DOCUMENT RESUME

ED 090 305 TM 003 593

AUTHOR Lai, Morris K.
TITLE A Noncentral Analysis of Variance Model Relating Statistical and Practical Significance.
REPORT NO R-A74-1
PUB DATE Apr 74
NOTE 35p.; Paper presented at the Annual Meeting of the American Education Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Analysis of Variance; Data Analysis; *Hypothesis Testing; *Mathematical Models; Probability; Research Problems; *Statistical Analysis; *Tests of Significance

ABSTRACT

When analysis of variance is used, statistically significant differences may or may not be of practical significance to educators. A large part of the problem is due to the fact that a "zero difference" null hypothesis can always be rejected statistically if the sample size is large enough. If, however, a method based on the noncentral F distribution is used, trivial differences cannot attain statistical significance. The (non-zero) null hypothesis is now rejected at the alpha level when the observed F exceeds the noncentral F cutoff point where the noncentrality parameter delta (sub0) is determined by the minimum practical difference set by the researcher. (Author)

TEACHER EDUCATION DIVISION PUBLICATION SERIES

A MONCENTRAL ANALYSIS OF VARIANCE MODEL RELATING STATISTICAL AND PRACTICAL SIGNIFICANCE

Morris K. Lei

Paper presented at the meeting of the American Educational Research Association, Chicago, April 1974

REPORT AZ4-1

FAR WEST LABORATORY FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
1855 Folsom Street, San Francisco, California, 94103, (415) 565-3000

A NONCENTRAL ANALYSIS OF VARIANCE MODEL RELATING STATISTICAL AND PRACTICAL SIGNIFICANCE

Introduction

Statement of the Problem

One of the most widely used methods of analyzing research data in the behavioral sciences is the analysis of variance (ANOVA), particularly the fixed effects model (Morrison & Henkel, 1969). Integrally tied in with this model is the idea of hypothesis testing in the form of tests of statistical significance. Of three types of statistical inference--point estimation, interval estimation, and hypothesis testing--behavioral scientists have devoted themselves almost exclusively to hypothesis testing (Heermann & Braskamp, 1970).

Several writers have criticized the current use of ANOVA (Selvin, 1957; DuBois, 1965; Bakan, 1966; Lykken, 1968; Fleiss, 1969; Overall, 1969). Other writers have suggested that with appropriate corrective steps, the basic ANOVA model is an exemplary method of analyzing data and obtaining meaningful results (Horst, 1967; Kempthorne & Doerfler, 1969; Winch & Campbell, 1969).

Some critics have argued that tests of significance, as done in ANOVA, essentially should not be used (e.g., Morrison & Henkel, 1970); however, the pervasive influence of tradition has been recognized (Sterling, 1959; Rozeboom, 1960; Lykken, 1968; Heermann & Braskamp, 1970). More recently Walker and Schaffarzick (1974) while reluctantly using the criterion of statistical significance to compare studies, expressed the hope for an improved methodology.

It would seem valuable to modify the ANOVA model such that some inherent weaknesses including those discussed by the aforementioned critics, are overcome. In particular, it would be desirable to relate practical significance more closely to statistical significance.

The notion of practical significance is complex in and of itself. There is no commonly accepted method of determining practicality. In educational research, where outcomes are not easily described in cost-benefit terms, it is often quite difficult to decide if a difference due to treatment is of educational or practical significance. Nevertheless, such assessments of practical significance are being made, and the current ANOVA model does not adequately handle the issue of practical significance.

An analysis of variance model, based on the noncentral F distribution, is presented in this paper as an attempt to improve upon the currently used ANOVA model, in particular in the area of the inadequate handling of practical significance.

Review of Related Research

Criticisms of Significance Testing

The literature critical of significance testing has appeared mainly in the past 15-17 years (Morrison & Henkel, 1970). Periodically researchers have been reminded that statistical significance does not necessarily imply practical significance (Selvin, 1957; DuBois, 1965; Mendenhall, 1968; Glass & Hakstian, 1969). In essence, what this warning says for the ANOVA case is that F tests with their associated p (for probability) level of significance are not sufficient means for assessing results. Nevertheless, reviewers sometimes use only significance levels when comparing results from several studies (e.g., Eysenck, 1960; Bracht, 1970).

Other authors have treated significant F values as implying sizable differences (Guilford, 1956; Mendenhall, 1968). Guilford (1956, p. 275) described the ANOVA results of a study:

> The F ratio for machines is significant beyond the .01 point, leaving us with considerable confidence that the machine differences, as such, have a real bearing upon the difficulty of the task.

Strictly speaking, such a significant F could have resulted where the differences were trivial (in the practical sense). The following theorem proves that for any predetermined (small) number, a statistically significant F (for J = 2) or t ratio can be obtained, but such that the differences due to treatment are less than that predetermined number.

Theorem: For any $\varepsilon > 0$, and $0 < |\bar{X}_1 - \bar{X}_2| < \varepsilon$, there exists an N_0 such that if sample size $N > N_0$, then t is statistically significant for an ordinary t-test. (In layman's terminology: With a large enough sample size, statistical significance is obtainable no matter how trivial the difference in means is.)

Proof: Let $\varepsilon > 0$ be given with $|\bar{X}_1 - \bar{X}_2| < \varepsilon$. Without loss of generality, assume: $s_1 = s_2 = s$ (Homogeneity of variance satisfied), and $n_1 = n_2 = n$ (equal cell size).

$\therefore N = n_1 + n_2 = 2n$

Then $|t| = (|\bar{X}_1 - \bar{X}_2|)/(s(2/n)^{1/2})$

$= (n^{1/2}(|\bar{X}_1 - \bar{X}_2|)/(s(2)^{1/2})$

Require that $n^{1/2} > 10s(2^{1/2})/(|\bar{X}_1 - \bar{X}_2|)$

$\Rightarrow n > 200s^2/|\bar{X}_1 - \bar{X}_2|^2$

Therefore, if $N = 2n > 400s^2/|\bar{X}_1 - \bar{X}_2|^2$, then

$$|t| > \frac{10s2^{1/2}}{|\bar{X}_1 - \bar{X}_2|} \cdot \frac{|\bar{X}_1 - \bar{X}_2|}{s\sqrt{2}} = 10$$

t is statistically significant

Q. E. D.

Because of the reliance on statistical significance in current research methodology, a misleading picture appears in the literature. A classical example was described by Bakan (1966). Suppose H_0 (the null hypothesis) is true in the population. Accordingly if tests of significance are carried out by 100 independent researchers, those 95 (approximately) who do not get statistical significance probably will not bother to publish their findings. The five who do attain statistical significance will be more inclined to publish their findings and they will make Type I errors.

Part of the misinterpretation of p values is due to a misunderstanding of what the probabilities relate to. Camilleri (1962) defined three types of probability: (1) intrinsic probability between population variables (e.g. in a population of scores what is the probability of a score being greater than one population standard deviation above the population mean), (2) auxiliary probability between a sample and a population (e.g. maximum likelihood estimates), and (3) inductive probability relating to the probable validity of a hypothesis; that is, scientific inference. He asserted that significance tests have been used for assessing inductive probability when really they are more appropriate for auxiliary probability.

Morrison & Henkel (1969) argued persuasively that, statistically speaking, most research does not qualify from the standpoint of legitimate use of significance tests. They presented the following paradigm:

Type of Sampling Technique	Type of Population Sampled: Specified	Unspecified
Probability	A	B
Nonprobability	C	D

The only legitimate use of significance tests is with studies in Category A.

Several writers have criticized the all-or-none method involved with significance testing (Rozeboom, 1950; Bakan, 1966; Meehl, 1967). Science progresses by adjustments of degree of belief rather than firm decisions. Bakan feels that the tests have little if anything to contribute to scientific inference. He does agree with Rozeboom that the tests are appropriate for making null hypothesis decisions.

R. A. Fisher (1959, p. 44) issued a caution about the interpretation of significance levels:

> They (tests of significance) do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test....

Accepting the Null Hypothesis

Some writers have advocated the accepting of H_0 if a significant statistic is not observed (Walker & Lev, 1953; Guilford, 1956; Guenther, 1964; Kirk, 1968; Glass & Stanley, 1970). Yet statisticians often have warned against such practices unless the power (probability of rejecting H_0 when the alternative hypothesis is true) is known (Berkson, 1942; Peatman, 1963; Mendenhall, 1968). Cohen (1969), however, has shown that in typical psychological research, power of greater than .90 would require larger samples than are usually available.

To emphasize the inappropriateness of accepting H_0 without knowing the power, the proof of a simple theorem is presented which states that

for a given level of significance, there exist normal distributions such that the F or t statistic will not be significant, but the size of the effects will be larger than any predetermined number.

THEOREM: There exist distributions satisfying the ANOVA assumptions such that the null hypothesis is not rejected, but the means differ by more than any pre-given number. (In layman's terminology: If, upon a non-significant test statistic, you accept H_0, then you may be calling a huge difference a "zero difference.")

Proof: (2-sample case) Let $\epsilon > 0$ be given. Require that $|\bar{X}_1 - \bar{X}_2| > \epsilon$. Without loss of generality, assume $s_1 = s_2 = s$ and $n_1 = n_2 = n$.

Then $t = (\bar{X}_1 - \bar{X}_2)/s(2/n)^{\frac{1}{2}} = n^{\frac{1}{2}}(\bar{X}_1 - \bar{X}_2)/\sqrt{2}s$

Let $s = |\bar{X}_1 - \bar{X}_2| \cdot n^{\frac{1}{2}}$

Then $t = (n^{\frac{1}{2}}(\bar{X}_1 - \bar{X}_2)/(\sqrt{2}|\bar{X}_1 - \bar{X}_2| \cdot n^{\frac{1}{2}})$

$= 1/\sqrt{2}$

$= \pm .707$ (not significant)

Q. E. D.

This proof indicates that a researcher who accepts H_0 may be calling an essentially infinite difference a "zero difference." McNemar's (1962) suggestion of using three regions (acceptance, suspended judgment, and rejection), depending on the size of the p, does not overcome this objection.

Conventional Rejection Levels

The subservience to using conventional levels (e.g., .01 or .05) was criticized over 30 years ago along with the very phrasing of "test of significance." (Snedecor, 1942). Despite more current warnings about

the ready acceptance of conventional significance levels (McNemar, 1962; Winer, 1962; Slough, 1963; Skipper, Guenther, & Nass, 1967; Labovitz, 1968), the American Psychological Association Publication Manual (1967) advocates the use of asterisks to indicate the various conventional levels. DuBois (1968) has contended that conventional levels promote objectivity.

Rosenthal and Gaito reported a "cliff" effect where researchers showed a greatest loss of confidence between p=.05 and p=.10 (Rosenthal & Gaito, 1963). However, a subsequent replication did not find this effect (Beauchamp & May, 1964). Both studies noted that students and faculty in the field of psychological research expressed more confidence (degree of belief in research findings) in the same p values based on 100 than on 10 cases in the sample, despite the fact that this meant that the smaller sample usually exhibited a larger difference.

Estimating Sample Size

Much attention has been devoted to the estimation of sample size with regard to detecting differences as statistically significant. Winer (1962) and Cohen (1969) produced tables that are difficult to use. A simpler method was presented by Overall and Dalal (1968). Most of the writers in this area have empnasized sample size or power with regard to obtaining statistical significance rather than sharpening of estimates. For example, Cohen (1969) defined "power" as the probability that an investigation would lead to statistically significant results. Such a definition implies that the purpose of increasing power is to increase the probability of obtaining statistical significance. It also means that absurdities logically follow; for example, a larger sample is sometimes deemed less desirable than a smaller one without any

mention of cost effectiveness (Hays, 1963). If, however, power is defined in terms of sharpness of estimates rather than ability to detect differences as statistically significant, then such an anomaly does not arise, for in this case, the larger the sample, the better the estimate. Furthermore, if the statistical level of significance is related to the level of practical significance, then the importance of sample size is placed in proper perspective. Since a zero null hypothesis can always be rejected with a large enough sample under the ordinary ANOVA model (see page 4), the decision making depends more on N than on the estimation of the parameters. The noncentral ANOVA that is proposed in this paper will result in a closer relationship between estimation and decision making.

Confidence Intervals

The most commonly accepted method of treating practical significance statistically is the use of confidence intervals around linear combinations of means. These confidence intervals are described in most educational statistics books, but are offered more as options than as recommended and expected procedures (Guilford, 1956; McNemar, 1962). Nevertheless, several educational researchers are beginning to use confidence interval procedures--in particular, the post-hoc methods of Tukey and Scheffé (Scheffé, 1959). These post-hoc procedures control the Type I error (rejecting a null hypothesis when, in fact, it is true) for all contrasts, and as such are a definite improvement over the practice of computing several t-tests. However, since post-hoc contrasts are computed after a statistically significant F test without regard for practical significance, much effort may be spent on putting bounds around

trivial results. What is needed is an F-type test which is significant in the statistical sense only if the results are also of practical significance.

Another confidence interval approach has been advocated by Rosenkrantz (1972). The use of "direct confidence intervals" can promote the control of the probability of indecisive results and thus provide opportunity to study the weaknesses of the model under consideration.

Measures of Association

Another accepted means of attempting to relate practical significance with statistical significance is the use of measures of association like ω^2 (Hays, 1963), which represent the percent of variance explained (Nunnally, 1960; Duggan & Dean, 1968). These measures, unlike the p values associated with the F test, are relatively independent of sample size (Kennedy, 1970).

Summary of Related Research

Tests of statistical significance, as commonly used, have been frequently criticized. Suggested improvements have also met with criticism particularly because of the continued lack of relationship between statistical and practical significance. Well known writers have advocated inappropriate methodology. Confidence intervals and measures of association were seen as two ways of assessing practical significance. What is also needed is an F-type test that relates statistical and practical significance.

Noncentral Analysis of Variance

Related Research and Theory

ANOVA's peculiar characteristic of sometimes resulting in statistical but not practical significance leads to the following situation: a statistical rejection of the null hypothesis can coincide with (1) a practical difference, or (2) a trivial difference.

Some researchers regard levels of significance as indications of the degree of certainty in the results. This certainty, however, refers to the probability that the true difference is not exactly zero. It does not refer to the size of the difference. With a large enough sample, a difference can be miniscule and yet the p value could easily imply that it is very certain that the true difference is not exactly zero.

Consider also the following example:

$\bar{X}_1 = 7,\ n_1 = 2,\ \bar{X}_2 = 2,\ n_2 = 2,\ S_{pooled} = 1.$

Then $t = (\bar{X}_1 - \bar{X}_2)/S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}} = \frac{5}{(1)^{\frac{1}{2}}} = 5.$

If the low p value ($p < .05$) obtained is used as a measure of certainty, then this example shows a case where one would be "certain" about the results based on a sample of only four subjects.

The most commonly accepted solution to the statistical-practical significance dilemma seems to be one of first ascertaining statistical significance and second assessing the practical significance of any statistically significant results. In essence, the statistical test does not necessarily match up with the practical one.

Since practicality often is assessed post hoc (after the statistical test), it is reasonable to ask for an a priori (before the test)

assessment so that a more appropriate null hypothesis can be used. The unquestioning acceptance of always using a zero difference null hypothesis has been criticized by several writers (Grant, 1962; Kerlinger, 1964; Cohen, 1969).

For the two sample t-test, it has been suggested (Dixon & Massey, 1969; Pena, 1970) that if d represents a practical difference, then the test statistic is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

In this case the use of the ordinary t statistic would amount to asking the wrong question. Instead of asking whether there is a difference at all, researchers usually should be asking whether or not there is an educational or practical difference. Instead of asking whether a Datsun gets better mileage than a Cadillac, we should be asking how many more gallons a Datsun gets and whether this difference is of practical importance.

Using Dixon and Massey's model, if a researcher obtains a statistically significant difference then it will also be of practical significance (i.e., greater than the preassigned value of d). Basically this procedure results in the test of the appropriate (non-zero) null hypothesis.

As indicated earlier, analysis of variance needs a similar procedure since trivial differences may be statistically significant, and Tukey's or Scheffe's confidence interval procedures (Scheffé, 1959) would merely be putting bounds around trivial differences. Fortunately, the noncentral parameter δ of the noncentral F distribution provides an analog to the d used in the t statistic just described. Again if the minimum practical

difference is greater than zero, then use of the ordinary F test amounts to asking the wrong question.

Once a researcher has determined what constitutes a practical difference, then the next problem is to associate this difference with the appropriate noncentrality parameter. If this δ is correctly determined, then the new model guarantees that statistical significance will be related to practical significance. The influence of sample size on the F value is no longer a problem since as N increases, δ also increases in such a way that the critical F value is automatically adjusted upwards to compensate for the increase in the F statistic due to the larger sample size.

Estimating the Noncentrality Parameter Associated with a Practical Difference

Kirk (1968) defined the noncentrality parameter[1] as

$$\delta = \sqrt{\sum_{j=1}^{J} n_j(\mu_j - \mu)^2/\sigma_e^2} = \sqrt{\sum_{j=1}^{J} n_j \alpha_j^2/\sigma_e^2}$$

where J = number of treatments
n_j = number of subjects in the jth treatment
μ_j = mean of the jth treatment
μ = grand mean
σ_e^2 = error variance

The noncentrality parameter expresses the size of the effects in terms of the differences between the various group means and the grand mean. When one speaks of practical differences, he is usually referring to the differences between treatments rather than the difference between each treatment and the overall mean. Of course, if a researcher could relate what he considers a practical difference with the squared

[1] Tang's (1938) classic on power defined the noncentrality parameter as $\phi = \delta/\sqrt{J}$.

differences between the group means and the grand mean, then he could directly substitute into the formula for δ, and thus determine the non-centrality parameter associated with a practical difference.

Since, however, differences among treatments is the more common approach, these differences will be related to δ by the following theorem[2]:

[2] Gini (undated) proved a similar theorem for the relationship between $\sum_{j=1}^{J} \alpha_j^2$ and $\sum_{i<\kappa<J} (\mu_i - \mu_\kappa)^2$.

THEOREM: $J \sum_{i=1}^{J} \alpha_i^2 = \sum_{i<k\leq J} (\mu_i - \mu_k)^2$

where J = number of treatments, μ_i = mean of ith treatment,

$\alpha_i = \mu_i - \mu_{..}$, where $\mu_{..} = \left(\sum_{i=1}^{J} \mu_i \right) / J$.

Proof:

$$\sum_{i=1}^{J} \alpha_i^2 = (\mu_1 - \mu_{..})^2 + (\mu_2 - \mu_{..})^2 + \ldots + (\mu_J - \mu_{..})^2$$

$$= \left(\mu_1 - \frac{\mu_1 + \mu_2 + \ldots + \mu_J}{J}\right)^2 + \ldots + \left(\mu_J - \frac{\mu_1 + \mu_2 + \ldots + \mu_J}{J}\right)^2$$

$$= \sum_{i=1}^{J} \mu_i^2 - \frac{2}{J} \sum_{i=1}^{J} \left(\mu_i^2 + \sum_{i \neq j} \mu_i \mu_j\right) + J \left(\frac{\sum_{i=1}^{J} \mu_i^2 + 2 \sum_{i<j} \mu_i \mu_j}{J^2} \right)$$

$$= \left(1 - \frac{2}{J} + \frac{1}{J}\right) \sum_{i=1}^{J} \mu_i^2 - \frac{4}{J} (\mu_1\mu_2 + \ldots + \mu_{J-1}\mu_J) + \frac{2}{J} (\mu_1\mu_2 + \ldots + \mu_{J-1}\mu_J)$$

$$= \frac{J-1}{J} \sum_{i=1}^{J} \mu_i^2 - \frac{2}{J} \sum_{i<j} \mu_i \mu_j$$

$$\therefore J \sum_{i=1}^{J} \alpha_i^2 = (J-1) \sum_{i=1}^{J} \mu_i^2 - 2 \sum_{i<j} \mu_i \mu_j$$

$$= (\mu_1^2 - 2\mu_1\mu_2 + \mu_2^2) + (\mu_1^2 - 2\mu_1\mu_3 + \mu_3^2) + \ldots +$$

$$(\mu_2^2 - 2\mu_2\mu_3 + \mu_3^2) + \ldots + (\mu_{J-1}^2 - 2\mu_{J-1}\mu_J + \mu_J^2)$$

$$= (\mu_1 - \mu_2)^2 + (\mu_1 - \mu_3)^2 + \ldots + (\mu_2 - \mu_3)^2 + \ldots + (\mu_{J-1} - \mu_J)^2$$

$$= \sum_{i<k\leq J} (\mu_i - \mu_k)^2$$

Q. E. D.

The main purpose of this theorem is to enable the researcher to relate practical differences, which can be expressed in terms of $(\mu_j - \mu_k)$, to δ, which is a function of $(\mu_j - \mu..)^2$. It would be difficult if the researcher had to translate practical differences in terms of $(\mu_j - \mu..)$.

Given K treatments let M = the minimum practical average difference between treatments, expressed in terms of absolute values.

$$M = \frac{\sum_{i<k\leq J} |\mu_j - \mu_k|}{\binom{J}{2}}$$ average of pairwise differences (absolute values)

Consider the case where cell sample sizes are equal. Since $\delta^2 = \left(n \sum_{j=1}^{J} \alpha_j^2\right) / \sigma_e^2$ and the previous theorem showed that

$$\sum_{j=1}^{J} \alpha_j^2 = \sum_{i<k\leq J} \frac{(\mu_j - \mu_k)^2}{J} = \left[\sum_{i<k\leq J} \frac{(\mu_j - \mu_k)^2}{\binom{J}{2}}\right] \cdot \left(\frac{J-1}{2}\right)$$

then a reasonable trial substitution for $\sum_{j=1}^{J} \alpha_j^2$ is $M^2[(J-1)/2]$ The square of the average of the pairwise differences is to be substituted for the average of the (differences)^2. So we have

$$\delta^2 = \left(\sum_{j=1}^{J} n\alpha_j^2\right)/\sigma_e^2 \quad , \quad \hat{\delta}^2 = \frac{n M^2 \left(\frac{J-1}{2}\right)}{\sigma_e^2}$$

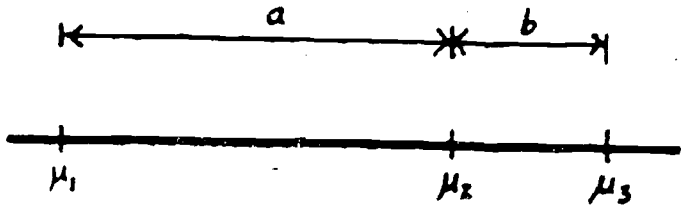
For J = 3, we have $\hat{\delta}^2 = \frac{n M^2 \left(\frac{3-1}{2}\right)}{\sigma_e^2} = \frac{n M^2}{\sigma_e^2}$

Besides using the minimum practical average M, it is also possible to substitute the individual minimum pairwise differences if these can be stated by the researcher. In addition, orthogonal a-priori contrasts may be performed with a δ being determined by the particular contrast. Post-hoc contrasts like Scheffé's can still be used as currently practiced since they control the Type I

error for all contrasts, independent of the truth of the null hypothesis (Scheffé, 1959).

Let $R = \frac{\delta^2}{\hat{\delta}^2}$ = a measure of how good an approximation results from using M^2 in place of $\sum_{j=1}^{J} \alpha_j^2$. By using several sample sizes, means, and variances, a computer program was used to compute several such ratios. R turns out to be a function of the relative distances between the means; for example, the lowest ratio (and therefore, the best estimate) was obtained when the means were equally distant (e.g., $\mu_1 = 7$, $\mu_2 = 10$, $\mu_3 = 13$). The worst estimate occurred when two means were as far away from the third means as possible (e.g., $\mu_1 = 0$, $\mu_2 = 15$, $\mu_3 = 15$). In between, the ratio was exactly determined by the variable $e = |a - b|/(a + b)$ where $a = |\mu_1 - \mu_2|$, $b = |\mu_2 - \mu_3|$ and $\mu_1 \leq \mu_2 \leq \mu_3$. Accordingly δ^2 can be readily determined by multiplying $\hat{\delta}^2$ by R.

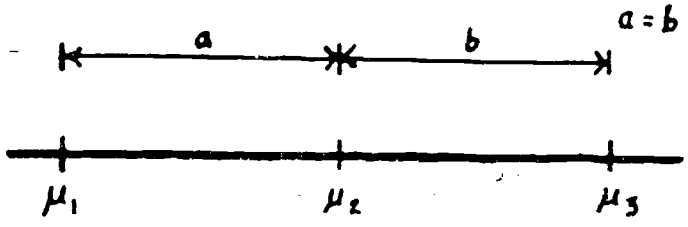
Figure 1: R in Terms of Position of Means



$e = |a-b|/(a+b)$

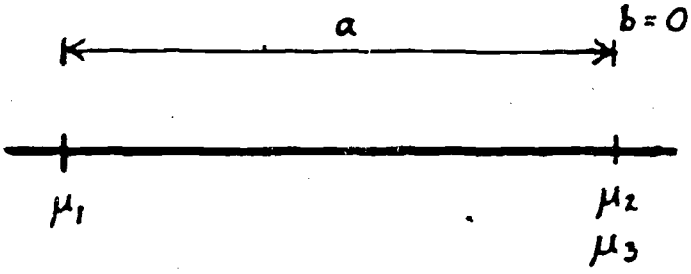
Special cases:

i) $e = 0 \Rightarrow R = 1.13$ (best estimate)



$e = |a-b|/(a+b)$
$= 0/(a+b) = 0$

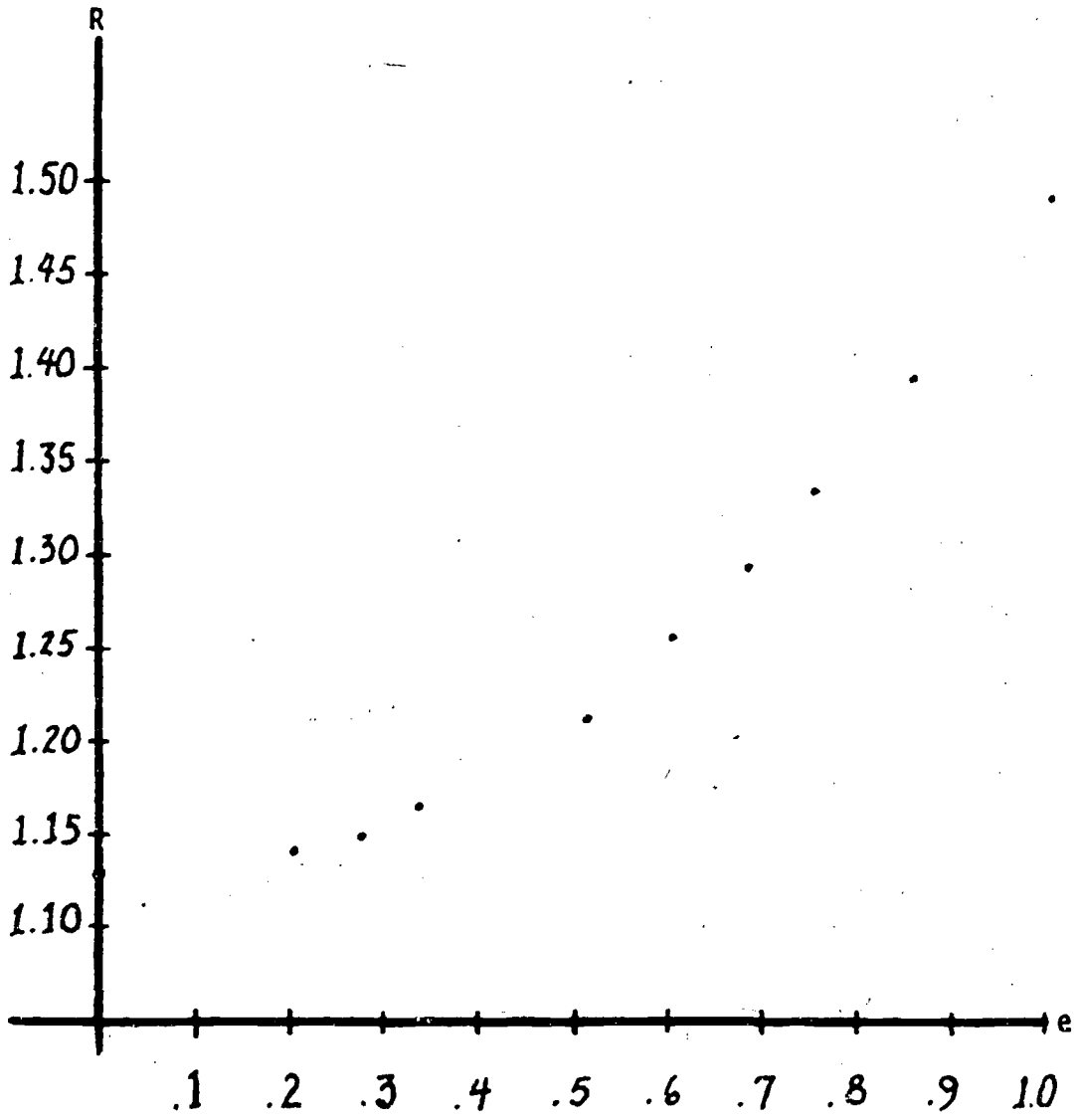
ii) $e = 1 \Rightarrow R = 1.5$ (worst estimate)



$e = |a-b|/(a+b)$
$= a/a$
$= 1$

Figure 2: $R = \delta^2/\hat{\delta}^2$ as a Function of

$e = |(\mu_2-\mu_1)-(\mu_3-\mu_2)|/(\mu_3-\mu_1)$



Performing the Noncentral F Test

Once the a-priori practical difference is used to determine δ_o, an ordinary F test is performed to test the hypotheses:

$H_0 : \delta^2 \leq \delta_o^2$ (there is no practical difference)

$H_1 : \delta^2 > \delta_o^2$ (there is a practical difference)

Instead of rejecting H_0 when the observed $F > F_{\nu_1, \nu_2}(1-\alpha)$, now H_0 is rejected when the observed $F > F'_{\nu_1, \nu_2, \delta_o}(1-\alpha)$, the noncentral F cutoff point.

Patnaik's Approximation of Noncentral F

The noncentral F distribution (denoted by F') has been tabled only partially (Johnson & Welch, 1939; Barton, David, & O'Neill, 1960; Severo & Zelen, 1960; Tiku, 1966). Unlike central F, F' cannot, in general, be expressed in closed form (Wishart, 1932; Price, 1964). A reasonably complete F' table would probably be too unwieldy for practical use (there are 389 pages in Resnikoff and Lieberman's Tables of the Non-central t-distribution, 1957). Of the several approximation procedures developed, Patnaik's (1949) seems the most usable since it utilizes the already available and familiar central F tables.

Although Patnaik's method involves laborious computation (Feldt & Malmoud, 1958; Grubbs, Coon, & Pearson, 1966), the resulting formulas for the fixed effects ANOVA case are relatively simple. The accuracy of Patnaik's approximation has been verified in several studies (Pearson, 1952; Tukey, 1957; Sankaran, 1963; Seber, 1963). A brief outline of the method appears in the Appendix to Scheffé's The Analysis of Variance (1959).

Derivation of Patnaik's Approximation for ANOVA

It can be shown that $E(\chi'^2_{\nu,\delta}) = \nu + \delta^2$ and variance $(\chi'^2_{\nu,\delta}) = 2\nu + 4\delta^2$ where $\chi'^2_{\nu,\delta}$ = noncentral χ^2 with noncentrality parameter $\delta = \sqrt{\sum_{j=1}^{J} n_j \alpha_j^2 / \sigma_e^2}$ (Scheffé, 1959)[1]

A possible approximation of $\chi'^2_{\nu,\delta}$ is $c\,\chi^2_{\tilde{\nu}}$.

Equating means and variances of the two distributions, we get

$c\tilde{\nu} = \nu + \delta^2$ (since $E(\chi^2_{\tilde{\nu}}) = \tilde{\nu}$) and $c^2(2\tilde{\nu}) = 2\nu + 4\delta^2$

$\Rightarrow c^2\tilde{\nu} = \nu + 2\delta^2$ since var $(\chi^2_{\tilde{\nu}}) = 2\tilde{\nu}$. Solve for c and $\tilde{\nu}$.

$$\tilde{\nu}(c^2 - c) = \delta^2$$

$$\tilde{\nu}(c^2 + c) = 2\nu + 3\delta^2$$

$$\delta^2(c^2 + c)/(c^2 - c) = 2\nu + 3\delta^2$$

$$\delta^2(c+1)/(c-1) = 2\nu + 3\delta^2$$

$$\delta^2(c+1) = (c-1)(2\nu + 3\delta^2)$$

$$c(\delta^2 - 2\nu - 3\delta^2) = -\delta^2 - 2\nu - 3\delta^2 \Rightarrow c = \frac{(-4\delta^2 - 2\nu)}{(-2\delta^2 - 2\nu)} = \frac{2\delta^2 + \nu}{\delta^2 + \nu}$$

$$\tilde{\nu} = \delta^2/c(c-1) = \delta^2(\delta^2+\nu)(\delta^2+\nu)/(2\delta^2+\nu)\delta^2$$

$$= (\delta^2+\nu)^2/(2\delta^2+\nu)$$

Noncentral F can be considered as $(U_1/\nu_1)/(U_2/\nu_2)$ where U_1 is $\chi'^2_{\nu_1,\delta}$ and U_2 is $\chi^2_{\nu_2}$

So $F'_{\nu_1,\nu_2,\delta} = \dfrac{\chi'^2_{\nu_1,\delta}/\nu_1}{\chi^2_{\nu_2}/\nu_2} = \dfrac{c\,\chi^2_{\tilde{\nu}}/\nu_1}{\chi^2_{\nu_2}/\nu_2} = c\,\nu_1^{-1}\tilde{\nu}\left(\dfrac{\chi^2_{\tilde{\nu}}/\tilde{\nu}}{\chi^2_{\nu_2}/\nu_2}\right) = c\,\nu_1^{-1}\tilde{\nu}\,F_{\tilde{\nu},\nu_2}$

$$c\tilde{\nu} = \frac{2\delta^2 + \nu_1}{\delta^2 + \nu_1} \cdot \frac{(\delta^2+\nu_1)^2}{(2\delta^2+\nu_1)} = \delta^2 + \nu_1 \Rightarrow F'_{\nu_1,\nu_2,\delta} \cong \frac{1}{\nu_1}(\delta^2+\nu_1)F_{\tilde{\nu},\nu_2}$$

where $$\tilde{\nu}_1 = \frac{[(J-1)(F-1) + (J-1)J]^2}{2(J-1)(F-1) + (J-1)} = \frac{F^2}{2F-1}$$

---

[1] Scheffé's problem IV.4 has an error in it. The expression $\Pr\{z \le (x-\delta)(1+\frac{x^2}{2\nu})^{-1}\}$ should read $\Pr\{z \le (x-\delta)(1+\frac{2x^2}{\nu})^{-\frac{1}{2}}\}$.

With the formulas for ν_1 and $F'_{\nu_1, \nu_2, \delta}$ a noncentral analysis of variance can now be performed. In the following illustrative example, notice that the observed F statistic would be significant for an ordinary ANOVA.

Illustrative Example

Noncentral ANOVA:

J=3 N=60 n=20 σ_e^2=25

1. Researcher states that the average difference between pairs of treatment must be greater than 10 in order for there to be a practical significance.

2. The sample means are 38.9, 51.6, 53.3; $\hat{F} = 47.8$

$a = |\bar{X}_1 - \bar{X}_2|$, $b = |\bar{X}_2 - \bar{X}_3| \Rightarrow a = 12.7$, $b = 1.7$

$e = |a-b|/(a+b) = 11.0/14.4 = .76 \Rightarrow R = 1.35$

3. $\hat{\delta}^2 = \frac{nM^2}{\sigma_e^2} = \frac{20(10^2)}{25} = 80$

$\therefore \delta^2 = R \cdot \hat{\delta}^2 = 1.35(80) = 108$

4. $\tilde{\nu}_1 = F^2/(2F-1) = 2284.8/94.6 = 24.2$

5. $F'_{\nu_1, \nu_2, \delta}(.95) = \frac{1}{\tilde{\nu}_1}(\delta^2 + \nu_1) F_{\tilde{\nu}_1, \nu_2}(.95)$

$= \frac{1}{2}(108+2) F_{24.2, 57}(.95)$

$= 55(1.71) = 93.5$

Since $\hat{F} < 93.5$, do not reject "H_0: There is no practical difference."

Monte Carlo Test of Noncentral ANOVA

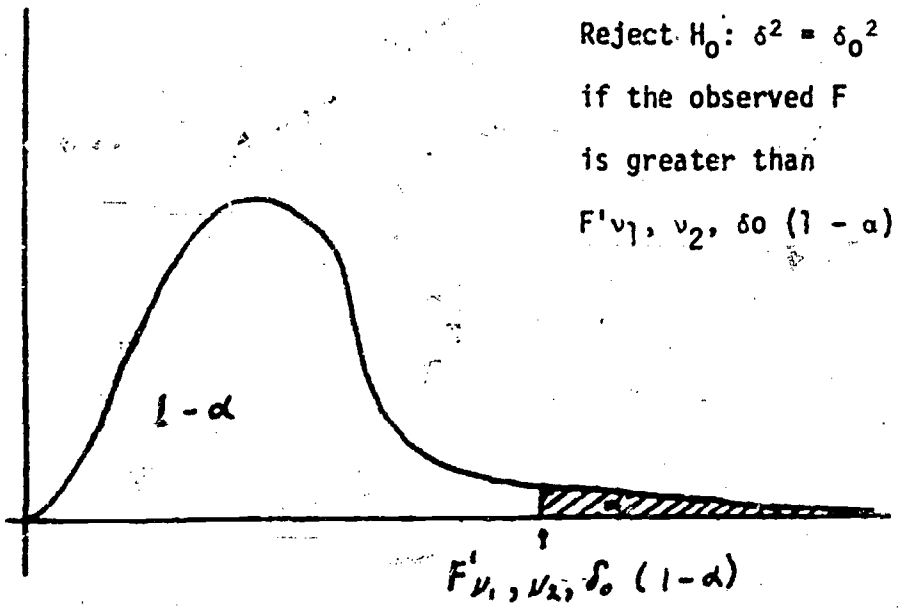
Rationale:

The CAL DEVIATE computer program (Hutchinson, 1967) was used to generate pseudorandom samples from three normal populations

with the following parameters: $\mu_1 = 40$, $\mu_2 = 50$, $\mu_3 = 55$; $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 25$; $n_1 = n_2 = n_3 = 20$. The Monte Carlo test for noncentral ANOVA is based on the following rationale: (1) Suppose a researcher has been able to postulate a minimum average practical difference among the three groups. A method described earlier relates this functionally to δ_0 where the researcher wants a statistically significant result to imply that the average differences among groups are such that $\delta > \delta_0$. (2) If the populations are set up such that the population $\delta = \delta_0$, then a verification of the model would require that for a Type I error rate of α, $100\alpha\%$ of the time the observed F would exceed the tabled or computed noncentral F. Figure 3 presents pictorially what is happening.

Figure 3

Rejection Region for Noncentral ANOVA



Notice that if the ordinary ANOVA were used, a statistically significant result would occur much more often than 100α% of the time even though the true differences are not sufficient to be of practical significance.

Procedure:

One hundred analyses of variance were run using the BMD01V program. The population is set up so that it barely misses meeting the criterion of having practical difference. In ANOVA hypothesis testing language:

H_0: $\delta^2 \leq \delta^2_0$ there is no practical difference

H_1: $\delta^2 > \delta^2_0$ there is a practical difference

where $\delta^2_{pop} = \delta_0^2$. Since $\bar{X}$ (sample mean) is an unbiased estimate of μ (population mean), the grand means of the three entire samples generated are used in calculating δ^2_{pop}. Each $\hat{\mu}_i$ represents the mean of all the data generated from the ith population.

$\hat{\mu}_1 = 39.95$, $\hat{\mu}_2 = 50.44$, $\hat{\mu}_3 = 54.99$

$\hat{\mu}.. = 48.46$ = grand mean of all the samples combined

$\alpha_1 = 6.53$, $\alpha_2 = 1.98$, $\alpha_3 = -8.51$ ($\alpha_j = \hat{\mu}_j - \hat{\mu}..$)

$\alpha_1^2 = 42.64$, $\alpha_2^2 = 3.92$, $\alpha_3^2 = 72.25$ $\Rightarrow \sum_{j=1}^{3} \alpha_j^2 = 118.81$

$$\delta^2 = \left(n \sum_{j=1}^{3} \alpha_j^2\right) / \sigma_e^2$$

$$= \frac{20\ (118.81)}{25}$$

$$= 94.84$$

The rejection point is $F'_{\nu_1, \nu_2, \delta}(1 - \alpha) = \hat{F} \cdot F_{\tilde{\nu}_1, \nu_2}(1-\alpha)$ where $\hat{F}$ = F observed. $F_{pop} = 1 + \delta^2/2$, $\tilde{\nu}_1 = \hat{F}^2/(2\hat{F} - 1)$.

These formulas were derived earlier in this paper.

$$F_p = 1 + \frac{94.84}{2} = 48.42, \quad \nu_i = \frac{(48.42)^2}{2(48.42) - 1} = 24.5$$

$$F'_{\nu_1, \nu_2, \delta}(1 - \alpha) = (48.42)\, F_{24.5,57}(1 - \alpha)$$

$$= 102.65 \text{ for } \alpha = .01$$

$$= 82.31 \text{ for } \alpha = .05$$

$$= 73.11 \text{ for } \alpha = .10$$

Since 100 samples were run, approximately 1 F value should exceed 102.65, about 5 should exceed 82.31, and about 10 should exceed 73.11.

Table 1 compares the expected with the actual number of F values exceeding the various cut-off points.

Table 1:

Summary of Monte Carlo Test of Noncentral ANOVA

α	Expected number exceeding cut off	Actual number exceeding cut off
.01	1	0
.05	5	8
.10	10	11

The misfit for $\alpha = .05$ is not as bad as it seems, since of the 8 exceeding 82.31, three were barely above that value (82.36, 82.56, and 82.89).

Summary

The Monte Carlo test, in general, verified that the non-central ANOVA procedure is operating at near the appropriate Type I error rate. Notice that an ordinary ANOVA procedure would

have yielded 100 statistically significant results where the population has imposed upon it the characteristic of no practical significance.

Summary and Discussion

Some of the common inappropriate uses of the traditional analysis of variance and also the shortcomings inherent in the ANOVA model itself have been described. The ANOVA model was modified to integrate practical significance with statistical significance. The modified version, based on the noncentral F distribution, included procedures for estimating the required noncentrality parameter δ, given that the researcher can state a priori what constitutes a minimum practical difference among the group means.

This proposed noncentral ANOVA would seem to be an improvement over ANOVA in several aspects:

1. No longer can trivial (in the educational sense) results attain statistical significance. Hence, the illogic of the concept of "too large a sample" does not exist apart from cost effectiveness.

2. The researcher is forced to relate numerical scores with practicality instead of analyzing scores in and of themselves.

3. Post-hoc contrasts (e.g., Scheffé, Tukey) are computed only around non-trivial (in the educational sense) results.

4. A statistical rejection can no longer be followed by two contradictory outcomes. In ordinary ANOVA, statistical significance can go with (1) no practical significance or (2) a practical difference. With noncentral ANOVA, statistical significance goes only with practical significance because the appropriate hypothesis is being tested.

If noncentral ANOVA becomes widely used, it would be desirable to have easily used noncentral F tables where a researcher need only

specify ν_1, ν_2, and δ to obtain the corresponding noncentral F value. The various partial tables now in existence are geared mainly for power calculations and not readily usable (e.g., Tang, 1938; Cohen, 1969).

The overall importance of the procedures presented in this study is the bringing together of statistics and practicality. This synergism enables not only more meaningful presentation of results, but also the powerful use of statistics in a complementary rather than ritualistic way.

The use of noncentral ANOVA can improve the quality of data analysis, while at the same time be straightforward enough for understanding and use by practitioners. E. S. Pearson (1938, p. 471) aptly described the importance of noncomplex concepts for users:

> If the object of the mathematical statistician is to provide tools for practical use, it seems important that the connexion between the abstract and the perceptual should be expressible in terms of the simplest possible probability concepts.

Noncentral ANOVA would seem to meet this criterion while at the same time provide a means of eliminating some of the crucial shortcomings of the currently used ANOVA model.

BIBLIOGRAPHY

American Psychological Association. Publication Manual. Washington, D.C.: APA, 1967.

Bakan, D. The test of significance in psychological research. Psychological Bulletin, 1966, 66, 423-437.

Barton, D. E., David, F. N., & O'Neill, A. F. Some properties of the distribution of the logarithm of non-central F. Biometrika, 1960, 47, 417-431.

Beauchamp, K. L., & May, R. B. Replication report: Interpretation of levels of significance by psychological researchers. Psychological Reports, 1964, 14(1), 272.

Berkson, J. Tests of significance considered as evidence. Journal of the American Statistical Association, 1942, 37, 325-335.

Bracht, J. Experimental factors related to aptitude-treatment interaction. Review of Educational Research, 1970, 40(5), 627-645.

Camilleri, S. F. Theory, probability, and induction in social research. American Sociological Review, 1962, 27, 170-178.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, 1969.

Dixon, W., & Massey, F., Jr. Introduction to Statistical Analysis. New York: McGraw-Hill, 1969.

DuBois, P. H. An Introduction to Psychological Statistics. New York: Harper & Row, 1965.

Duggan, T. J., & Dean, C. W. Common misinterpretations of significance levels in sociology journals. The American Sociologist, 1968, 3, 45-46.

Eysenck, H. J. The concept of statistical significance and the controversy about one-tailed tests. Psychological Review, 1960, 67(4), 269-271.

Feldt, L. S., & Manmoud, M. W. Power function charts for specifications of sample size in analysis of variance. Psychometrika, 1958, 23(3), 201-210.

Fisher, R. A. Statistical Methods and Scientific Inference. New York: Hafner, 1959.

Fleiss, J. L. Estimating the magnitude of experimental effects. Psychological Bulletin, 1969, 72, 273-276.

Gini, C. W. Variabilita e mutabilita, contributo allo studio delle distribuzioni e relazioni statische studi Economico-Gioridici della R. Universita di Cagliari, undated.

Glass, G., & Hakstian, A. Measures of association in comparative experiments: Their development and interpretation. American Educational Research Journal, 1969, 6, 403-414.

Glass G. V., & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Grant, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, 1962, 69, 54-61.

Grubbs, F. E., Coon, H. J., & Pearson, E. S. On the use of Patnaik type chi approximations to the range in significance tests. Biometrika, 1966, 53, 248-252.

Guenther, W. C. Analysis of Variance. Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1956.

Hays, W. Statistics. New York: Holt, Rinehart, & Winston, 1963.

Heermann, E. F., & Braskamp, L. A. Readings in Statistics for the Behavioral Sciences. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Horst, P. Psychological Measurement and Prediction. Belmont, California: Wadsworth, 1966.

Hutchinson, D. CAL DEVIATE. Computer Center, University of California, Berkeley, 1967.

Johnson, N. L., & Welch, B. L. Applications of the non-central t-distribution. Biometrika, 1939, 31, 362-389.

Kempthorne, O., & Doerfler, T. E. The behavior of some significance tests under experimental randomization. Biometrika, 1969, 56, 231-247.

Kennedy, J. J. The eta coefficient in complex ANOVA designs. Educational and Psychological Measurement, 1970, 30, 885-889.

Kerlinger, F. Foundations of Behavioral Research. New York: Holt, Rinehart, & Winston, 1964.

Kirk, R. Experimental Design: Procedures for the Behavioral Sciences. Belmont, California: Brooke/Cole, 1968.

Labovitz, S. Criteria for selecting a signif cance level: A note on the sacredness of .05. The American Sociologist, 1968, 3, 220-222.

Lykken, D. Statistical significance in psychological research. Psychological Bulletin, 1968, 70, 151-159.

McNemar, Q. Psychological statistics. New York: John Wiley & Sons, 1962.

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

Mendenhall, W. Introduction to linear models and the design and analysis of experiments. Belmont, California: Wadsworth, 1968.

Morrison, D. E., & Henkel, R. E. Significance tests reconsidered. The American Sociologist, 1969, 4, 131-140.

Morrison, D. W., & Henkel, R. E. The significance test controversy. Chicago: Aldine, 1970.

Nunnally, J. The place of statistics in psychology. Educational and Psychological Measurement, 1960, 20, 641-650.

Overall, J. Classical statistical hypothesis testing within the context of Bayesian theory. Psychological Bulletin, 1969, 71, 285-292.

Overall, J., & Dalal, S. N. Empirical formulae for estimating appropriate sample sizes for analysis of variance designs. Perceptual and Motor Skills, 1968, 27(2), 363-367.

Patnaik, P. B. The noncentral X^2 and F-distributions and their approximations. Biometrika, 1949, 36, 202-232.

Pearson, E. S. Comparison of two approximations to the distribution of the range in small samples from normal populations. Biometrika, 1952, 39, 130-136.

Pearson, E. S. Note on Professor Pitman's contribution to the theory of estimation. Biometrika, 1938, 30, 471-474.

Peatman, J. Introduction to applied statistics. New York: Harper & Row, 1963.

Pena, D. A significant difference of opinion with the Coats position. Educational Researcher, 1970, 11, 9-10.

Price, R. Some non-central F distributions expressed in closed form. Biometrika, 1964, 51, 107-122.

Resnikoff, G. J., & Lieberman, G. J. Tables of the non-central t-distribution. Stanford: Stanford University Press, 1957.

Rosenkrantz, R. The significance test controversy. Educational Researcher, 1972, 1(12), 10-14.

Rosenthal, R., & Gaito, J. The interpretation of levels of significance by psychological researchers. Journal of Psychology, 1963, 55(1) 33-38.

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

Sankaran, M. Approximations to the non-central chi-square distribution. Biometrika, 1963, 50, 199-204.

Scheffé, H. The analysis of variance. New York: John Wiley & Sons, 1959.

Seber, G. The non-central chi-squared and beta distributions. Biometrika, 1963, 50, 542-544.

Selvin, H. C. A critique of tests of significance in survey research. American Sociological Review, 1957, 22, 519-527.

Severo, N. C., & Zelen, M. Normal approximation to the chi-square and non-central F probability function. Biometrika, 1960, 47, 411-416.

Skipper, J. K., Guenther, A. C., & Nass, G. The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. The American Sociologist, 1967, 1, 16-18.

Slough, D. A. Experimental precision and tests of hypotheses. Psychological Record, 1963, 13(2), 221-226.

Snedecor, G. W. The use of tests of significance in an agricultural experiment station. Journal of the American Statistical Association, 1942, 37, 383-386.

Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance--or vice versa. Journal of the American Statistical Association, 1959, 54, 30-34.

Tang, P. C. The power function of the analysis of variance tests with tables and illustrations of their use. In J. Neyman, & E. S. Pearson (Eds.), Statistical Research Memoirs, Vol. II. London: University of London, 1938.

Tiku, M. L. A note on approximating to the non-central F distribution. Biometrika, 1966, 53, 606-610.

Tukey, J. W. Approximations to the upper 5 percent points of Fisher's β distribution and non-central χ^2. Biometrika, 1957, 44, 528-530.

Walker, D. F., & Schaffarzick, J. Comparing curricula. Review of Educational Resarch, 1974, 44(1), 83-111.

Walker, H. M., & Lev, J. Statistical inference. New York: Henry Holt, 1953.

Winch, R. F., & Campbell, D. T. Proof? No. Evidence? Yes. The significance of tests of significance. The American Sociologist, 1969, 4, 140-143.

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

Wishart, J. A. A note on the distribution of the correlation ratio. Biometrika, 1932, 24, 441-456.