the throws, the empirical relative frequency of fives (4/6) is so much larger than the hypothetical relative frequency (1/6) that an observer might suspect that the die is biased to yield fives. The observer's suspicion is interpretable as his or her degree of belief that the die is biased to yield fives. If a specific numerical value can be assigned to this degree of belief, the Bayesian approach can be used. The complement of this numerical value is the degree of belief that the die is not biased to yield fives. If subsequent throws of the die yield an empirical relative frequency of fives that is more likely to be obtained, given the specified numerical degree of belief than given its complement, the observer's degree of belief that the die is biased for fives increases. Otherwise, the degree of belief decreases (Hays, 1963, pp. 297-299).

**Steps in the Standard Method**

The standard method of statistical inference is formulated in the present subsection in terms of seven distinct steps, based on a summary by Lindquist (1956, p. 49 and other pages cited below). Criticisms of the steps are discussed in later sections.

*Step 1.* Formulate a null hypothesis to be tested. Unless the research is entirely exploratory, this step should also include formulating an alternative hypothesis (Fisher, 1966, pp. 12, 21; Hays, 1963, p. 250). The alternative hypothesis, which is often called an "experimental hypothesis" when it is explicitly formulated, may refer to an outcome hoped for on the basis of a theoretical prediction, or an expectation based on previous research, or merely an intuition. Implicit in Step 1 is prior formulation of a problem to be solved or a topic to be studied, which also includes selection of effects or behaviors to be observed.

*Step 2.* Select an acceptable "level of significance," or "alpha." Alpha is the risk of a Type I error, which is rejecting a true null hypothesis. This step should also include selecting an acceptable level of risk ("beta") of a Type II error, which is failing to reject a false null hypothesis (Lindquist, pp. 66-68). Fisher (1966, p. 17) said that this concept is meaningful only when the test involves a series of hypotheses about a population value. However, he was only partially correct, as shown below, and although this substep is often

omitted in practice, it should be taken.

Alpha is arbitrarily selected a priori, but in psychology it is conventionally set at 0.05. A minimum value of beta can also be arbitrarily selected a priori: Researchers can estimate the standard deviation of the measure to be used, arbitrarily select a minimum value of beta and a sample size, and on the basis of these values compute the minimum magnitude of effect that should be statistically significant. Alternatively, researchers can estimate the standard deviation, arbitrarily select a minimum value of beta and the minimum magnitude of effect that they would want to be identified as statistically significant, and compute the sample size that is needed to attain the preselected minimum value of beta (e.g., Walker & Lev, 1953, pp. 72-76, 163-167). The same formulas can be used to estimate beta after the data have been collected; for this purpose, the sample size is known and the standard deviation and magnitude of the effect are estimated from the data.

Based on a review of published research, Sedlmeier and Gigerenzer (1989) concluded that the average actual risk of a Type II error is extremely large, about 0.60. Many critics of the standard method have accepted this conclusion (e.g., Cohen, 1994; Hopkins et al., 1998; Hunter, 1997). Indeed, Hunter (1997) cited the figure as universal rather than an average: The standard method "has been shown to have a 60% error rate" and "the error rate for the significance test is 12 times larger than researchers think it is" (p. 3). Actually, however, the risk of a Type II error depends on the magnitude of the true effect (e.g., Abelson, 1997; Estes, 1997; Hays, 1963, p. 270; Winer, 1962, p. 12) and although the magnitudes of true effects can be estimated or, as indicated in the preceding paragraph, set at an a priori minimum value, the true magnitudes are indeterminable. Therefore, the actual risk of a Type II error is also indeterminable (Estes, 1997; Lindquist, 1956, p. 72). This point provides the justification for Fisher's statement mentioned above; but despite this justification, selecting a desired value of beta is possible and, as shown later, can be useful.

*Step 3.* Select a way to quantify (measure) the observations. If the observations are categorical and not otherwise quantifiable, use number of cases per category as the quantification. Use the measure of deviation selected in the next step to compute the deviation of the observations from the null hypothesis.

*Step 4.* Select a measure of the extent to which observations deviate from the null hypothesis, using a measure that has a determinable sampling distribution. Determine the sampling distribution of the measure on the assumption that the null hypothesis is true.

*Step 5.* Use the selected alpha to separate a "region of acceptance" from a "region of rejection" in the hypothetical sampling distribution

determined in Step 4. The region of acceptance is sometimes called the "region of nonrejection" (e.g., Hays, 1963, p. 271; Winer, 1962, p. 12) to indicate that the null hypothesis may be *retained as tenable* rather than *accepted* when the obtained deviation falls in this region.

The region of rejection can be located entirely in one tail of the hypothetical sampling distribution or partly in both tails, depending on the nature of the hypothesis being tested. For example, if the experimental hypothesis is that a particular kind of training yields stimulus equivalence, the null hypothesis can be that it does not yield stimulus equivalence and a one-tailed test can be used. The best procedure, however, is usually to divide the region of rejection equally between the two tails. Even when the stated experimental hypothesis can be supported only by deviations in one of the tails, a two-tailed test should be used if deviations opposite to the experimental hypothesis would be interpretable. In the example, the training might yield reversal of the expected responses--if equivalence would be indicated by $S_1$--$R_1$ and $S_2$--$R_2$, reversal would be $S_1$--$R_2$ and $S_2$--$R_1$. If the test is one-tailed, the reversal is not detectable; the statistically justified conclusion is not that reversal occurred, but only that stimulus equivalence was not obtained.

*Step 6.* Obtain a random sample of relevant observations from a relevant population.

*Step 7.* If the obtained measure of deviation falls in the region of acceptance, the outcome of the study is statistically nonsignificant and the null hypothesis is retained (or accepted). If the obtained measure of deviation falls in the region of rejection, the outcome of the study is statistically significant, the null hypothesis is rejected, and a plausible alternative hypothesis is accepted.

## Real and Fancied Problems in the Steps

### Step 1: Meaningfulness of Null Hypotheses

***Alleged implausibility of null hypotheses.*** Ward Edwards (1965) objected to null hypothesis testing on the following argument: "Many null hypotheses tested by classical procedures [i.e., the standard method] are scientifically preposterous, not worthy of a moment's credence even as approximations. If a hypothesis is preposterous to start with, . . . why test it?" (pp. 401-402). Edwards did not specify how null hypotheses might be scientifically preposterous and he did not give any examples; but in any case, the answer to his question is that the standard method usually involves the hope that the null hypothesis will be rejected, thus permitting acceptance of an experimental hypothesis that has been theoretically predicted, empirically

expected, or intuited.

A related argument is that the "nil" hypothesis--the null hypothesis that populations do not differ (Cohen, 1994)--is not true unless it is true for an infinite number of decimal places (Loftus, 1996) or at least for a large number of decimal places (Cohen, 1994). This argument is misleading because it cannot legitimately be limited to "nil" hypotheses (which actually are not usefully distinguished from other null hypotheses and therefore do not deserve a separate name). The argument must encompass any hypothesized relation among populations; for example, if the hypothesis is that the difference between two population means is 10 units, the argument implies that this hypothesis is true only if the difference is 10.0 followed by zeros to an indefinitely large number of decimal places. The argument therefore implies that no hypothesis in physics, chemistry, psychology, or any other science can be true because the degree of precision required by the argument has not been attained in any science and in principle cannot be attained (Popper, 1974, p. 280). Given that the required degree of precision cannot be attained, the argument implies that no scientific hypothesis is falsifiable and therefore either (a) science is actually pseudoscience because it cannot conform to Popper's (1983, passim) "falsifiability" criterion for the designation *science* or (b) the argument under consideration is irrelevant to science. I think the latter alternative is obviously the more reasonable one.

A null hypothesis will not be rejected if it is false only in a decimal place beyond detectability by available instruments. In fact, even if psychologists had instruments that could detect almost infinitesimal differences, group research methodology would make testing the null hypothesis worth while because the variability of the measurements would almost surely be so large that the obtained effect would not be statistically significant. The concept of *power* is relevant to this point; it is discussed later.

***Argument by Paul Meehl.*** Meehl (1967) argued against testing the null hypothesis because it is "[quasi-] always false" in biological, social, and behavioral sciences (pp. 108, 110; his brackets). His premise was that "it is highly unlikely that *any* psychologically discriminable stimulation which we apply to an experimental subject would exert literally *zero* effect upon any aspect of his performance" (p. 109). His argument is undermined by three problematic phrases in the premise and by irrelevance of nonzero effects that are theoretically or practically negligible, but it is not problematic under certain conditions.

(a) Meehl's phrase "highly unlikely" presumably accounts for the bracketed "quasi," which in Latin means "as if" but in Meehl's use fits the standard English meaning "almost" or "nearly" (*Oxford*, 1989, pp. 1001-1002). However, just as Watson (1913) failed to see the contradiction (noted

by Bergmann, 1956) between denying the existence of mental images and, in a footnote on the same page, admitting the possibility of "a sporadic few," Meehl failed to see that his own argument was undermined by the phrase "highly unlikely." The phrase undermines his argument because it admits the possibility that any given null hypothesis may be true and therefore requires that all null hypotheses be tested.

(b) Meehl's phrase "psychologically discriminable stimulation" excludes subthreshold magnitudes of stimulation, but the only evidence that stimulation is subthreshold is that no aspect of performance is affected. Consequently, Meehl's argument is tautological: Psychologically discriminable stimulation has *by definition* an effect on some aspect of performance; therefore, if the stimulation is psychologically discriminable, the null hypothesis referring to the affected aspect is false by this definition. Conversely, stimulation that is not psychologically discriminable has *by definition* no effect on any aspect of performance; therefore, if the stimulation is not psychologically discriminable, the null hypothesis is true by this definition.

A premise used by W. Edwards, Lindman, and Savage (1963) is subject to this criticism although they used "real" instead of "psychologically discriminable." They said: "Convention asks, 'Do these two programs differ at all in effectiveness?' Of course they do. Could any real difference in the programs fail to induce at least some slight difference in their effectiveness?" (pp. 215-216).

(c) Meehl's phrase "any aspect" is ambiguous because in the context of his premise it could mean either "every aspect" of performance or "some aspect" of performance. The first meaning makes the premise obviously false because of empirical evidence that stimulation which is psychologically discriminable with respect to some aspects of performance can be subthreshold with respect to other aspects. For example, a stimulus that has been previously experienced may be superthreshold with respect to affective responses but subthreshold with respect to verbal indicants of recognition (e.g., Kunst-Wilson & Zajonc, 1980). The second possible meaning of "any aspect" is no better because it undermines the relevance of the premise: Any given researcher is interested in selected aspects of performance, and even if Meehl's premise were correct, it would be irrelevant whenever the affected aspects were not the aspects selected for study (Hagen, 1997).

(d) Even if Meehl's premise were correct with respect to aspects of performance that were being studied, it would be irrelevant whenever the effects were theoretically or practically negligible or so small that they were not detectable--a point that W. Edwards et al. (1963) acknowledged with respect to their own premise. An analogy is that the trajectory of a bullet fired from a gun on the earth is influenced by gravitational attraction of not only the

earth but also all other bodies in the universe, yet predictions of the trajectory will not be measurably wrong if the other bodies are ignored (Nagel, 1961, footnote 8, pp. 560-561). Also, the shooter's body may react measurably to the firing of the gun, but the vast majority of the bodies in the universe will not react in any measurable way.

(e) Meehl's premise is not problematic under three conditions: (i) The efficacy of a treatment is being tested; (ii) Meehl's premise is correct with respect to aspects of performance that are relevant to treatment efficacy; and (iii) the effects are large enough to warrant rejecting the null hypothesis. In this case, the researcher finds (with a two-tailed test) either that the treatment has the desired effect or that it has the opposite effect.

**Step 2: Interpretation of Alpha and Beta**

Several authors have misinterpreted alpha and beta. For example: (a) Cortina and Dunlap (1997, p. 166) said that alpha "is the Type I error rate, regardless of whether or not the null is true"; (b) Rosnow and Rosenthal (1996b, p. 254) said that the obtained probability is "the obtained probability of a Type I error in a test of statistical significance"; (c) Walker and Lev (1953, p. 62) said, "The probability of rejecting a hypothesis is called the power of the test" (italicized in original); and (d) several authors said that power should be maximized (e.g., Fisher, 1966, p. 22; Hays, 1963, p. 287; Walker & Lev, 1953, p. 62). For other examples, see Pollard and Richardson (1987). The first three misinterpretations result from neglecting to mention that alpha and beta are *conditional* probabilities: Alpha is the Type I error rate, conditional on a true null hypothesis; and beta is the Type II error rate, conditional on a false null hypothesis. The complement of beta (1 - beta) is also a conditional probability; it is the *power* of the test, that is, the probability of rejecting the null hypothesis, conditional on its being false (e.g., Cortina & Dunlap, 1997; Pollard & Richardson, 1987; Walker & Lev, 1953, p. 60).

The crucial points are: (a) The Type I error is *rejecting* a *true* null hypothesis and therefore it cannot occur if the null hypothesis is *retained*, regardless of whether the null hypothesis is true or false, and it cannot occur if the null hypothesis is *false*, regardless of whether the null hypothesis is retained or rejected. Consequently, if the null hypothesis is true, the probability of a Type I error is alpha and the concepts of Type II error and power are irrelevant. (b) Conversely, the Type II error is *retaining* a *false* null hypothesis and therefore it cannot occur if the null hypothesis is *rejected*, regardless of whether the null hypothesis is true or false, and it cannot occur if the null hypothesis is *true*, regardless of whether the null hypothesis is retained or rejected. Consequently, if the null hypothesis is false, the probability of a Type

II error is beta, power is equal to one minus beta, and the concept of Type I error is irrelevant. (c) A test has too much power if it identifies a trivial effect as statistically significant; the aim should therefore be to *optimize* power rather than to *maximize* it. (Each of the following authors, among others, discussed some of these points, but none discussed all of them: Bakan, 1967, chap. 1; Cortina & Dunlap, 1997; Fisher, 1956, pp. 41-46; Frick, 1996; Hays, 1963, pp. 280-281; Hopkins et al., 1998; Neyman & Pearson, 1928; Pollard & Richardson, 1987; Schmidt, 1996; Walker & Lev, 1953, pp. 60-63, 163; Winer, 1962, p. 11).

Frick (1996) summarized recommendations to minimize "the total probability of making an error," that is, alpha plus beta (p. 387); but the foregoing considerations cast some doubt on these recommendations. Of course, a researcher does not know in advance whether the null hypothesis is true or false and therefore needs to guard against both kinds of error. However, only one kind of error can be made about any one null hypothesis (Schmidt, 1996). Consequently, although reducing the probability preselected as alpha increases beta when the null hypothesis is false, it cannot affect beta when the null hypothesis is true.

Therefore, researchers should use the degree of rigor with respect to Type I errors that has become standard, which is the conventional alpha of 0.05, or they should explicitly argue for a different alpha persuasively enough to convince readers that a different degree of rigor is needed. They should also try to use a beta commensurate with the smallest *meaningful* deviation of sample observations from the null hypothesis. Experiments are conducted for many different reasons, as Sidman (1960, pp. 4-40) showed, and different reasons might well entail different definitions of a meaningful deviation. For example, theoretical meaningfulness might be consistent with a smaller deviation than practical meaningfulness.

**Step 3: Measurement**

Step 3 is problematic because humans tend to make errors in observing, measuring, scoring, transcribing, and the like (e.g., Fisch, 1998). These problems are not discussed herein, not because they are unimportant, but because they are not unique to the standard method of statistical inference and because careful training of the observers, scorers, and so on should minimize these errors whether the standard method or some other method is used.

**Step 4: Hypothetical Sampling Distribution**

In Step 4, the sampling distribution of the measure of deviation is

determined on the basis of a mathematical model. A problem is that the sample of observations obtained in Step 6 is often discrepant from this model, thus implying that the population from which the sample was obtained is also discrepant from the model. For example, if two treatments are being compared and the null hypothesis is that their effects are the same, and if the measure of deviation is, say, $F$ in an analysis of variance, then the mathematical model requires assuming that the samples of observations were drawn (a) at random from two populations of treatment effects that (b) have identical means, (c) are normally distributed, and (d) have identical variances.

The first assumption is discussed later, in the subsection on Step 6. The assumption of identical means is usually tested because it is usually the null hypothesis, but the distribution and variance assumptions are usually not tested even though tests are available. A rationale for not testing the latter assumptions is that Monte Carlo research has demonstrated that for many measures of deviation used in psychological research, even fairly large discrepancies from normal distributions and equal variances can be ignored because the actual sampling distributions are acceptably close to the hypothetical sampling distributions (e.g., Hays, 1963, pp. 321-322, 378-379; Lindquist, 1956, pp. 73-86). However, when repeated-measures (within-groups) designs are used, as in most behavior analytic research, the actual sampling distributions are sensitive to smaller discrepancies from the underlying model, especially with respect to homogeneity of covariances, which is assumed in the model underlying repeated-measures designs with more than two conditions repeated (Hertzog & Rovine, 1985; McCall & Appelbaum, 1973; Winer, 1962, pp. 369-374). Analysis of covariance is also sensitive to relatively small discrepancies from the underlying model, and one assumption that is often violated is homogeneity of regression (Lindquist, 1956, pp. 328-330).

**Step 5: Regions of Acceptance and Rejection**

Ward Edwards (1965) argued that the standard method is "always violently biased against the null hypothesis" (p. 400) and that researchers should avoid this bias by formulating their predictions such that retaining the null hypothesis confirms the predictions. This argument is flawed: (a) The alleged bias against the null hypothesis occurs only in highly constrained situations (Wilson, Miller, & Lower, 1967). (b) Rejecting the null hypothesis implicates the logically valid argument of denying the consequent, but accepting the null hypothesis implicates the logical fallacy of affirming the consequent (these logical arguments are discussed later). (c) The standard method does not by itself permit acceptance of the null hypothesis when the measure of deviation falls in the region of acceptance (e.g., Fisher, 1966, p. 16).

Rozeboom (1960) arrived at the same recommendation as W. Edwards from a different argument, but he also overlooked this point. It is discussed later. (Wilson et al., 1967, gave additional arguments against W. Edwards's recommendation.)

### Step 6: Sampling

In Step 6 of the standard method, the nature of the topic or the population under investigation may make obtaining a truly random sample of observations difficult or even impossible. Two problems arise. One problem is that some statistics, such as *F* and *t*, are based on a mathematical model in which the sample means and variances are independent. The means and variances are necessarily independent if samples are drawn at random from a normally distributed population (Lindquist, 1956, chap. 2), but they are not necessarily independent if the samples are not actually random. In the latter case, the means and variances may still be independent, but the assumption of independence needs to be supported by empirical evidence that the means and variances are not significantly correlated in a reasonably large set of samples.

The other problem is that generalization of findings from a sample to a population is justified only if the sample is a good representation of the population. Most of the random samples in a sampling distribution are good representations; therefore, any given random sample is likely to be a good representation. Consequently, if a sample is not actually random, generalization is strictly justifiable only to an hypothetical population defined as "the population from which the obtained sample is a random sample." Few researchers give serious thought to how this population might differ from the population of real interest.

The problem of generalization also arises in single-subject research. On the one hand, single-subject researchers usually include several subjects in order to demonstrate generality (e.g., Perone, 1994); but on the other hand, these researchers usually want to generalize their findings beyond these specific subjects. They want to generalize their findings to a population of similar individuals, but this generalization is justified only if the behavior of the sample of subjects is indeed similar to behavior in the population. Put another way, the generalization requires that the behavior of the sample is a good representation of the behavior of the population; this requirement is met, in principle, if the sample is a random sample from the population.

### Step 7: Testing Null Hypotheses

Loftus (1996) said that "carefully crafted" objections to reliance on

testing the null hypothesis have been published periodically, and he lamented that they "just kind of dissolve away in the vast acid bath of existing methodological orthodoxy" (p. 162). Actually, regardless of how carefully crafted they were, many of the objections were either flawed or irrelevant and, as Cohen (1995) and Hagen (1997, 1998) said, others referred not to the method as such but to misuse of the method. For example, a common criticism is that the standard method of statistical inference is illogical, but the alleged logical flaws in the method are actually psychological flaws in some users of the method. Another common comment is that the method has been "discredited," but this word is unwarranted; the warranted word is "criticized." Several alleged flaws are discussed in the present subsection. Another common criticism refers to the availability of better methods; this criticism is discussed in a later section.

***Direction of effect.*** Kaiser (1960) said that the standard method is flawed because when a two-tailed test of the null hypothesis is used, "we cannot logically make a directional statistical decision or statement when the null hypothesis is rejected on the basis of the difference in the observed sample means" (p. 160; italicized in original). Meehl (1967) enthusiastically endorsed this position, R. J. Harris (1997) found it entirely reasonable, and Schmidt (1996, p. 122) implicitly endorsed it in commenting that the standard method involves the null hypothesis but no alternative hypothesis.

Kaiser's point reminds me of an episode in the movie *No Time for Sergeants*: Ben Whitledge and Will Stockdale are Air Force recruits and Ben has chided Will for reacting to a female captain as a female rather than as an officer; later, Will sees her across a room and says to the sergeant, "I don't notice whether it's a man or a woman or what. All I see is a captain, and that's all." The sergeant infers that Will has bad eyesight. (This episode in the movie is a bowdlerization of the episode in the original novel, by Hyman, 1954, pp. 114-119.) Analogously, anyone who denies seeing the direction of the obtained means seems likely to be feigning ignorance or to have bad eyesight.

The flaw in Kaiser's argument is revealed by considering the rationale underlying tests of the null hypothesis. These tests are implicitly based on the commonsense expectation that uncommon things--things which rarely happen--do not happen to *us* (Bakan, 1967, p. 5; Fisher, 1966, p. 14). This expectation justifies the commonsense assumption that if we know that an event actually occurred and we know nothing else about it, we are justified in believing that the event was not a *rare* event. Of course, if we do know something else about the event, we may be justified in believing that a rare event actually occurred; if our lottery ticket wins the jackpot, our prior knowledge about lotteries will lead us to believe that a rare event happened to us. (If the prior knowledge can be quantified as a personal probability, the

Bayesian approach can be used to quantify the belief that a rare event actually happened.) However, if we observe a snowfall in Death Valley in July and know nothing else about Death Valley in July, we are justified in believing that this snowfall was not a rare event.

In tests of the null hypothesis, *rare* is given a precise definition--for a two-tailed test, an obtained two-tailed probability equal to or less than alpha. Despite this refinement, the commonsense assumption is still used in the standard method because prior knowledge is ignored: Any *obtained* effect is assumed not to be rare (i.e., not to have a true probability of alpha or less). Based on the commonsense assumption, whenever an event would be rare if the null hypothesis were true, we conclude that the null hypothesis is false. Furthermore, the logic of this decision not only permits but demands a conclusion about the direction of the true effect--the obtained effect would be even rarer if the true effect lies in one of the two possible directions away from the null value, but it would be less rare if it lies in the other possible direction. Based on the commonsense assumption that obtained events are not rare events, the conclusion must be that the true effect lies in the direction in which it is less rare (i.e., has a probability greater than alpha).

***Truth of the null hypothesis.*** Several authors argued that testing null hypotheses is useless because the null hypothesis is usually or always false, even if only to the "tiny" degree (Cohen, 1994, p. 1000) discussed above in connection with Step 1 (Bakan, 1967, pp. 6-8, 29; Baril & Cannon, 1995; Cohen, 1994, 1995; W. Edwards et al., 1963; Loftus, 1993, 1996; Meehl, 1967; Thompson, 1998). The argument is flawed in two major ways.

The argument is flawed because, as Hagen (1997, 1998) commented, it assumes that everything is related to everything else and therefore it requires that "all measurable human characteristics--indeed, temperament, intelligence, health, and even age at marriage and length of life--would have to be related, at least to some degree, to the position of the planets when one was born and the distribution of leaves in one's teacup" (Hagen, 1998, p. 801). He added that unless the assumption has this cosmic scope, it does not support the conclusion that testing null hypotheses is useless.

The argument is also flawed because in the standard method of statistical inference the issue is not whether the null hypothesis is true but whether it can be rejected on the basis of a statistical test. Rejecting it on the argument that it is false because of some true but negligible deviation from the null value or because it cannot be true for an indefinitely large number of decimal places is trivial. Furthermore, if a one-tailed test is being used and the deviation from the null value falls in the region of acceptance, then regardless of the magnitude of the deviation, the null hypothesis is retained. Evidently, Bakan, Baril, and Cannon, and the others cited above assumed a two-tailed

test. The point that should be made is that if the null hypothesis is false because of some small true effect, two-tailed statistical tests will lead to rejecting the null hypothesis at a rate somewhat higher than one-half of alpha and consequently at the same rate will lead to accepting a false alternative hypothesis that refers to a large effect.

A related argument is that the null hypothesis is not worth testing because it will always be rejected if the sample size is large enough (e.g., Cohen, 1994; Thompson, 1998). However, this argument is implicitly based on the assumption that the null hypothesis is false; if the null hypothesis is true, it will be rejected at the rate determined by alpha regardless of the sample size (Hagen, 1997).

***Acceptance/rejection versus believability.*** Another objection to the standard method is that the goal of research is not to make an accept/reject decision about the null hypothesis but to effect a change in the believability of a proposition (e.g., Cohen, 1994; Rozeboom, 1960). This objection is correct but misleading. Effecting a change in the believability of a proposition is indeed a goal of most research, and the standard method does not yield a personal probability (degree of belief). The standard method nevertheless deals with this goal, but in the Discussion section rather than in the Results section.

The accept/reject (or retain/reject) decision about the null hypothesis has only an ancillary role, which is subservient to the goal of *describing* observed phenomena. The actual *findings* of a study are the descriptive statistics, not the inferential statistics (Cohen, 1994; Fisher, 1956, p. 4; Michael, 1974). The findings are the obtained means or other descriptive summaries of the observations, such as confidence interval estimates, error bars, and estimated effect sizes (e.g., Loftus, 1993; Rosenthal, 1993)[2]. Therefore, decisions about the findings should be based on more than the outcome of the statistical test. This point is discussed in the next subsection.

## Other Real and Fancied Problems in the Standard Method

### Statistical Conclusion Validity

***Definition.*** The issue raised at the end of the preceding subsection is about the "statistical conclusion validity" of the findings (Cook & Campbell, 1979, chap. 2). Statistical conclusion validity means that a given instance of accepting or retaining the null hypothesis is not a Type II error and a given

[2] Confidence interval estimates, error bars, and estimated effect sizes are considered to be descriptive statistics even though they involve assumptions about the population, such as symmetrical distribution.

instance of rejecting the null hypothesis is not a Type I error. Statistical tests indicate whether an outcome is statistically nonsignificant or significant, but they do not provide the required evidence about statistical conclusion validity. Rather, the researcher must provide this evidence by argument, which should take into account the magnitude of the obtained effects. Unfortunately, few researchers provide these arguments except when they are disputing the nonsignificance of a desired effect or the significance of an undesired effect. Consequently, the use of statistical inference has probably put a lot of erroneous decisions into the journals (e.g., Hopkins et al., 1998), but the fault is in the users, not in the method.

***Plausible arguments.*** Many commentators (e.g., Malgady, 1998; Schmidt, 1996) have pointed out that obtaining a nonsignificant effect is not sufficient evidence that the null hypothesis is true. The reason is that in the null sampling distribution, the measure of deviation falls in the region of acceptance all but alpha proportion of the time (1 - alpha) if the null hypothesis is true, but it also falls in this region some large proportion of the time (but less than 1-alpha) if the null hypothesis is false and the true effect falls in a wide range within the region of acceptance (e.g., A. L. Edwards, 1967, pp. 212-213; Lindquist, 1956, p. 67).

Nevertheless, a nonsignificant effect can be used as one premise in an argument that the null hypothesis is true. To be plausible, the argument should include at least the first two of the following points and preferably more: (a) The obtained effect is very small; for example, the differences among the obtained means are so small that even if the differences were real they would be negligible. Put another way, the estimated effect size is very small. Schmidt (1996, p. 120) commented erroneously that the size of an obtained effect is not reported in a "truly traditional analysis," but even a cursory examination of Fisher's works shows that his examples include descriptions of effects. (b) The obtained effect is statistically nonsignificant. (c) The obtained effect would have been statistically nonsignificant even if alpha had been larger; that is, the obtained effect does not "approach significance." (d) Power was sufficient to detect any nonnegligible effect. (e) The obtained effect would be theoretically or empirically anomalous if it were real.

Converse points would provide a plausible argument that the null hypothesis is indeed false and that an alternative hypothesis is true: (a') The obtained effect is large, or the null value is far outside the confidence interval estimate of the effect. (b') The obtained effect is statistically significant. (c') The obtained effect would have been statistically significant even if alpha had been smaller; that is, the obtained effect is "highly significant." (d') Power was not excessive (this point would be supported by Point a'). (e') The obtained effect was predicted or is plausibly explainable or is consistent with other

empirical evidence. (Cook and Campbell, 1979, chap. 2, discussed additional arguments.)

If an obtained effect falls in the region of acceptance, the null hypothesis cannot legitimately be rejected even "marginally" or "weakly" because an obtained effect is either nonsignificant or significant. The phrase "marginally significant" is meaningful only if it is used to indicate that the obtained probability is precisely equal to alpha, which designates the *margin* between the regions of acceptance and rejection. As Rosnow and Rosenthal (1996b) commented, Abelson (1996) slipped in saying that an obtained probability of 0.08 "weakly" supported rejecting the null hypothesis. However, if the obtained effect falls in the region of acceptance but is nevertheless large, the null hypothesis should be *retained* rather than *accepted* because in this case the power of the test was evidently too low to constitute an adequate test.

Conversely, if the obtained effect is small but nevertheless falls in the region of rejection, the logic of the standard method requires that the null hypothesis be rejected. However, in this situation the finding could be held to need replication on the argument that the test evidently had too much power, resulting in acceptance of an effect that is (or seems to be) theoretically or practically negligible. In this situation, using a sufficiently smaller alpha would make the obtained effect nonsignificant, but alpha cannot be changed legitimately after the fact and the obtained effect must therefore be reported to be significant. A good approach is to report an estimate of the effect size (always a good idea, as noted by Cohen, 1994; Hagen, 1998; Rosenthal, 1993). If it is small enough, the researcher can justifiably argue that although the effect is statistically significant, it can be ignored because it is so small. Of course, readers may disagree about effects that are "small enough" to justify the argument. In any case, however, the argument is not that the effect is not real, which would require arguing that a rare event actually occurred, but that the researcher explicitly or implicitly overestimated the sample size needed to obtain a desirable level of power of the test.

***A fictitious example.*** A fictitious experiment described by Loftus (1996) is relevant to the foregoing points. In the fictitious experiment the obtained means in two conditions, each with 20 subjects, were 5.05 and 5.03 on a 10-point scale and the $t$ for this difference was "practically zero" (p. 167). Loftus argued that the null hypothesis should not be accepted unless the power of the test was large, and he said, "[1] it can be shown easily that, given a particular mean difference, the smaller the $t$ value, the lower is power--and hence, [2] the less appropriate it is to accept the null hypothesis" (p. 167; bracketed index numbers added). Part [1] is accurate, but it does not justify Part [2] because the concept of power is relevant only if the null hypothesis is false in a nontrivial sense. The best point estimate of a true mean is the

obtained mean because the obtained mean is "unbiased," "consistent," "efficient," and "sufficient" (Hays, 1963, pp. 197-201). Therefore, the best point estimate of the true mean difference in Loftus's fictitious experiment is the obtained mean difference of 0.02. A difference of 0.02 on a 10-point scale is almost surely meaningless and therefore raising the issue of power in this case is almost surely trivial.

**Alleged Hybridization**

Cohen (1994) and Loftus (1996), among other critics of the standard method (e.g., Gigerenzer, 1993; Sedlmeier & Gigerenzer, 1989), said that the standard method is an unfortunate hybridization of methods developed by Fisher and by Neyman and Pearson. Neither Cohen nor Loftus cited a specific work by Neyman and Pearson, but the method Neyman and Pearson described in 1928 is actually the same as Fisher's (1966) method and both are consistent with the now standard method. Granted, Fisher and Neyman and Pearson criticized each other quite severely, but the criticisms were actually niggling, as are the differences between the approaches. Examples are given in the following paragraphs.

1. Fisher (1956, pp. 41-42) rejected the Neyman and Pearson concept of Type I and II error rates as relative frequencies in replications of a study, but he (p. 82) inadvertently resolved the disagreement by distinguishing between an actual series of replications from a real population and an hypothetical series of replications from an hypothetical population. That is, the disagreement disappears when a given actual sample is "regarded by an act of imagination" as drawn from an hypothetical population of samples. These considerations also indicate the sense in which statistical significance is sometimes said (e.g., by Hagen, 1997; Melton, 1962) to indicate that a finding is probably replicable. This interpretation does not mean that the same value of the statistic used will be obtained; it means that if a Type II error did not occur, the findings obtained in replications are likely to be in the same tail of the region of rejection--that is, statistically significant and in the same direction as the original finding (e.g., Bakan, 1967, p. 15). Put another way, it means that 95% of the 95% confidence intervals calculated in replications can be expected to include the true population value (Cohen, 1995), provided that a Type II error did not occur. This proviso is needed because the confidence interval is centered on the obtained value, but if a Type II error occurred, the confidence interval should be centered on the null value.

2. Loftus (1996), Gigerenzer (1993), and Sedlmeier and Gigerenzer (1989) said that Fisher used the phrase "level of significance" to mean "obtained probability" and Neyman and Pearson used this phrase in the now

standard sense of a preselected alpha. Actually, Fisher usually used "level of significance" and "significance" to mean alpha (1956, e.g., pp. 42, 60, 62, 66, 81; 1966, e.g., pp. 13-14, 25, 57, 187-189, 196-197), but he indeed sometimes used these terms to mean "obtained probability" (e.g., 1956, p. 49 and perhaps p. 39). Therefore, this difference dissolves into carelessness on Fisher's part.

3. Schmidt (1996) said, "[1] The concept of statistical power does not exist in Fisherian statistics. [2] In Fisherian statistics, the focus of attention is solely on the null hypothesis. [3] No alternative hypothesis is introduced" (p. 122; bracketed index numbers added). Sentences [1] and [3] are incorrect (see Fisher, 1966, pp. 12, 21-22) and Sentence [2] is correct but misleading because of the adverb "solely." Fisher (1966, e.g., p. 16) emphasized the Type I error rate and Neyman and Pearson (1928) emphasized the Type II error rate, but both Fisher (1966, p. 17) and Neyman and Pearson (1928) acknowledged both kinds of error. Fisher's (1966, p. 17) rationale was that the Type II error rate cannot be specified unless the experimental hypothesis specifies a precise value of the parameter being tested.

4. Bakan (1967, pp. 25-27) pointed out that Fisher's approach leads formally to a decision to reject versus inconclusiveness and Neyman and Pearson's approach leads to a decision to reject versus a decision to accept. The reason is that Fisher's approach is implicitly based on deductive logic, in which the null hypothesis can be validly rejected if it is inconsistent with the obtained effect and no logically valid conclusion about the null hypothesis is possible if it is consistent with the obtained effect (this logic is discussed in the next subsection). In contrast, the Neyman and Pearson approach is based on decision theory, in which samples from a batch of products are tested as a way to decide whether the batch should be rejected as flawed or accepted as unflawed. However, as Bakan commented, the basic ideas are the same in both approaches (p. 26) and Neyman and Pearson only "explicated what was already implicit is the work of the Fisher school" (p. 27).

### Deductive Logic in Null-Hypothesis Testing

Several commentators have argued that deductive logic is not involved in the standard method of statistical inference (e.g., Cohen, 1994; Falk, 1998; Hagen, 1997, 1998; Tryon, 1998). Actually, however, deductive logic provides the basic rationale of the standard method--the rationale for testing the null hypothesis rather than the experimental hypothesis that the researcher hopes to confirm. The standard method is consistent with syllogistic reasoning, as shown below. Syllogistic reasoning provides no valid way to confirm the antecedent clause in a conditional (if-then) proposition but it provides one valid

way to falsify that clause. Therefore, the standard method puts the null hypothesis rather than the experimental hypothesis in the antecedent clause. The issues are discussed in the following paragraphs.

***Cohen's argument***. Cohen (1994) argued that the logic of the standard method of statistical inference is flawed because the reasoning is probabilistic. He presented the following syllogism as a model of this reasoning.

[Syllogism 1] If the null hypothesis is correct [i.e., true], then these data are highly unlikely.

These data have occurred.

Therefore, the null hypothesis is highly unlikely. (Cohen, 1994, p. 998)

Cohen used the following syllogism as an analogy to show the flaw in Syllogism 1:

[Syllogism 2] If a person is an American, then he is probably not a member of Congress.

This person is a member of Congress.

Therefore, he is probably not an American. (ibid.)

Baril and Cannon (1995) and Cortina and Dunlap (1997) criticized Cohen's syllogisms but overlooked the fatal flaws, which are that (a) both syllogisms are logically invalid and (b) Syllogism 1 has a valid form but Syllogism 2 does not and therefore the valid form of Syllogism 1 is not challenged by any form of Syllogism 2. Both syllogisms are logically invalid because both violate the logical principle of the excluded middle. In bivalued logic, which is used in syllogistic reasoning, a proposition is either true or false and these truth values are mutually exclusive and exhaustive. Thus, the truth values represented by phrases such as "probably true" and "probably false" (or "highly unlikely," "probably not," etc.) are not legitimately assignable to the conclusions of syllogisms.

In the following revision of Syllogism 1 the conclusion is reworded to be consistent with bivalued logic and the premises are rewritten to conform to the wording actually used in the standard method of statistical inference--the premises refer to "rare event" rather than likelihood. The latter change does not affect the validity of the syllogism.

[Syllogism 1'] If the null hypothesis is true, then the obtained event is a rare event.

The obtained event is not a rare event.

Therefore, the null hypothesis is false.

The minor change from "highly unlikely" in the conclusion of Syllogism 1 to "false" in the conclusion of Syllogism 1' changes the argument from invalid to valid. Syllogism 1' is an instance of the valid argument of denying the consequent ("modus tollens").

The excluded middle in the conclusion of Syllogism 2 can be eliminated

by making the conclusion a negation--"Therefore, he is not an American"-- but this change does not make the syllogism valid and the syllogism has no valid form (compare Werkmeister, 1948, chap. 9, 11, 12). The reason is that the reference to probability in the consequent of the major premise cannot be eliminated without changing the meaning of this premise. The predicate term of the syllogism--"American"--is undistributed in the major premise because the word "probably" allows the possibility that some Americans are members of Congress. However, the conclusion stated in Syllogism 2 and the revised conclusion suggested above are illicit unless "American" is distributed in the major premise. The point may be more obvious if the major premise is changed to the following logically equivalent proposition: "Most Americans are not members of Congress." (The reference to probability is carried by "Most" in this version; it would be carried by "Some" in strictly formal logic, but the flaw would be the same.)

***Logic in the standard method.*** As indicated above, rejection of the null hypothesis is valid when the obtained outcome would be a rare event if the null hypothesis were true. Falk (1998, p. 798) said that the standard method of statistical inference is "a probabilistic imitation of modus tollens," but in fact the reasoning is not probabilistic even though probabilities are implicit in each of the three statements in Syllogism 1′. In the major premise, "rare event" refers to an empirical outcome, that is, finding that the obtained discrepancy is "an event with a probability of alpha or less"; but the premise is "If A, then rare event," not "If A, then probably rare event." The minor premise is a denial of the consequent in the major premise, and the denial is based on the commonsense assumption that obtained events are not rare events. However, the minor premise is not probabilistic; it is "Not rare event" rather than "Probably not rare event." Finally, the conclusion is "Therefore not A," which is not probabilistic even though it is implicitly accepted with the proviso "within limits of the risk of a Type I error."

Syllogism 1′ is therefore an accurate representation of the logic implicit in the standard method of statistical inference. However, in practice the researcher does not stop with the conclusion that the null hypothesis is false but rather proceeds to another syllogism:

[Syllogism 3] If alternative hypothesis *x* is true, then the obtained event is not a rare event.

The obtained event is not a rare event.

Therefore, alternative hypothesis *x* is true.

As in Syllogism 1′, the consequent in the major premise of Syllogism 3 refers to the empirical outcome and the minor premise is based on the commonsense assumption that obtained events are not rare events. However, Syllogism 3 is invalid in deductive logic--it is the fallacy of affirming the

consequent. Nevertheless, it is the basis of inductive reasoning (e.g., Cohen, 1994). Unless a study is entirely exploratory, it is intended to test a theoretically based prediction, an empirically based expectation, or an intuition and this test has the form of Syllogism 3. This test is therefore interpretable as based on the deductive fallacy of affirming the consequent or on valid inductive reasoning.

## Alternative Approaches

As mentioned earlier, the standard method of statistical inference has been criticized as less useful than alternative approaches that are available. The most commonly cited alternatives are discussed in the present section.

### Model Fitting

Nesselroade and McArdle (1997) and Granaas (1998) criticized the standard method because more precise approaches are often available. These approaches involve fitting the data to precisely specified mathematical-statistical models. These approaches are indeed precise, but they also involve testing a null hypothesis--the goodness-of-fit hypothesis that the data fit the model (Granaas, 1998; McArdle & Nesselroade, 1994). A persisting problem in this approach is that in practice the goodness-of-fit tests usually have too much power, that is, unless a very small alpha is adopted, the tests detect trivial discrepancies as statistically significant.

### Meta-Analysis

Another alternative approach is meta-analysis, which involves aggregating effect sizes obtained in studies that are related to a selected topic. Several authors (e.g., Lipsey & Wilson, 1993; Schmidt, 1996) argued that individual studies are inconclusive and that only meta-analysis can integrate results across studies. Failure to realize that any one study is inconclusive has probably sometimes led researchers to abandon a topic prematurely on the basis of one persuasive study (e.g., Hopkins et al., 1998), but meta-analysis may not be the solution. I have argued elsewhere (Reese, in press a) that many meta-analyses are problematic. One reason is that most of the studies that are integrated were published and therefore usually had effects large enough to be statistically significant. Another reason is that the studies are usually heterogeneous in potentially important ways, including populations sampled, procedures and manipulations or tests used, and interactions assessed (e.g., main effects and interactions of race and sex of participants are often not

analyzed). Another reason, noted by Abelson (1997), is that "cause sizes" are ignored but need to be analyzed.

I would add still another reason: An effect revealed by meta-analysis is arguably real, but it can be trivial. Schmidt (1996) gave two examples to illustrate the usefulness of meta-analysis. In one, the assumed true effect of a drug was 0.50 in *z* units (i.e., half a standard deviation above the null value) and in the other the estimated true correlation of a clerical aptitude test with job performance was 0.22. I question whether these effects are meaningful. For example, if a drug had an effect of half a standard deviation on Stanford-Binet IQ, the control group mean would be 8 IQ points less than the experimental group mean (e.g., 102 vs. 110); an effect this small is unlikely to have any practical implications. In Schmidt's second example, the correlation of 0.22 means that the clerical aptitude test accounted for only about 5% of the variance in job performance.

In short, meta-analysis usually has the same problems that traditionally made the phrase "cross-study comparison" a pejorative epithet. In order to deal with such problems the meta-analyst needs to use the traditional, somewhat subjective approach to reviewing the literature--identifying possible moderator variables and looking for relevant evidence in the studies reviewed. Lipsey and Wilson (1993) used this approach.

**Parameter Estimation**

A third alternative approach is to report confidence interval estimates and point estimates of population parameters. These estimates have been argued to be more informative--and more useful for later meta-analysis--than are the outcomes of statistical tests (e.g., Cohen, 1994; Schmidt, 1996). However, parameter estimates and statistical tests are informative in different ways because they have different functions, as noted earlier (compare Fisher, 1956, p. 57-60). Many research reports include confidence interval estimates, as Schmidt (1996) commented, and most include the obtained means, which as noted earlier are the best point estimates of the true (population) means. Estimates are indeed useful for later meta-analysis, but until enough reports are available to make a meta-analysis worthwhile and informative, the outcomes of statistical tests provide information that is useful. This information is especially useful, as Abelson (1997) pointed out, when it is surprising.

Effect sizes should be estimated and reported whether or not the null hypothesis is tested and if it is tested, whether it is retained, accepted, or rejected. One reason is to facilitate later meta-analyses, another is to let readers apply their own criteria for determining whether an obtained effect is "large" or "small." Effect sizes can be estimated in many different ways,

summarized by, for example, Rosenthal (1993), Rosnow and Rosenthal (1996a), and Tatsuoka (1993). A method recommended by Rosnow and Rosenthal (1996a) involves computing a "counternull" value and using it as one end of a confidence interval estimate of the true effect size and using the hypothetical null value as the other end. The counternull value is the value that would have yielded the same probability for the obtained effect if the counternull value had been used instead of the null value; that is, the obtained effect has the same probability in the null and counternull distributions. A problem is that the null-counternull interval overestimates the true effect size because it capitalizes on chance--it is a one-tailed interval in the direction of the obtained effect. Another problem is that the level of confidence in the interval is determined by the obtained two-tailed probability; if the obtained two-tailed probability is $p$, the percentage confidence is $100 \times (1 - p)$. For example, if the obtained probability is 0.20, the calculations yield an 80% confidence interval. Although 80% may seem to be a high level of confidence, standard practice is to use a 95% confidence interval, based on the use of 0.05 as a two-tailed alpha. Whenever this alpha is used and the obtained effect falls in the region of acceptance--as it does when computation of the null-counternull interval is useful--the 95% confidence interval computed in the standard way (e.g., Hays, 1963, pp. 287-291) will always include values on both sides of the null value.

Another recommended method of estimating effect sizes is to use the Pearson correlation between treatments and treatment effects (e.g., Rosnow & Rosenthal, 1996a). A problem that might arise is that the Pearson correlation assesses the magnitude of linear regression but not the magnitude of nonlinear regression. The problem does not arise in studies with only two treatments because regression is necessarily linear when only two data points are involved. However, if more than two treatments are involved, the researcher should test the linearity of the regression. If the test was not done and cannot be inferred from the reported data, the meta-analyst should not pool the outcome with the outcomes of studies in which regression was shown to be linear.

Another estimate, the final example herein, is based on the obtained value of Student's $t$ and the sample size per group. Assuming equal $N$:

$$d = \frac{t}{\sqrt{2/N}}$$

where $d$ is the estimated effect size. In Schmidt (1996, footnote 4, p. 125), this estimate was misprinted as $2t/\sqrt{N}$.

**The Bayesian Approach**

According to Cohen (1994), many researchers believe that when the null hypothesis is rejected, the probability that it is actually true is alpha. This belief is mistaken even though it may seem reasonable because alpha is the probability that rejecting the null hypothesis is a Type I error if the null hypothesis is *true*. Cohen said that researchers want to know the "inverse probability" $p(H_0|D)$, which is the conditional probability that the null hypothesis is true, given the data obtained. The standard method does not yield this inverse probability (e.g., Cohen, 1994; Fisher, 1956, pp. 37, 43-46; Hagen, 1997; Winkler, 1993). Rather, as Cohen noted, it yields $p(D|H_0)$, which is the conditional probability of obtaining the actually obtained data if the null hypothesis is true. Therefore, researchers who want to know the inverse probability should not use the standard method or, if they use it, they should not misinterpret it as indicating the inverse probability.

Many critics of the standard method have favored the Bayesian approach, based on the work of an early 18th century clergyman and mathematician, Thomas Bayes, which he declined to publish during his lifetime--it was published posthumously by his friend Richard Price (Fisher, 1956, pp. 8-9, 1966, p. 6). The critics seem to favor this approach because unlike the standard method, it assigns a value to the inverse probability $p(H_0|D)$. This probability is based in part on an a posteriori probability derived from the outcome of the research and in part on an a priori probability that the theory, prediction, or intuition is true--the researcher's prior degree of belief. At one extreme--absolute prior certainty--no empirical test is needed, and at the other extreme--no prior belief--the Bayesian approach cannot be used.

The central problem with the Bayesian approach is that the degree of belief is usually determined by each researcher who holds a prior belief and is therefore usually at least subjective (e.g., Frick, 1996; Hays, 1963, p. 299; McGraw, 1995; Wilson et al., 1967) and possibly also arbitrary and biased (Frick, 1996) and "in the nature of mathematical slight-of-hand" (Fisher, 1966, p. 198). The standard method does not have this problem because degrees of belief are not considered; in fact, taking the subjective personal element out of hypothesis testing is the major goal of the standard method. Nevertheless, as Fisher (1966, pp. 194, 198) said, if legitimate a priori probabilities are available, the Bayesian approach should be used. These probabilities might be degrees of belief, as Hagen (1997) said, or they might necessarily be relative-frequency probabilities, as Hays (1963, p. 116) implied. In any case, as Hays (p. 299) said, if an objective way to assess degrees of belief were developed, the Bayesian approach could be useful and might come to be used routinely. Unfortunately, more than three decades after Hays's statement, no objective

way to assess degrees of belief has been developed.

## CONCLUSION

My arguments in this article indicate that many of the flaws attributed to the standard method of statistical inference are actually flaws in researchers who use the method in unthinking, reflex ways. All the criticisms of the standard method can be challenged, and none should be accepted without deep study. They are best interpreted as warnings about how this method can be misused rather than as indications of inherent flaws in the method. Group researchers and behavior analysts who use the method and who have responded appropriately to the warnings can obtain highly useful information about the findings of a study. A major point, however, is that the *findings* of a study are the obtained means or other descriptive statistics, not the outcomes of the statistical tests. A comment by Nesselroade and McArdle (1997, p. 25) about mathematical modeling procedures is also relevant to the standard method of statistical inference, so much so that I quote it here as *l'envoi*:

> [Statistical methods] are the tools and not the craftsmen; they are the instruments and not the musicians. As with quality tools and fine musical instruments, full realization of the promise of these analytic devices requires a high level of familiarity and knowledge on the part of users. The more skillfully these instruments are used, the more impressive and valuable are the outcomes that they help to produce, and the greater will be the gains from applying them to substantive problems and issues. (p. 25; bracketed phrase added)

Neyman and Pearson (1928, p. 232) made exactly the same point.

## REFERENCES

Abelson, R. P. (1996). Vulnerability of contrast tests to simpler interpretations: An addendum to Rosnow and Rosenthal. *Psychological Science, 7,* 242-246.

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12-15.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation.* San Francisco, CA: Jossey-Bass.

Baril, G. L., & Cannon, J. T. (1995). What *is* the probability that null hypothesis testing is meaningless? *American Psychologist, 50,* 1098-1099.

Bergmann, G. (1956). The contribution of John B. Watson. *Psychological Review, 63,* 265-276.

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

Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist, 50,*

1103.
Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.
Cornwell, D., & Hobbs, S. (1976, March 18). The strange saga of little Albert. *New Society*, pp. 602-604.
Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*, 161-172.
DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.
Edwards, A. L. (1967). *Statistical methods* (2nd ed.). New York: Holt, Rinehart & Winston.
Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin, 63*, 400-402.
Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 79*, 193-242.
Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science, 8*, 18-20.
Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53*, 798-799.
Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21*, 111-123.
Fisher, R. A. (1956). *Statistical methods and scientific inference*. New York: Hafner.
Fisher, R. A. (1966). *The design of experiments* (8th ed.). Edinburgh:Oliver & Boyd.
Frick, R. W. (1996).The appropriate use of null hypothesis testing.• *Psychological Methods, 1*, 379-390.
Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren, & C. Lewis (Eds). *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311- 339). Hillsdale, NJ: Erlbaum.
Granaas, M. M. (1998). Model fitting: A better approach. *American Psychologist, 53*, 800-801.
Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.
Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53*, 801-803.
Harris, B. (1979). Whatever happened to Little Albert? *American Psychologist, 34*, 151-160.
Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8*, 8-11.
Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart, & Winston.
Hertzog, C., & Rovine, M. (1985). Repeated-measures analysis of variance in developmental research: Selected issues. *Child Development, 56*, 787-809.
Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21*, 125-137.
Hubbard, R. (1995). The earth is highly significantly round ($p < .0001$). *American*

*Psychologist, 50*, 1098.

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

Hyman, M. (1954). *No time for sergeants*. New York: Random House.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67*, 160-167.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207*, 557-558.

Kyburg, H. E., Jr., & Smokler, H. E. (1964). Introduction. In H. E. Kyburg, Jr., & H. E. Smokler (Eds.), *Studies in subjective probability* (pp. 1-15). New York: Wiley. (Original work published 1961).

Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Loftus, G. R. (1993). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers, 25*, 250-256.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161-171.

Malgady, R. G. (1998). In praise of value judgments in null hypothesis testing . . . and of "accepting" the null hypothesis. *American Psychologist, 53*, 797-798.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen, & H. W. Reese (Eds.), *Life-span developmental psychology:Methodological contributions* (pp. 223-267). Hillsdale, NJ: Erlbaum.

McCall, R. B., & Appelbaum, M. I. (1973). Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development, 44*, 401-415.

McGraw, K. O. (1995). Determining false alarm rates in null hypothesis testing research. *American Psychologist, 50*, 1099-1100.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647-653.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace, & World.

Nesselroade, J. R., & McArdle, J. J. (1997). On the mismatching of levels of abstraction in mathematical-statistical model fitting. In H. W. Reese & M. D. Franzen (Eds.), *Biological and neuropsychological mechanisms: Life-span developmental psychology* (pp. 23-49). Mahwah, NJ: Erlbaum.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika, 20A*, 175-240.

*Oxford English dictionary* (2nd ed.; prepared by J. A. Simpson, & E. S. C. Weiner; Vol. 12). (1989). Oxford: Oxford University Press.

Perone, M. (1994). Single-subject designs and developmental psychology. In S. H. Cohen, & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological contributions* (pp. 95-118). Hillsdale, NJ: Erlbaum.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102,* 159-163.

Popper, K. R. (1974). Scientific reduction and the essential incompleteness of all science. In F. J. Ayala, & T. Dobzhansky (Eds.), *Studies in the philosophy of biology: Reduction and related problems* (pp. 259-284). Berkeley: University of California Press.

Popper, K. R. (1983). *Realism and the aim of science* (from the *Postscript to the logic of scientific discovery*; W. W. Bartley, III, Ed.). Totowa, NJ: Rowman & Littlefield.

Prytula, R. E., Oster, G. D., & Davis, S. F. (1977). The "rat rabbit" problem: What did John B. Watson really do? *Teaching of Psychology, 4,* 44-46.

Reese, H. W. (1998). Utility of group methodology in behavior analysis and developmental psychology. *Mexican Journal of Behavior Analysis, 24,* 137-151.

Reese, H. W. (in press a). Strategies for replication research exemplified by replications of the Istomina study. *Developmental Review.*

Reese, H. W. (in press b). The Watson and Rayner "Little Albert" study. *Monographs of the Society for Research in Child Development.*

Rosenthal, R. (1993). Cumulating evidence. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519-559). Hillsdale, NJ: Erlbaum.

Rosnow, R. L., & Rosenthal, R. (1996a). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331-340.

Rosnow, R. L., & Rosenthal, R. (1996b). Contrasts and interactions redux: Five easy pieces. *Psychological Science, 7,* 253-257.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416-428.

Savage, L. J. (1964). The foundations of statistics reconsidered. In H. E. Kyburg, Jr., & H. E. Smokler (Eds.), *Studies in subjective probability* (pp. 173-188). New York: Wiley. (Original work published 1961).

Schlesinger, G. N. (1991). *The sweep of probability.* Notre Dame, IN: University of Notre Dame Press.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115-129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in*

*psychology*. New York: Basic Books.

Tatsuoka, M. (1993). Effect size. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 461-470). Hillsdale, NJ: Erlbaum.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53*, 796.

Walker, H. M., & Lev, J. (1953). *Statistical inference*. New York: Holt.

Watson, J. B. (1913). Image and affection in behavior. *Journal of Philosophy Psychology and Scientific Methods, 10*, 421-428.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*, 1-14.

Werkmeister, W. H. (1948). *An introduction to critical thinking*. Lincoln, NE: Johnsen.

Wilson, W., Miller, H. L., & Lower, J. S. (1967). Much ado about the null hypothesis. *Psychological Bulletin, 67*, 188-196.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

Winkler, R. L. (1993). Bayesian statistics: An overview. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 201-232). Hillsdale, NJ: Erlbaum.