CONSEQUENCES OF FAILURE TO MEET ASSUMPTIONS UNDERLYING THE FIXED EFFECTS ANALYSES OF VARIANCE AND COVARIANCE

Gene V Glass

Laboratory of Educational Research
University of Colorado

Percy D. Peckham

University of Washington

James R. Sanders

Indiana University

The effects of violating the assumptions underlying the fixed-effects analyses of variance (ANOVA) and covariance (ANCOVA) on Type-I and Type-II error rates have been of great concern to researchers and statisticians. The major effects of violation of assumptions are now well known, after nearly four decades of research. Early summaries and reviews by Hey (1938), Garret and Zubin (1943), Grant (1944), and Gourlay (1955) and more recent reviews by Bradley (1963), Atiqullah (1967), Elashoff (1969) and Scheffé (1959, Ch. 10) can be extended and updated with recent research which provides closure to an area of active inquiry. (For a review of the effects of violation of the assumptions of the random-effects ANOVA—a subject not reviewed here—the reader is directed to Scheffé, 1959, pp. 334-337 and Box & Anderson, 1962.) Asking whether ANOVA and ANCOVA assumptions are satisfied is not idle curiosity. The assumptions of most mathematical models are always false to a greater or lesser extent. The relevant question is not whether ANOVA assumptions are met exactly, but rather whether the plausible violations of the assumptions have serious consequences on the validity of probability statements based on the standard assumptions. Applied statistics in education and the social sciences experienced a largely unnecessary hegira to non-parametric statistics during the 1950s. Increasingly during the 1950s and early 1960s the fixed-effects, normal theory ANOVA was replaced by such comparable nonparametric techniques as the Wilcoxon test, Mann-Whitney U-test, Kruskal-Wallis one-way ANOVA, and the Friedman two-way ANOVA for ranks (Siegel, 1956). The flight to non-parametrics was unnecessary principally because researchers asked "Are normal theory ANOVA assumptions met?" instead of "How important are the inevitable violations of normal theory ANOVA assumptions?"

The problem of what happens to levels of significance and power when the assumptions underlying an ANOVA or ANCOVA model are violated presents considerable difficulty to the mathematical statistician. In one sense, this is to be expected. Two criteria by which mathematicians select assumptions for their procedures are *credibility* and *manageability*. A credible assumption is one that is likely to be met by actual data, for example, a behavioral scientist would have little use for a procedure or model that depended heavily on the assumption that a variable had a rectangular distribution (all scores equally likely). A manageable assumption is one which simplified many mathematical derivations and operations. The widespread assumption of normality of observations on variables is perhaps the best example of an assumption that is credible, in many instances, and manageable. Many groups of observations made in the social sciences have one mode, a large proportion of central scores, and very few scores deviating greatly from the central scores; the fact that the mean and variance of samples from a normal distribution are statistically independent simplifies much of the statistical theory which rests upon the normality assumption.

Not surprisingly, then, things quickly become complicated when one inquires about the effects of violation of an assumption that has become important because it is credible and manageable. To answer such inquiries, one must necessarily make other assumptions that—although possibly more credible—are certainly less manageable (mathematically). " . . . we realize that standards of rigor possible in deducing a mathematical theory from certain assumptions generally cannot be maintained in deriving the consequences of departures from these assumptions" (Scheffé, 1959, p. 331).

Thus, easy mathematical generalizations will be scarce on the following pages, and numerous tables reporting specific (that is, non-general) results must be relied upon. Because of the breakdown of most mathematical attempts to deal with problems of assumptions violation, many of the findings are arrived at through computer simulation of inferential tests. Unlike mathematical argument which can be easily checked for validity, the validity of most computer simulations is difficult to check. At the conclusion of this paper guidelines for reporting computer simulation studies in this area of research are presented.

The paradigm for past research on violation of ANOVA and ANCOVA assumptions has conformed in varying degrees to the following:

For Level of Significance

1. Given a true null hypothesis, the $1 - \alpha$ percentile point in the F-distribution with $J - 1$ and $N - J$ d.f.$({}_{1-\alpha}F_{J-1, N-J})$ is found. (This percentile point will be the value exceeded by $(100\alpha)\%$ of the F-ratios obtained in an ANOVA when the null hypothesis is true *and the ANOVA assumptions are met.*)

2. By empirical or mathematical means, the actual percent of F-ratios exceeding ${}_{1-\alpha}F_{J-1,N-J}$ is found when the null hypothesis is true and the variances are heterogeneous or the populations are non-normal or both.
3. The *nominal** significance level, α, and the *actual* significance level, the percent of F's exceeding ${}_{1-\alpha}F_{J-1,N-J}$, are compared.

For Power

1. Given a *particular false* null hypothesis, the probability is determined of rejecting the null hypothesis when the ANOVA assumptions are met. This probability is the proportion of F-ratios exceeding ${}_{1-\alpha}F_{J-1,N-J}$ for a given false null hypothesis; it is called the "nominal power."
2. By empirical or mathematical means, the actual percent of F-ratios exceeding ${}_{1-\alpha}F_{J-1,N-J}$ is found for a given false null hypothesis when the population distributions are non-normal or have unequal variances.
3. The *nominal* power, found in step no. 1, is compared with the *actual* power, found in step no. 2.

When empirical, simulation techniques are used, actual levels of significance or power contain sampling error; the comparison of theoretical and actual levels must take this error into account. A practical solution to this problem is to evaluate the standard error of a proportion for samples of a size equal to the number of replications taken in the simulation, and to take this inferential instability into account in any "theoretical versus actual" comparison. Suppose that the distribution of a particular test statistic is insensitive to an assumption violation built into a simulation. The investigator performs n replications; and counts the proportion p of test statistics exceeding the $100(1-P)$ percentile in the appropriate probability distribution. The standard error of p, the empirically determined actual significance level, is known to be $\sqrt{(1-P)P/n}$; thus it would be unwise to attach much importance to a theoretical versus actual significance level or power difference which is less than about two standard errors of p. Since $P = .95$ and $.99$ are frequently of interest for checking the validity of Type I error rates, the graphs of the standard error of p for $P = .95$ and $.99$ and various values of n are given in Figure 1 for ready reference in interpreting data presented later in this paper. Note, however, that the standard error of p increases as P tends toward .50; thus we should be more tolerant of a given deviation from a theoretical probability that occurs near the center of the probability scale (for example, for power

*Several significant terms are defined in the Glossary at the end of this paper.

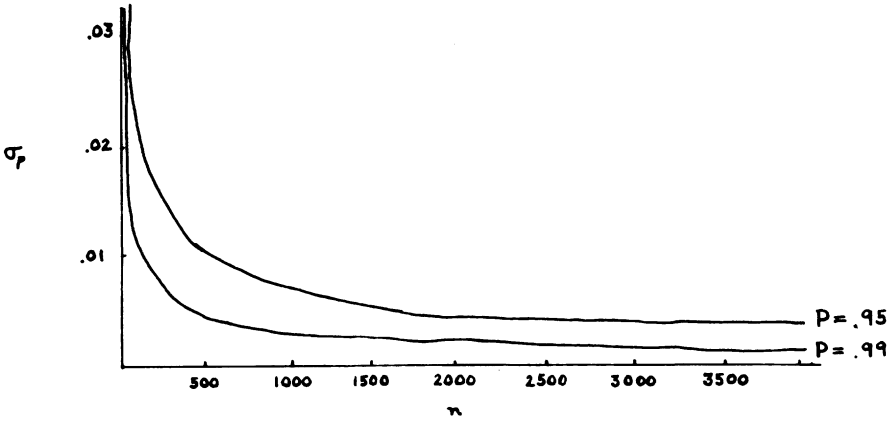


FIG. 1: Standard error of p for P = .95 and .99 and various values of n.

values near .50) than one that occurs near the extremes (for example, for levels of significance .05 and .01).

Fixed-effects ANOVA Model

The following assumptions are made in a simple one-way fixed-effects model ANOVA:

$$y_{ij} = \mu + \alpha_j + e_{ij} \tag{1}$$

$$e_{ij} \sim NID(0, \sigma^2) \tag{2}$$

$$\Sigma \alpha_j = 0 \tag{3}$$

The first assumption is that of *additivity*. An observation may be thought of as the simple sum of three components: μ, the common location parameter; α_j, the incremental or decremental effect of treatment j on the dependent variable for all observations in group j; and e_{ij}, the error of the (i,j)th observation. One must be aware of differing definitions of the term "additivity." There is an unfortunate ambiguity in the word "additivity" in statistical literature. In two and higher-way fixed-effects layouts, "additivity" often refers to the absence of any interaction. A two-way ANOVA model is spoken of as being "additive" when $E_k(y_{ijk}) = \mu + \alpha_i + \beta_j$, that is, the expected score in the (i,j)th cell is the simple sum of main effects and no interaction terms. This usage is not particularly descriptive, even though it is quite common; it seems sufficient and clear merely to speak of "having no interaction" in such cases. The second meaning of "additivity" when applied to an ANOVA model is used to distinguish models involving only sums of effects (either main or interaction) from models involving

non-additive functions of effects. For example, the functional relationship $s = \frac{1}{2}gt^2$ is a multiplicative (and non-stochastic, incidentally) relationship, and no additive model can be expected to fit very well a relationship between distance, time, and gravitational force.

The treatment and error components in a model for a comparative experiment could conceivably combine multiplicatively. However, if this case should arise, log transformations will yield an additive model. Tukey (1949) presented a procedure for testing for non-additivity in an ANOVA situation. (Scheffé, 1959, provided a unique interpretation of Tukey's procedure.) He suggested consideration of a transformation of the data followed by a new analysis of the transformed variable if the occurrence of a large non-additivity mean square is found. Cochran (1947) indicated that the principal effect of non-additivity is loss of information and that this will be trivial unless the error variance is very low or there is a very serious departure from additivity. We may conclude that, in reality, the violation of the additivity assumption should be of little concern for the researcher. For more material on this same question see Cox (1959, pp. 14-22) and Schlesselman (1971).

The third assumption is of no concern; it is not altogether necessary since it is merely a consequence of choosing to express y_{ij} in three terms (μ, α_j, e_{ij}) instead of two ($\beta_j = \mu + \alpha_j$ and e_{ij}, for example). In fact, the condition $\Sigma\alpha_j = 0$ is more properly regarded as a *restriction* adopted to achieve a unique solution to the least-squares normal equations rather than an *assumption* about the fit between reality and a mathematical model.

The second assumption states that the e_{ij}s have a normal distribution with a population mean (expectation) of 0 and variance of σ^2, and they are independent. We can consider three distinct violations of this assumption: (a) non-normality, (b) different variances from group to group, (c) nonindependence. It is a consequence of (a) — (c) that a huge number of observations taken under the J treatments should have nearly a normal distribution within each group and the variance of this distribution from one group to the next should be the same, σ^2. Furthermore, if repeated samples are drawn and a scatter diagram of the sample means for any one of the $J(J-1)/2$ pairs of treatments is constructed, it will show zero correlation of the means (across samplings of J means) because of the independence assumption.

For some non-normal distributions a transformation can be found which brings the data more closely in line with the normal distribution. The use of such "normalizing" transformations has often been discussed as a means of more nearly satisfying the ANOVA assumptions (Bartlett, 1947; Gaito, 1959; Curtiss, 1943; Mueller, 1949; Box & Cox, 1964; Nelder, 1964; Schlesselman, 1971). For reasons which will become apparent later in this paper, the payoff of normalizing transformations in terms of more valid probability statements is low, and they are seldom considered to be worth the effort (see Games & Lucas, 1966). This general conclusion needs to be qualified: trans-

formations—particularly of frequency of elapsed time data are useful in equalizing group variances in cases of unequal n's, the need for which will be discussed later; secondly, data transformations may be useful for purposes unrelated to inferential robustness, e.g., removal of interactions. We shall not discuss normalizing transformations further.

Failure to satisfy the independence assumption can be serious. The correlated or dependent-groups t-test is an appropriate statistical technique (if only two treatments are being compared) when nonindependence of the e_{ij}s exists. In the behavioral sciences, one often speaks of the problem of "repeated measures" (that is, observing the same persons under more than one treatment) when working under conditions of non-independent samples. The problem is identical to problems in analysis of Scheffé's mixed-effects model ANOVA (see Scheffé, 1959, Chap. 8). For reviews of the problem of correlated observations in the ANOVA see Greenhouse and Geisser (1959) and Lana and Lubin (1963). Collier, Baker, Mandeville and Hayes (1967) performed an extensive simulation study of the effects of heterogeneous variances and covariances in the mixed-effects model. Mandeville (1969) investigated the application of the mixed-model analysis to dichotomous data.

Nonindependence of errors can have serious effects on the validity of probability statements in the t-test or ANOVA. As Cochran (1947) reported, positive correlations will yield a more liberal test, while negative correlations will result in a more conservative test. The higher the correlations, the more deviant the actual significance level from the nominal level. Scheffé (1959) demonstrated that the effect of serial correlation on inferences about means can be serious. The data are presented in Table 1. We present these data first to emphasize that the violation of the independence assumption, which we shall not discuss further in this paper, is far more serious than the violation of the assumptions which we will discuss. Table 1 is a pointed reminder of the more critical side of the problem of violation of ANOVA assumptions.

Heterogeneous Variances

First we shall look at some effects of heterogeneous variances on the level of significance of the F-test. Hsu (1938) was one of the first

TABLE 1*

Effect of Serial Correlation on True Probability of Nominal 95 Percent Confidence Interval for μ Not Covering True μ for Large n

ρ	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
Probability of interval not capturing μ	1.10^{-5}	0.002	0.011	0.028	0.050	0.074	0.098	0.120	0.140

* Scheffé (1959, p. 339).

statisticians to obtain concise mathematical results in this area. (His work is reprinted in the more readily obtainable reference Scheffé, 1959, p. 353.*) Hsu determined the actual probability of a significant result at the .05 level for various values of the ratio of σ_1^2 to σ_2^2 in a two-tailed *t*-test. Hsu's data are reported in Table 2 along with more recent results by Scheffé (1959) and Pratt (1964) to the same problem. Scheffé's findings concern the effect of heterogeneous variances on α for large samples in the two-group case.

As an example of the interpretation of Table 2, consider the entry in the first row under 0.2. This entry, .178, is the probability of obtaining a significant *t*-ratio when the null hypothesis is true, the nominal α is .05, $n_1 = 15$ and $n_2 = 5$, and $\sigma_1^2 = .2\sigma_2^2$. Thus, one is nearly three and one-half times more (.18 = 3.6 · .05) likely to commit a Type-I error than he supposes when $n_1 = 15$, $n_2 = 5$, and $\sigma_1^2 = .2\,\sigma_2^2$.

One of the most striking features of Table 2 is that when $n_1 = n_2 = 7$, the nominal α (.05) and the actual probability of rejecting the null hypothesis are almost exactly the same. In the column headed "1" in Table 2, $\sigma_1^2 = \sigma_2^2$, i.e., the variances are homogeneous and no

*An error was made in Scheffé's construction of Table 10.4.1 (p. 353 in Scheffé, 1959) based on Hsu's data. Three figures appear in the wrong columns in the last row of the table. Table 2 in the present paper is the correct transcription of Hsu's table from the original source (and incorporates the correction of a typographical error which appears in row 1, column 5 of Hsu's original Table I).

TABLE 2*

Actual Probability of a Type-I Error with a Two-tailed t-test for Various Sample Sizes and Ratios of Sample Sizes and Values of the Population Variances When the Nominal Significance Level Is .05

		σ_1^2/σ_2^2								
n_1	n_2	*0*	*0.1*	*0.2*	*0.5*	*1*	*2*	*5*	*10*	∞
15	5	.317	.230	.178	.098	.05	.025	.008	.005	.002
5	3	.216	.145	.103	.072	.05	.038	.031	.030	.031
7	7	.072	.063	.058	.051	.05	.051	.058	.063	.072
n_1/n_2										
1		.05		.05	.05	.05	.05	.05		.05
1.5**		.109			.081	.05	.027			.016
2		.17		.12	.08	.05	.029	.014		.006
5		.38		.22	.12	.05	.014	.002		.00001
∞		1.00		.38	.17	.05	.006	.00001		0

* Entries in top half of table are due to Hsu; remaining entries due to Scheffé, except as otherwise indicated.

** From Pratt (1964).

assumption is violated. One sees that under these circumstances the nominal and actual significance levels are equal, as would be expected.

Box (1954a, 1954b) obtained some of the early mathematical results on the problem of the effect on α of heterogeneous variances in the one-way ANOVA. Many of these results appear in Scheffé (1959, p. 354). They are reprinted here as Table 3.

The following is an example of how Table 3 is read: if 3 treatments are compared and $n_1 = 9$, $n_2 = 5$, and $n_3 = 1$, and if the population variances are in the ratio 1:1:3 (e.g., $\sigma_1^2 = 10$, $\sigma_2^2 = 10$, $\sigma_3^2 = 30$), then the probability of a Type-I error is actually .17 when the experimenter thinks it is .05.

Box's results evidence the same trend noticed in Hsu's data (Table 2): when n's are equal, the *actual* and the *nominal* level of significance agree quite closely. It is noteworthy that Box found an actual significance level of .12 under the conditions in the last line of Table 3. This rather serious distortion in the nominal probability of a Type-I error could be due to the small n (3), the large number of groups (7), or the strange pattern and extreme violation of the homogeneous variances assumption. Whatever the cause, we find it significant to note that subsequent investigators have not extended Box's work in the direction of this curious finding. The conventional conclusion that heterogeneous variances are not important when n's are equal seems to have boundary

TABLE 3

*Effect of Heterogeneous Variances on the Probability of a Type-I Error in the One-way ANOVA for a Nominal Significance Level of .05**

No. of Groups, J	*Ratio of σ^2's*	*Sample Sizes, n*	*Actual Probability of Type-I Error*
3	1:2:3	5,5,5	.056
	(For example, $\sigma_1^2 = 10$,	3,9,3	.056
	$\sigma_2^2 = 20$, $\sigma_3^2 = 30$.)	7,5,3	.092
		3,5,7	.040
3	1:1:3	5,5,5	.059
	(For example, $\sigma_1^2 = \sigma_2^2 = 10$,	7,5,3	.11
	$\sigma_3^2 = 30$.)	9,5,1	.17
		1,5,9	.013
5	1:1:1:1:3	5,5,5,5,5	.074
	(For example, $\sigma_1^2 = \sigma_2^2 =$	9,5,5,5,1	.14
	$\sigma_3^2 = \sigma_4^2 = 10$, $\sigma_5^2 = 30$.)	1,5,5,5,9	.025
7	1:1:. . . :1:7	3,3, . . .,3	.12

* Data due to Box (1954a).

conditions like all other conclusions in this area, and the boundary conditions may not have been sufficiently probed.

There exist further data which corroborate the trends indicated in Tables 2 and 3. Some of the earliest empirical work in this area was performed by D. W. Norton; his findings will be reported in the section dealing with heterogeneous variances and non-normality. Later work which duplicated much previous research in many respects was reported by Boneau (1960). Young and Veldman (1963) reported empirical results quite like those contained in the above tables. They found that for $n_1 = n_2 = 10$ and $\sigma_1^2/\sigma_2^2 = 219/69$, 252/36, 279/9, and 287/1, the actual probabilities of a Type-I error when one is working at the nominal .05 level are approximately .050, .054, .066, and .062, respectively.

One of the rare mathematical findings on the question of heterogeneous variances is due to van der Vaart (1961). He showed that given large sample sizes and a true null hypothesis, the t-statistic for the independent groups test of difference between means is distributed as zd, where z is the unit normal variable and d is given by

$$d = \left[\frac{\left(\frac{n_1}{n_1 + n_2}\right) + \left(1 - \frac{n_1}{n_1 + n_2}\right)\frac{\sigma_2^2}{\sigma_1^2}}{1 - \left(\frac{n_1}{n_1 + n_2}\right) + \left(\frac{n_1}{n_1 + n_2}\right)\frac{\sigma_2^2}{\sigma_1^2}} \right]$$

By substituting various values of sample sizes and variances into the formula for d, one can obtain some idea of the effect of heterogeneous variances and its interaction with sample sizes. For example, when $n_1 = n_2$, $d = 1$ regardless of the values of the variances. The behavior of d is in accord with the empirical findings above.

The following general conclusions seem justified:

1. When n's are unequal and variances are heterogeneous, the actual significance level may greatly *exceed* the nominal significance level *when samples with smaller n's come from populations with larger variances.*
2. When n's are unequal and variances are heterogeneous, the actual significance level may be greatly *exceeded by* the nominal significance level *when samples with smaller n's come from populations with smaller variances.*

The problem of testing $\mu_1 = \mu_2$ when variances are heterogeneous is a classic one in the history of statistics. Behrens (1929) first addressed the problem and presented a solution which was extended by Fisher (1935) to what has come to be called the Behrens-Fisher solution. The Behrens-Fisher solution created controversy because of its "fiducial probability" rationale (see Pearson & Hartley, 1966). Welch (1947) derived an approximate test of $\mu_1 = \mu_2$ under $\sigma_1^2 \neq \sigma_2^2$ by

adjusting the degrees of freedom of Student's t-distribution so that it approximates the exact sampling distribution in question. Welch's technique is presented in several textbooks; and, as the results above suggest, it should be used when heterogeneous variances are suspected and n_1 and n_2 differ. For an empirical study of various methods of coping with unequal variances and unequal n's in two-group t-test, see Kohr (1970). Welch (1951) also suggested a method for correcting the F-statistic in the fixed-effects ANOVA when variances and n's are not equal.

Non-normality

It has often been reported that violation of the normality assumption should be of little concern. Rider (1929) and Pearson (1929, 1931) found little effect of non-normality on the two-tailed t- and F-tests, respectively, provided that the degrees of freedom for residual variance are not too small. Cochran (1947) indicated that the consensus of studies up to that time was that no serious errors were introduced by non-normality in the significance levels of the F-test or the two-tailed t-test. Hack (1958) supplied more raw data to demonstrate just how deviant the populations of random samples must be from normality before the F-test is affected. Scheffé (1959) pointed out that β_2, the kurtosis, and, to a lesser degree, β_1, the skewness, are the most important indicators of the extent to which non-normality affects the usual inferences made in the ANOVA. He showed that for the fixed-effects model, the distribution of t is independent of the form of the population for large n and hence the inferences about the mean, μ, which are valid in the case of normality must be correct for large n regardless of the form of the population.

In the following discussion of effects of non-normality, the degree of violation of normality will often be described partially in terms of the skewness and kurtosis of a non-normal distribution. A common measure of skewness is denoted by β_1 and has the following definition:

$$\sqrt{\beta_1} = \frac{E(X - \mu)^3}{\sigma^3}. \tag{4}$$

When $\beta_1 = 0$, the distribution is symmetrical. It is positive when the right tail of the distribution is elongated and the left tail is not; it is negative when the reverse is true.

The "flatness" or "peakedness" of a distribution is reflected by the kurtosis, β_2:

$$\beta_2 = \frac{E(X - \mu)^4}{\sigma^4}. \tag{5}$$

For a normal distribution, $\beta_2 = 3$. When β_2 exceeds 3, the distribution is said to be "leptokurtic"; the graph of its frequency distribution is peaked and has "thick" tails. When β_2 is less than 3, the

distribution is "platykurtic"; the graph of its frequency distribution tends to be flat with thinner tails.

It is difficult to generalize about the skewness and kurtosis of distributions likely to be encountered in practice. Of course, the skewness and kurtosis of important theoretical distributions are well-known; the normal and the rectangular distributions have zero skewness and kurtosis of 3 and 1.8, respectively. The binomial distribution with parameters n and P has skewness $(1 - 2P) / \sqrt{nP(1 - P)}$ and kurtosis $3 + [1 - 6P(1 - P)] / [nP(1 - P)]$, which approach 0 and 3, respectively, as n increases. When $n = 1$, hence the distribution is a two-point binomial distribution, the kurtosis ranges from a minimum value of 1 at $P = ½$, a value of 3 at $P = .21$ or .79, and it increases without bound beyond 3 as P approaches 0 or 1. Student's t-distribution is symmetrical and becomes increasingly leptokurtic as the degrees of freedom approach zero. Empirical estimates of skewness and kurtosis are scattered among statistical literature. Kendall and Stuart (1963, p. 57) reported the frequency distribution of age at marriage for over 300,000 Australians; the skewness and kurtosis were 1.96 and 8.33, respectively. The distribution of heights of 8,585 English males (see Glass & Stanley, 1970, p. 103) had skewness and kurtosis of —0.08 and 3.15, respectively. Student (1927) noted that observation error in routine chemical analysis generally followed a slightly leptokurtic distribution. Lord (1952) presented some actual distributions of test scores along with the associated measures of skewness and kurtosis. Three of Lord's eight distributions are graphed in Figure 2 together with the values of $\sqrt{\beta_1}$ and β_2.

E. S. Pearson (1931) presented early empirical results which showed that for the ANOVA with two groups the actual and nominal probabilities of a Type-I error are nearly equal when skewed distributions are sampled. Similarly, departure from the mesokurtosis ($\beta_2 = 3$) of the normal distribution had little effect. Pearson also compared the nominal α and empirical probabilities of a Type-I error for $J = 5$, 10 and $n = 4$, 5, 10 when the populations were highly positively skewed and not mesokurtic. Again, failure to meet the normality assumption gave no cause for concern.

Norton's (1952; see also Lindquist, 1953) early empirical study is of particular interest as he compared nominal and actual percentiles at nine different points in the F-distribution (see Table 4). The discrepancies were generally quite minor even for very small samples ($n = 3$).

Boneau (1960) compared actual with nominal .05 and .01 significance levels for various sample sizes and violations of the assumptions of homogeneous variances and normality for the two-group t-test. The actual significance levels are based on samples of 1,000 t-ratios. Entire frequency distributions of the t-ratios obtained when t-tests are performed on non-normal populations or populations with unequal variances are reproduced in the 1960 article by Boneau.

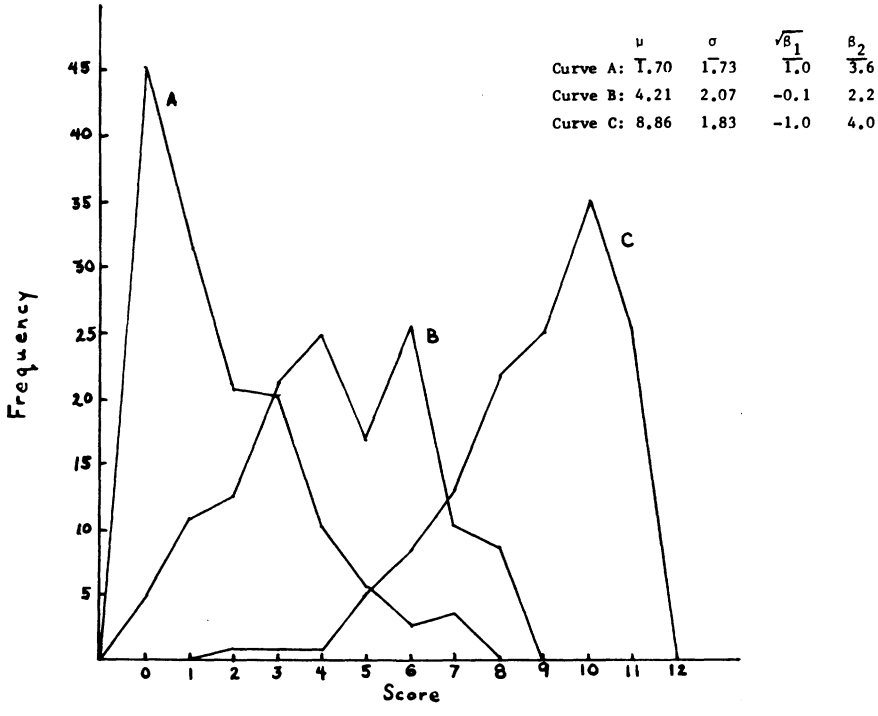
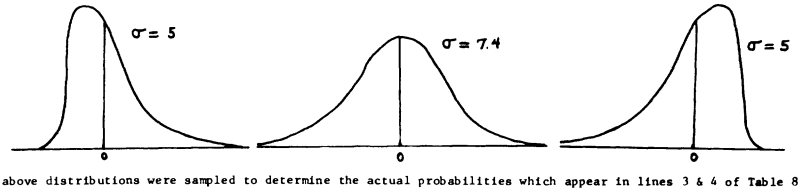
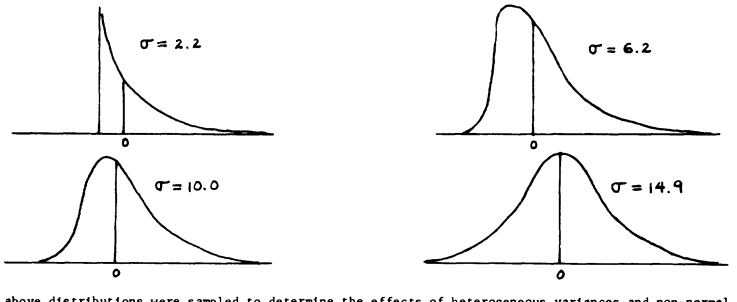


FIG. 2: Illustration of various values of skewness and kurtosis of score distributions. (After Lord, 1952).



(The above distributions were sampled to determine the actual probabilities which appear in lines 3 & 4 of Table 8.)



(The above distributions were sampled to determine the effects of heterogeneous variances and non-normal distributions. See lines 5 & 6 of Table 8.)

FIG. 3: Graphs of distributions sampled in the Norton study upon which data in Table 8 are based.

TABLE 4

*Effects of Non-normality on the Null Distribution in the One-factor Analysis of Variance**

Nominal Probability of Type-I Error:					*.50*	*.25*	*.20*	*.10*	*.050*	*.025*	*.010*	*.005*	*.001*
Description of Population	*J*	*n*	*Skewness*	*Kurtosis*	*Actual Probabilities of Type-I Error (Empirically Estimated)***								
Normal***	4	5	0	3.0	.51	.25	.20	.10	.056	.029	.014	.007	.002
Leptokurtic	3	3	0	7.0	.53	.28	.23	.13	.078	.046	.028	.015	.008
Leptokurtic	4	5	0	7.0	.54	.29	.22	.11	.066	.038	.016	.010	.004
Rectangular	3	3	0	1.7	.48	.25	.20	.12	.061	.032	.018	.009	.001
Mod. Skewed	4	5	0.5	3.7	.51	.25	.20	.10	.052	.028	.013	.007	.001
Ext. Skewed	3	3	1.0	3.8	.50	.24	.19	.10	.048	.021	.008	.004	.001
Ext. Skewed	4	5	1.0	3.8	.51	.26	.20	.10	.048	.023	.010	.007	.001
J-Shaped	3	3	1.4	3.6	.51	.24	.19	.09	.048	.026	.010	.005	.002

* Data due to Norton; see Lindquist (1953, p. 82).

** Each empirical probability is based on 3,000 cases; the probabilities within each row are based on the same 3,000 cases.

*** This case was included as a check on the accuracy of the empirical sampling procedure.

We shall give only a few of the more easily reproduced findings from Boneau's work (see Table 5).

TABLE 5

*Actual Probabilities of Type-I Errors in the One-way ANOVA with Two Groups for Various Non-normal Populations**

Population 1			*Population 2*			*Nominal Level of Significance*	
Shape	σ_1^2	n_1	*Shape*	σ_2^2	n_2	*.05*	*.01*
Normal	1	5	Normal	1	5	.053**	.009**
Exponential***	1	5	Exponential	1	5	.031	.003
Exponential	1	15	Exponential	1	15	.040	.004
Rectangular	1	5	Rectangular	1	5	.051	.010
Rectangular	1	15	Rectangular	1	15	.050	.015
Normal	1	5	Rectangular	1	5	.056	.010
Normal	1	15	Rectangular	1	15	.056	.010
Exponential	1	5	Normal	1	5	.071	.019
Exponential	1	15	Normal	1	15	.051	.014
Exponential	1	25	Normal	1	25	.046	.013
Exponential	1	5	Rectangular	1	5	.064	.033
Exponential	1	15	Rectangular	1	15	.056	.016

* Data due to Boneau (1960).

** This case, for which the ANOVA assumptions are met, was included as a check on the empirical sampling procedure.

*** The exponential distribution looks something like the upper 1/3 of a normal distribution. See Glossary.

Sawrey (1958) noted that Cochran (1950) considered the problem of ANOVA on a dichotomous variable (perhaps the ultimate in non-normality) nonetheless amenable to treatment by conventional normal theory methods:

> Without having looked into the matter, I had once or twice suggested to research workers that the *F*-test might serve as an approximation even when the table consists of 1's and 0's. As a testimony to the modern teaching of statistics, this suggestion was received with incredulity, the objection being made that the *F*-test requires normality, and that a mixture of 1's and 0's could not by any stretch of the imagination be regarded as normally distributed. The same workers raise no objections to a χ^2 test, not having realized that both tests require to some extent an assumption of normality, and that it is not obvious whether F or χ^2 is more sensitive to the assumption (Cochran, 1950, p. 262).

Lunney (1970) studied the effect of this flagrant violation of the normality assumption. Several ANOVA designs were investigated with regard to the analysis of a dichotomous dependent variable. The investigation covered only fixed-effects ANOVA models with equal n's. However, the implications are great for researchers in fields where observations must be reported as dichotomous data. We shall only present Lunney's results for the one-way ANOVA. In Table 6 are presented the actual and nominal significance levels of the F-distribution for 2, 3, 4, and 5 levels of the factor.* The results of the study must not be generalized to cases where cell n's are unequal. The

*Probabilities of the occurrence of a "success" on a binomial trial were varied (.1, .2, .3, .4 and .5) for each configuration and sample size. Sample sizes employed were 3, 7, 11, 15, 19, 23, 27 and 31 observations per level. Lunney reported that if the probability of a "success" is between .2 and .8 and the number of degrees of freedom for the error term is equal to twenty or more, the actual Type I error rate does not vary greatly from the nominal Type I error rate using a dichotomous variable. If the probability of a "success" is around .1 or .9, forty degrees of freedom are needed for the error term.

TABLE 6

*Actual (Empirical) Percentile Values of the F-Distribution Associated With the One-Way ANOVA on a Binomial Variable (1000 replications)**

	Nominal Significance Level			
	.10	*.05*	*.025*	*.01*
J = 2	.101[a]	.048	.023	.010
	.016[b]	.007	.006	.004
J = 3	.099[a]	.051	.026	.011
	.010[b]	.007	.005	.003
J = 4	.098[c]	.047	.024	.010
	.011[d]	.009	.005	.004
J = 5	.099[c]	.051	.026	.010
	.009[d]	.008	.005	.003

* Data based on Lunney (1969).

[a] Average of actual significance levels over 27 simulations varying n and the binomial parameter.

[b] Standard deviation of actual significance levels over the same 27 simulations.

[c] Average of actual significance levels over 34 simulations varying n and the binomial parameter.

[d] Standard deviation of actual significance levels over the same 34 simulations.

robustness of the *F*-test on dichotomous data will *not* hold with unequal *n*'s.

Hsu and Feldt (1969) investigated the effect of scale limitations on Type I error rate in *F*-tests involving independent groups. Four scale lengths were studied: five, four, three, and two points. They found:

1. The *F*-test gave excellent control of Type I error with a five-point scale. Moderate heterogeneity of variance, platykurtosis ($\beta_2 < 3$), or skewness had little effect, with samples as small as 11 cases per group.
2. Not quite as precise control of Type I error rate was found with a four-point scale; however, the differences between actual and nominal significance levels were still quite small.
3. Control of Type I error rates with a three-point scale was slightly better than that for a four-point scale, though not as good as that for a five-point scale.

TABLE 7

Effect of Non-normal (Limited Score Values) on the Level of Significance of the Fixed-effects F-*test (Data due to Hsu and Feldt, 1969)*

Five-Point Scale					*Four-Point Scale*				
		Nominal Significance Levels					*Nominal Significance Levels*		
Populations Sampled	n	*.10*	*.05*	*.01*	*Populations Sampled*	n	*.10*	*.05*	*.01*
A,A,A,A	11	.1014	.0516	.0104	D,D,D,D	11	.1100	.0580	.0104
A,A,A,A	51	.0986	.0518	.0096	D,D,D,D	51	.1022	.0522	.0118
B,B	11	.1074	.0518	.0130	E,E	11	.1042	.0506	.0110
B,B	51	.1024	.0516	.0118	E,E	51	.0978	.0486	.0092
B,B,B,B	11	.0976	.0516	.0104	E,E,E,E	11	.0988	.0538	.0100
B,B,B,B	51	.1004	.0492	.0108	E,E,E,E	51	.1000	.0486	.0086
C,C	11	.0992	.0470	.0088	F,F	11	.0992	.0444	.0110
C,C	51	.1000	.0502	.0094	F,F	51	.0940	.0484	.0108
C,C,C,C	11	.1016	.0480	.0096	F,F,F,F	11	.0974	.0532	.0104
C,C,C,C	51	.0974	.0522	.0108	F,F,F,F	51	.0926	.0524	.0116
A,B*	11	.1128	.0556	.0100	D,E*	11	.1020	.0542	.0094
A,B	51	.0996	.0504	.0106	D,E	51	.1172	.0584	.0120
A,B,B,B	11	.1040	.0550	.0132	D,E,E,E	11	.1038	.0546	.0134
A,B,B,B	51	.1016	.0494	.0128	D,E,E,E	51	.1092	.0542	.0134
Population	μ	σ^2	β_1	β_2	*Population*	μ	σ^2	β_1	β_2
A	2.00	1.04	0	2.51	D	1.50	0.89	0	2.10
B	2.00	0.54	0	3.09	E	1.50	0.45	0	2.78
C	2.35	1.15	−0.39	2.73	F	1.75	0.66	−0.08	2.38

* Note: The following four simulations also involve heterogeneous variances.

TABLE 7 (*Continued*)

Three-Point Scale					*Two-Point (Dichotomous) Scale*				
Populations Sampled	n	*Nominal Significance Levels* .10	.05	.01	*Populations Sampled*	n	*Nominal Significance Levels* .10	.05	.01
G,G,G,G	11	.1059	.0541	.0137	J,J	11	.1333	.0511	.0188
G,G,G,G	51	.0996	.0476	.0091	J,J	51	.0893	.0527	.0100
H,H	11	.0958	.0492	.0080	J,J,J,J	11	.1093	.0510	.0102
H,H	51	.1005	.0492	.0103	J,J,J,J	51	.0993	.0511	.0111
H,H,H,H	11	.0978	.0488	.0105	K,K	11	.1256	.0489	.0172
H,H,H,H	51	.1032	.0544	.0107	K,K	51	.0958	.0553	.0115
I,I	11	.1030	.0446	.0119	K,K,K,K	11	.1170	.0542	.0147
I,I	51	.1003	.0486	.0078	K,K,K,K	51	.0980	.0496	.0102
I,I,I,I	11	.1026	.0520	.0104	L,L	11	.1040	.0403	.0128
I,I,I,I	51	.1003	.0511	.0102	L,L	51	.1015	.0481	.0091
G,H*	11	.1032	.0503	.0131	L,L,L,L	11	.1011	.0514	.0104
G,H	51	.1076	.0493	.0099	L,L,L,L	51	.0980	.0507	.0093
G,H,H,H	11	.1026	.0550	.0157					
G,H,H,H	51	.1016	.0554	.0129					
Population	μ	σ^2	β_1	β_2	*Population*	μ	σ^2	β_1	β_2
G	1.00	0.66	0	1.15	J	0.50	0.25	0	1.00
H	1.00	0.33	0	3.03	K	0.40	0.24	0.41	1.17
I	0.85	0.43	0.16	2.29	L	0.25	0.19	1.15	2.33

* Note: The following four simulations also involve heterogeneous variances.

4. Much less control of Type I error rate was found with a two-point scale than with three-, four-, and five-point scales. Larger sample sizes resulted in more precise control of Type I error.

One of the rare illuminating mathematical findings concerning the effect of non-normality on the F-test derives from early work by E. S. Pearson (1931). Pearson showed that *under a true null hypothesis* the variances and covariance of MS_B and MS_W can be approximated by

$$\sigma^2_{MS_B} \doteq 2\sigma^4 [1 + \tfrac{1}{2}(\beta_2 - 3)(J - 1)/(nJ)]/(J - 1),$$

$$\sigma^2_{MS_W} \doteq 2\sigma^4 [1 + \tfrac{1}{2}(\beta_2 - 3)(n - 1)/n]/[J(n - 1)],$$

$$\sigma_{MS_B, MS_W} \doteq \sigma^4 (\beta_2 - 3)/(Jn).*$$

*Geary (1936) proved that *only* with the normal distribution will the sample mean and variance be independent.

The variance of a ratio of two variables can be approximated by (see K. Pearson, 1897):

$$\sigma^2_{\frac{x}{y}} \doteq \frac{\mu_x^2}{\mu_y^2}\left[\frac{\sigma_x^2}{\mu_x^2} + \frac{\sigma_y^2}{\mu_y^2} - \frac{2\sigma_{xy}}{\mu_x\mu_y}\right]. \tag{6}$$

Noting that $E(MS_B) = E(MS_W) = \sigma^2$ and substituting the approximate variances and covariance into (6) yields, after some algebra, the following approximation to the variance of MS_B/MS_W under the null hypothesis and with non-normal distributions:

$$\sigma^2_{MS_B/MS_W} \doteq \frac{2}{J-1} + \frac{2}{J(n-1)}. \tag{7}$$

Particularly noteworthy is that the original approximations to the variances and the covariance of MS_B and MS_W are independent of the skewness of the non-normal distributions and that the variance of the F-ratio is independent of the kurtosis because of compensating effects of β_2 on the variances and the covariance of the mean square. The question remains of how (7) compares to the variance of $F = MS_B/MS_W$ under the null hypothesis and assuming normal distributions.

From Kendall and Stuart (1963, p. 378), the variance of $F = MS_B/MS_W$ under the null and normal theory assumptions is

$$\sigma^2_{MS_B/MS_W} = \frac{2[J(n-1)]^2\,[(J-1) + J(n-1) - 2]}{(J-1)[J(n-1)-2]^2\,[J(n-1)-4]}. \tag{8}$$

Taking $J(n-1)$ "large," (8) reduced to

$$\sigma^2_{MS_B/MS_W} \doteq \frac{2}{J-1} + \frac{2}{J(n-1)}. \tag{9}$$

Hence, even though the distributions sampled are non-normal, the null distribution of $F = MS_B/MS_W$ has a variance which approximately equals the variance of the corresponding central F-distribution. Since one might also expect the mean of MS_B/MS_W to be near unity under a true null regardless of the non-normality of the distributions (and since for $J(n-1)$ "large" the mean of the central F-distribution is approximately unity), Pearson's findings are a compelling rationalization of the widely noted robustness of the fixed-effects ANOVA to non-normality.

Bradley (1963, 1966) created some doubt about the robustness of the t- and F-tests to violation of the normality assumption, especially at increasingly remote tail regions (that is, beyond .01). His point is well taken that it is risky to generalize the results of a few studies of α to any specific distribution and set of experimental conditions. However, we are unsympathetic to dramatizations of the lack of robustness of the ANOVA by appeal to small α's. Statements of significance at levels

beyond .001 ought not be taken too literally since minor violations of assumptions could easily distort the nominal .001 level into the .002 level (a 100% distortion!). Significance at the .01 level should be sufficiently convincing that appeal to untrustworthy smaller α's is unnecessary. Arguments that significance at smaller α-levels should be reported to emphasize strength of effects are unconvincing. Effects should be measured on the metric of the dependent variable and should be bracketed by appropriate confidence intervals; they should not be measured with α-levels.

Incautious statements concerning the robustness of the ANOVA to non-normality could send applied statistics off on a rerun of the unproductive 1950s stampede to nonparametric methods. Hawkridge (1970, p. 36) threatened the safety of the herd with this warning:

> The question of using t and F tests with such skewed distributions was brought to our notice during one particular study. . . . One of our staff at AIR . . . was statistical consultant for the study, and he went to some trouble to investigate the claims about the robustness of t and F tests. He showed that although Norton (in Lindquist, 1953) and Boneau (1960) had defended the robustness of t and F tests, the more recent work of Bradley . . . had raised new doubts about the violation of certain assumptions. This is not the place to go into detail about this debate, but Bradley's view . . . is that nonparametric statistics should be used when parametric assumptions are violated, rather than their normally more efficient parametric counterparts.

Studies of the effect of non-normality on the power of the F-test generally contain investigations of non-normality and α as special cases. The following tables (which appear later in this paper under "Effects of Non-normality on the Power of the ANOVA") contain further data on the effects of non-normality on the level of significance: Tables 9, 10, 12 and 13, and Figure 4.

Additional references: Bartlett (1935); Pitman (1937); Daniels (1938); Gayen (1949, 1950a, 1950b); Gronow (1951); Bradley (1952); Horsnell (1953); Hack (1958); Levene (1960); Atiqullah (1962); Baker, Hardyck, and Petrinovich (1966).

Heterogeneous Variances and Non-normality

Certainly one of the most ambitious early empirical studies of the effects of violation of the ANOVA assumptions was carried out by Norton as part of his dissertation research at the University of Iowa in the early 1950s. Norton's findings first appeared in print in Lindquist (1953).

We present Norton's results here, even though they show no new trends, for two reasons: (1) Norton investigated the dual effect of

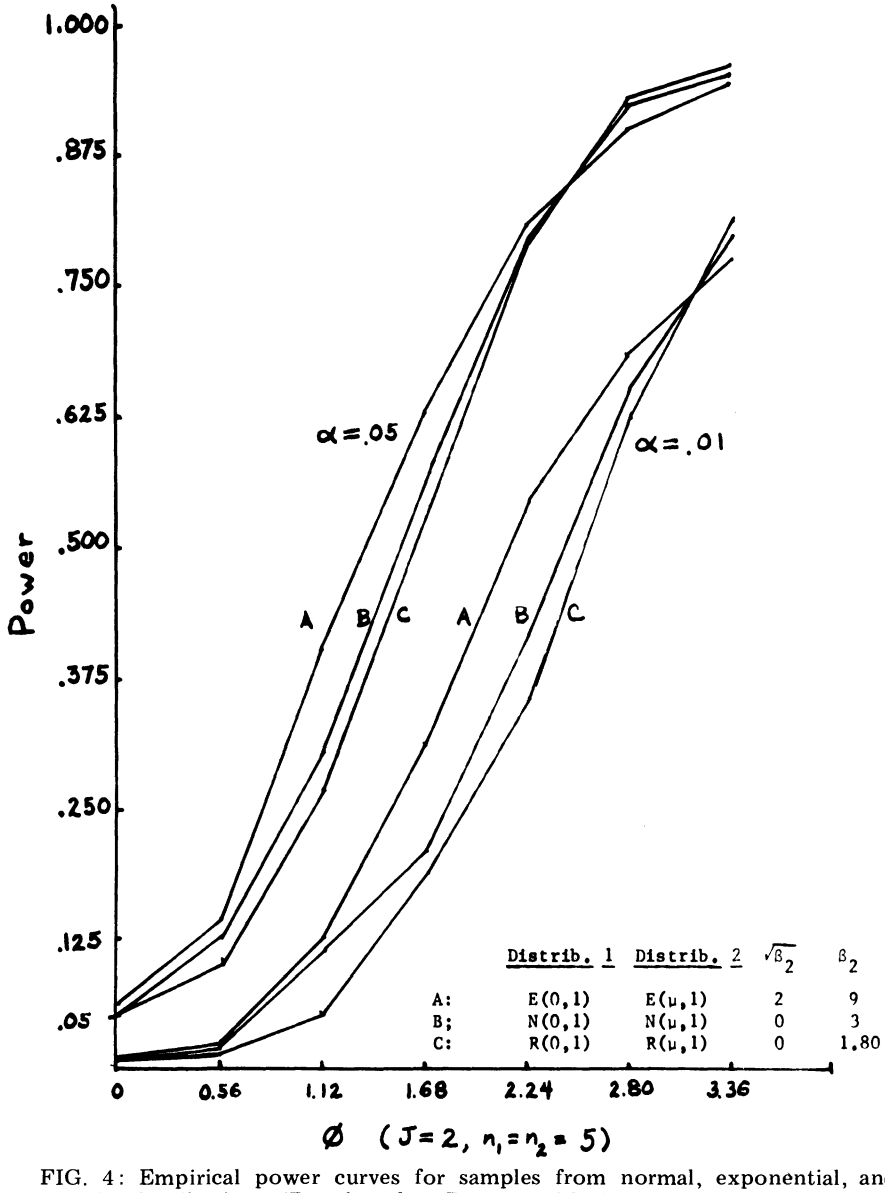


FIG. 4: Empirical power curves for samples from normal, exponential, and rectangular distributions. (Data based on Boneau, 1962.)

non-normality and heterogeneous variances, and (2) he compared actual and nominal significance levels for nine different points in the *F* distribution. Table 8 differs only trivially from the tables on pages 82 and 84 of Lindquist's text. The data in Table 8 support the conclusions drawn earlier in this paper for the one-way fixed-effects ANOVA. If the

TABLE 8

*Combined Effects of Heterogeneous Variances and Non-normality on the Null Distribution in the One-factor ANOVA**

Nominal Probability of Type-I Error:					*.50*	*.25*	*.20*	*.10*	*.05*	*.025*	*.01*	*.005*	*.001*
No. of Cases	*Description of Population***	*J*	*n*	*df*	*Actual Probabilities of Type-I Error (Empirically Estimated)*								
3333	Heterogeneous Variances ($\sigma_1^2 = 25$, $\sigma_2^2 = 100$,	3	3	2,6	.5012	.2748	.2301	.1347	.0726	.0429	.0213	.0123	.0036
3000	$\sigma_3^2 = 225$)	3	10	2,27	.4946	.2683	.2156	.1173	.0656	.0383	.0200	.0120	.0037
3333	Heterogeneous Forms	3	3	2,6	.4989	.2685	.2256	.1323	.0672	.0384	.0165	.0102	.0015
3333		3	6	2,15	.4978	.2625	.2169	.1137	.0681	.0393	.0198	.0129	.0039
3333	Heterogeneous Forms and Variances ($\sigma_1^2 = 4.84$,	4	3	3,8	.4884	.2853	.2463	.1588	.1002	.0633	.0357	.0216	.0084
	$\sigma_2^2 = 38.44$, $\sigma_3^2 = 100.00$												
3000	$\sigma_4^2 = 221.61$)	4	10	3,36	.4735	.2690	.2313	.1337	.0810	.0530	.0293	.0190	.0087

* Data due to Norton (see Lindquist, 1953).

** See Figure 3 for a description of the populations sampled.

Norton study has any shortcomings, it is that the effects of violation of the assumptions were studied only for ANOVA designs with equal n's. Boneau's study (1960) also included two simulations testing the combined effect on α of non-normality and heterogeneous variances. In sampling 5 cases from each of two rectangular distributions with variances 1 and 4, Boneau obtained .071 and .019 actual error rates and the nominal .05 and .01 significance levels. Under the same conditions but with exponential in place of rectangular distributions, actual significance levels of .083 and .017 were observed at the nominal .05 and .01 levels, respectively.

Additional references: Bhattacharjee (1968); Horsnell (1953); Neave and Granger (1968).

Effects of Non-normality and Heterogeneous Variances on the Power of the ANOVA

In this section most of what is known about the effects of non-normality and heterogeneous variances on the *power* of the ANOVA will be summarized. Previous sections in this paper dealt only with distortions in the sampling distribution of the F-ratio (arising from non-normality and/or heterogeneous variances) *when the null hypothesis is true.* The problem remains of the correspondence between the distribution of the F-ratio under various *false null hypotheses* when the ANOVA assumptions are satisfied and when they are not.

The deviation of a set of J population means from equality (for purposes of investigating the power of the F-test) is expressed in terms of a *non-centrality parameter.* Suppose that J population means have the values $\mu_1, \ldots, \mu_J$, and σ^2 is the population variance of any one population. The value of the non-centrality parameter for these J means is given by

$$\delta^2 = \frac{n\sum_{1}^{J} (\mu_j - \bar{\mu}.)^2}{\sigma^2}, \text{ where} \tag{10}$$

$$\bar{\mu}. = (\mu_1 + \ldots + \mu_J)/J.$$

A second conventional means of expressing the degree of violation of the null hypothesis is a simple function of the non-centrality parameter.

$$\phi = \sqrt{\frac{\delta^2}{J}} \tag{11}$$

The non-centrality parameter involves σ^2 which is assumed homogeneous across groups in the ANOVA model. A logical problem is

encountered in asking the question whether the power of the ANOVA is affected by heterogeneous variances. If the variances are heterogeneous, what does one use for the value of σ^2 in calculating ϕ and the nominal (theoretical) power? There is no "theoretical" power since there is no "theory" of the non-central F-distribution under heterogeneous variances. Horsnell (1953) and Donaldson (1968) dealt with the problem by substituting the average within-group variance, that is, $(\sigma_1^2, + \ldots + \sigma_J^2)/J$, for σ^2 in the formula for ϕ. A nominal power of some sort is thus obtained, but its meaning is different from the concept of "nominal power" met thus far. (We deal with this problem in detail under "Heterogeneous Variances and Power.")

Games and Lucas (1966) reported an empirical investigation of the effect of non-normality on the power of the fixed-effects ANOVA. Most of their results appear in Table 9. Each of the empirical probabilities in Table 9 is based on 1,000 F-ratios.

In general, the correspondence between the theoretical normal-theory power calculations and the empirical power values for non-normal populations is remarkably close. Populations of the form described in lines 6-8 of Table 9 produced empirical powers with sizable deviations from theoretical normal theory power, however. Specifically, the effect on power was great when the non-normal populations were extremely skewed or quite leptokurtic. Moderate departures from normality had no effect of practical importance however. (A further attempt to identify trends will be made later in the discussion of the theoretical non-empirical work done by Srivastava.) It must be held in mind when evaluating Games and Lucas' work that all three populations sampled were non-normal in the same way (that is, all three were leptokurtic or moderately skewed, or extremely skewed, and so on).

Perhaps the most extensive theoretical investigation of the power of ANOVA for non-normal populations is that of Srivastava (1959). Srivastava presented tables of the power of the F-test for various values of skewness and kurtosis of the underlying distribution. (The population distributions Srivastava considered were all members of a family of distributions that can be obtained by varying parameters in an *Edgeworth Series;* this series was presented by Edgeworth in 1904 as a means of representing skewed distributions.) Only a small part of Srivastava's findings are presented in Table 10, but those results presented there convey his findings in general.

In Srivastava's study, the actual power of the F-test for any combination of values of skewness, kurtosis, and ϕ depended also on precisely how the true population means were arranged along a scale. In other words, even though ϕ equals 2.0, the power for non-normal distributions will vary slightly when the three population means are 11, 11, and 14 and when they are 10, 12, and 14. These differences were so slight, however, that they have been disregarded in constructing Table 10. The entries in Table 10 for nonzero values of ϕ assume that two

TABLE 9

*Actual (Empirical) Probabilities of Rejecting the Null Hypothesis at the .05 Level for Various Non-normal Distributions and Values of ϕ**

Nature of Distribution	*Skewness*	*Kurtosis*	*Values of ϕ*									
			0		*0.5*		*1.0*		*1.5*		*2.5*	
Normal	0	2.90	.046**	.059	.098	.091	.200	.275	.430	.519	.851	.945
Slightly Skewed	.45	3.53	.052	.052	.077	.096	.193	.266	.410	.556	.848	.947
Square Root Trans.	0	2.91	.054	.046	.086	.098	.201	.286	.518	.575	.834	.946
Moderately Skewed	.64	3.53	.058	.054	.080	.100	.188	.273	.446	.563	.858	.943
Logarithm Trans.	0	2.82	.058	.057	.094	.097	.213	.268	.445	.594	.828	.941
Extremely Skewed	2.04	9.54	.048	.037	.108	.078	.264	.326	.541	.559	.854	.910
Reciprocal Trans.	.03	2.88	.059	.057	.134	.120	.325	.485	.536	.657	.792	.964
Leptokurtic	0	9.16	.049	.036	.101	.109	.269	.338	.522	.584	.874	.924
Rectangular	0	1.80	.070	.058	.106	.094	.209	.258	.400	.538	.856	.954
Theoretical (nominal) Power			*.050*	*.050*	*.087*	*.097*	*.206*	*.261*	*.417*	*.540*	*.851*	*.954*

* Data due to Lucas & Games (1966).

** In each of the five columns, the left entry is based on $J = 3$, $n = 3$ and the right entry is based on $J = 3$, $n = 6$.

TABLE 10

*Actual Probabilities (Theoretically Approximated) of Rejecting the Null Hypothesis at the .05 Level for Various Non-normal Distributions and Values of ϕ**

			Values of ϕ									
Skewness		*Kurtosis*	*0*		*1.0*		*1.5*		*2.0*		*2.5*	
1.	0	2.0	.054**	.052**	.249	.308	.531	.649	.819	.908	.967	.993
2.***	0	3.0	*.050*	*.050*	*.262*	*.319*	*.551*	*.659*	*.821*	*.907*	*.957*	*.988*
3.	0	3.5	.048	.049	.268	.325	.561	.664	.822	.907	.952	.986
4.	0	4.0	.046	.048	.275	.330	.571	.669	.823	.906	.947	.983
5.	0	5.0	.042	.045	.288	.342	.590	.679	.824	.906	.937	.978
6.	0	5.4	.041	.044	.290	.346	.598	.683	.825	.905	.933	.976
7.	.25	2.0	.055	.053	.248	.307	.532	.649	.820	.908	.967	.993
8.	.25	3.0	.051	.050	.261	.318	.552	.659	.822	.907	.957	.988
9.	.25	3.5	.049	.049	.267	.324	.562	.664	.823	.907	.953	.985
10.	.25	5.0	.044	.045	.287	.341	.591	.679	.825	.906	.938	.978
11.	.49	5.4	.043	.045	.291	.345	.600	.683	.827	.906	.934	.976

* Data due to Srivastava (1959).

** In each of the five columns, the left entry is for $J = 5$, $n = 3$ and the right entry is for $J = 5$, $n = 5$.

*** The normal distribution.

means are large and equal, two means are small and equal, and the fifth mean is midway between these groups.

Srivastava (1958) also investigated the influence of skewness and kurtosis, individually and together, on the power of the simple *t*-test; by "simple *t*-test" is meant the situation in which the null hypothesis is that the population mean, μ, equals some value a and the test statistic referred to the *t*-table is $(\bar{x}. - a)/(s/\sqrt{n})$. The "correlated" or "dependent-groups *t*-test" is an example of such a test. Srivastava's results show enlightening trends associated with the values of β_1 and β_2; some of these results appear in Table 11. Pearson (1929; also see Scheffé, 1959, footnote #32, pp. 350-351) reported that the effect of skewness in testing H_0: $\mu = a$ is as follows: if $\mu - a$ is positive and the distribution is positively skewed, then the actual power exceeds the nominal power; if $\mu - a$ is positive and the distribution is negatively skewed, then the actual power is less than the nominal power. Note, however, that this relationship is not borne out in Srivastava's data. It obtains in the one case ($\beta_2 = 3$) for which the nominal power is high (.868) but not in the other two cases (nominal powers of .580 and .236). (See Pearson (1958) for a comparison of Srivastava's and Pearson's 1929 results.)

Boneau (1962) presented simulation data incidentally in the context of comparing the power of the *t*-test and the Mann-Whitney

TABLE 11

*Actual Probabilities (Analytically Estimated) of Rejecting the Null Hypothesis that μ = a with a t-test at the .05 Level of Significance ($n = 10$)**

Kurtosis	Value of $\frac{\mu - a}{\frac{\sigma}{\sqrt{n}}}$	Skewness				
		−.36	−.16	0	.16	.36
2	1	.254	.245	.222	.195	.179
	2	.556	.559	.559	.555	.550
	3	.842	.850	.870	.898	.914
3	1	.268	.259	.236**	.208	.193
	2	.576	.579	.580**	.575	.571
	3	.839	.847	.868**	.895	.911
4	1	.282	.272	.250	.222	.206
	2	.597	.600	.601	.596	.591
	3	.837	.845	.866	.893	.909

* Data due to Srivastava (1958).

** The power of the test when the population is normal.

U-test which bear on the problem of the power of the one-factor ANOVA with two groups under non-normality. By combining data on Boneau's Figures 2, 8, and 9, the graphs in Figure 4 can be obtained. Each of the 42 points on the curves in Figure 4 is based on 1,000 replications of the simulation. In each simulation, Boneau randomly drew two samples of $n = 5$ cases each from either a pair of normal, rectangular or (negative) exponential distributions. The normal and rectangular distributions were scaled to mean zero and variance one; although the skewness and kurtosis of the rectangular distribution were not reported, they can easily be calculated from theory to be 0 and 1.80, respectively. Boneau chose the formula e^{-x} for the probability density of the (negative) exponential distribution. Presumably X takes on all non-negative values. It can be shown that the kth cumulant of this distribution is equal to $(k-1)!$, and thus that

$$\mu = \kappa_1 = 0! = 1,$$

$$\sigma^2 = \kappa_2 = 1! = 1,$$

$$\sqrt{\beta_1} = \kappa_3/\kappa_2^{3/2} = 2! = 2, \tag{12}$$

$$\beta_2 = (\kappa_4 + 3\kappa_2^2)/\kappa_2^2 = 3! + 3(1) = 9. \tag{13}$$

Boneau (1960, p. 52) reported only empirical estimates based on 5,000 sample values of the mean and variance of the exponential distribution. The estimated mean and variance were 0.0218 and 1.0475, respectively. It is clear that Boneau's exponential distribution had a mean of zero, though theoretically its mean is unity. This discrepancy cannot be resolved from the data and methodology Boneau reports.

The empirical power curves in Figure 4 were formed by varying the difference between μ_1 and μ_2 from 0 to 3 units in steps of 0.5, drawing two samples of size 5 from each population of test statistics exceeding ${}_{.95}F_{1,8}$ and ${}_{.99}F_{1,8}$.

Boneau's findings are in accord with those of Srivastava and Lucas and Games. The platykurtic rectangular distribution reduced the power slightly below that of normal theory; the highly leptokurtic negative exponential distribution increased the power above that of normal distribution theory. Boneau's exponential distribution corresponds—at least in the first four moments—fairly closely to Lucas and Games's "extremely skewed" distribution (line 6 of Table 9); the effects on power are similarly pronounced in these two situations.

Young and Veldman (1963a) performed a small empirical study in which the effect of skewness on the power of the two-tailed t-test (equivalent to the power of the one-way ANOVA, $J = 2$) was

investigated. Some of their results reported in an unpublished paper are presented in Table 12. The three skewed distributions sampled by Young and Veldman were formed by taking a normally distributed variable x and raising it to the third, sixth, and twelfth powers (x^3, x^6, x^{12}). Young and Veldman did not report the skewness and kurtosis of their populations. Furthermore they did not report the mean of the original normal distribution, so it was impossible to determine exactly the skewness and kurtosis of the powers of the normal variables. Young and Veldman did report, however, that the standard deviation of the original normal distribution was 12. Assuming that they quite likely chose a mean for the distribution which would make negative values of the variable rare, μ could have been as low as 50. We estimate that the skewness and kurtosis of the third, sixth, and twelfth powers of a normally distributed variable with mean 50 and standard deviation 12 are ($\beta_1 = 1.1, \beta_2 = 4.0$), ($\beta_1 = 2.5, \beta_2 = 9.8$) and ($\beta_1 = 4.7, \beta_2 = 28.2$), respectively. If our guess at Young and Veldman's choice for the mean is too low, our estimates of the skewness and kurtosis of the powers of the variables are too low also. If the original normal variables had a mean of 100, for example, the actual skewness and kurtosis of their powered variables were considerably larger than our approximations. Veldman indicated (personal communication) that information on what the mean of x actually was is no longer available.

TABLE 12

Actual (Empirical) Probabilities, based on 5,000 t-ratios, of Rejecting the Null Hypothesis with a Two-tailed t-test for Some Non-normal Distributions When the Probability of a Type-I Error Is .05. (J = 2, n = 10)*

	Value of $(\mu_1 - \mu_2)/\sigma_{\bar{x}_1 - \bar{x}_2}$				
Distributions Sampled	*0*	*1*	*2*	*3*	*4*
Normal $X \sim N(?,144)$	.05	.15	.48	.80	.95
X^3 ($\beta_1 = 1.1, \beta_2 = 4.0$)**	.03	.17	.49	.80	.94
X^6 ($\beta_1 = 2.5, \beta_2 = 9.8$)	.02	.27	.62	.82	.90
X^{12} ($\beta_1 = 4.7, \beta_2 = 28.2$)	.02	.65	.86	.91	.91

* Data due to Young and Veldman (1963b).

** Skewness and kurtosis calculations are based on the assumption that $\mu = 50$ for the original normal distribution (see text).

Young and Veldman's results indicate—presumably, for rather extreme skewness and kurtosis—that the actual power of the t-test is substantially greater than the theoretical power in the mid-range of the scale (.10 to .70, perhaps) but that it is slightly less than the theoretical power for large (near 1.0) values of the power. This same trend is apparent in Srivastava's data in Table 11.

Pearson (1929) presented data showing that the equal numbers of observations per group tended to diminish the effect of non-normality on power. The effect of kurtosis on power is apparently greater than the effect of skewness. In accord with one's expectations based on consideration of the central limit theorem, it appears that as n increases, all forms of non-normality have less and less effect on both power and significance level based on normal theory. Scheffé (1959) pointed out why one would expect that skewness would have less effect on the t-test of a difference, between means, especially with unequal group sizes, than on the test for a single mean.

Donaldson (1968; or see Heerman & Braskamp, 1970) investigated the effect of drawing samples from two or more non-normal parent populations (exponential and lognormal) on both the Type-I error rate and power. In Figure 5 the theoretical distributions for the three cases are shown. The non-normal distributions lead to conservative Type-I errors in all cases. These results agree with previous findings. In addition, he found that generally the F-test power for the non-normal distributions was higher than for the normal distribution, except for the very large values of power, for example, $1 - \beta$ above .90. This finding agrees with the findings of the studies presented above. It is noteworthy that the actual power exceeds the nominal power by as much as .20 in many instances. This discrepancy decreases with increasing n. Some of Donaldson's data appear in Table 13. Donaldson's study is exemplary of many of the best features of a simulation robustness study. Non-normal distributions are based upon theoretical distributions, rather than being *ad hoc*, inadequately described sets of scores; Monte

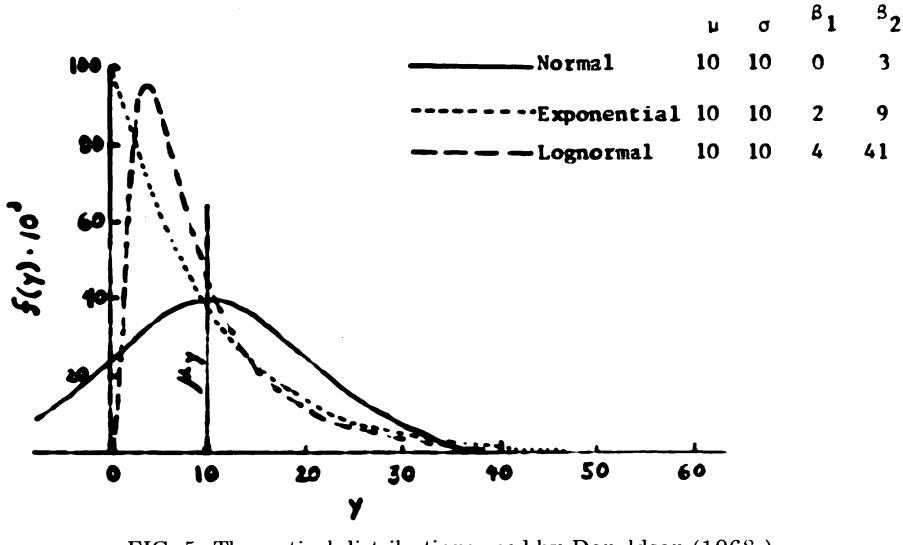


FIG. 5: Theoretical distributions used by Donaldson (1968.)

TABLE 13

Effect of Non-normality on the Power of the Fixed-effects ANOVA (Data based on Donaldson, 1968)

				Actual Power (Empirically Estimated)					
				*Theoretical (normal theory) power***		*Exponential distributions*		*Lognormal distributions*	
J	n	$\Delta\mu$*	ϕ	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
2	4	0	0	.05	.01	.042	.007	.031	.005
		5.000	0.50	(.089)	(.020)	.109	.025	.152	.037
		10.000	1.00	(.224)	(.065)	.310	.108	.419	.189
		15.000	1.50	.43	(.162)	.539	.266	.648	.396
		20.000	2.00	.66	.32	.723	.447	.788	.575
		25.000	2.50	.84	.50	.844	.615	.873	.704
2	8	0	0	.05	.01	.044	.007	.035	.005
		3.536	0.50	(.098)	(.025)	.115	.030	.145	.060
		7.072	1.00	(.260)	(.092)	.325	.128	.417	.201
		10.608	1.50	.50	(.244)	.573	.328	.663	.449
		14.144	2.00	.75	.46	.773	.553	.815	.653
		17.680	2.50	.91	.71	.887	.738	.899	.789
2	16	0	0	.05	.01	.046	.007	.038	.004
		2.500	0.50	(.102)	(.028)	.117	.030	.145	.060
		5.000	1.00	(.276)	(.105)	.325	.140	.391	.181
		7.500	1.50	.55	(.276)	.586	.349	.656	.438
		10.000	2.00	.78	.55	.797	.595	.824	.667
		12.500	2.50	.93	.78	.917	.791	.917	.818
2	32	0	0	.05	.01	.049	.009	.046	.007
		1.768	0.50	(.108)	(.030)	.113	.029	.130	.036
		3.536	1.00	(.289)	(.114)	.307	.129	.361	.112
		5.304	1.50	.55	(.307)	.575	.336	.627	.409
		7.072	2.00	.80	.58	.798	.596	.818	.655
		8.840	2.50	.93	.80	.933	.807	.918	.820
4	4	0	0	.05	.01	.041	.010	.035	.008
		2.500	0.56	(.110)	(.028)	.117	.032	.158	.049
		5.000	1.12	(.324)	(.120)	.400	.172	.513	.266
		7.500	1.68	.68	.36	.711	.453	.770	.574
		10.000	2.24	.92	.68	.885	.714	.887	.769
		12.500	2.80	.99	.89	.960	.867	.949	.876
4	8	0	0	.05	.01	.040	.007	.034	.006
		1.768	0.56	(.119)	(.034)	.129	.034	.153	.047
		3.536	1.12	(.386)	(.165)	.441	.216	.519	.296
		5.304	1.68	.76	.50	.770	.555	.808	.637
		7.072	2.24	.96	.84	.935	.822	.913	.821

TABLE 13 (*Continued*)

				Actual Power (Empirically Estimated)					
				*Theoretical (normal theory) power***		*Exponential distributions*		*Lognormal distributions*	
J	n	$\Delta\mu$*	ϕ	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
4	16	0	0	.05	.01	.047	.009	.037	.006
		1.250	0.56	(.129)	(.036)	.136	.039	.152	.047
		2.500	1.12	(.416)	(.194)	.449	.224	.510	.288
		3.750	1.68	.78	.55	.764	.590	.817	.647
		5.000	2.24	.97	.88	.956	.867	.941	.871
4	32	0	0	.05	.01	.045	.007	.044	.008
		0.884	0.56	(.131)	(.041)	.131	.036	.145	.040
		1.768	1.12	(.434)	(.211)	.454	.218	.491	.271
		2.652	1.68	.81	.60	.804	.607	.816	.645
		3.536	2.24	.98	.90	.967	.888	.952	.884

* $\Delta\mu$ is the difference between means (or successive means for $J = 4$).

** Theoretical (nominal) power values were read from the Pearson-Hartley charts. Values in parentheses are empirical estimates (10,000 replications) and were taken from Donaldson's tables because the P-H charts could not be used to obtain accurate readings at these points.

Carlo procedures are documented; empirical probabilities are based upon 10,000 replications; conclusions are carefully drawn and supported by mathematical argument. The researchers in this area would do well to attempt to emulate Donaldson's methods.

Additional reference: David and Johnson (1951).

Heterogeneous Variances and Power

The question of the effect of heterogeneous variances on the power of the fixed-effects model F-test cannot be addressed in the same manner that the effects of violation of normality on α and $1 - \beta$ or heterogeneous variances on α were determined. The paradigm of research in this area has been to compare actual levels of significance or power (under some violation of the assumptions) with the theoretical α or $1 - \beta$ calculated assuming that the assumptions are met. Determining the actual power given normality and heterogeneous variances presents no unique difficulties. If analytic attempts fail (as they have), then simulation techniques can be employed to estimate actual power levels. *However, there exists no method by which the theoretical power of the F-test can be determined when error variances are heterogeneous.* The power of the test depends on α, J, n and ϕ; but ϕ is a function of σ^2,

the common (that is, homogeneous) variance among the J groups. Unless a single, common value of the error variance can be specified, the value of ϕ cannot be determined; without a value for ϕ, the theoretical power cannot be determined. Hence, the question of the effect of heterogeneous variances on the theoretical power of the F-test is meaningless. Surprisingly, then, one does find published studies ostensibly addressed to this question. On closer examination, however, these studies prove to be addressed to a different question, which is often implicit and not clearly identified. In each of the three studies reviewed below, the indeterminant value of σ^2 in ϕ was replaced by the average among-groups variance (i.e., $\bar{\sigma}^2_{\cdot} = (\sigma^2_1 + \ldots + \sigma^2_J)/J$), and the "theoretical" power was determined using this average variance in the formula for ϕ. Thus, the question answered in these studies is whether standard power charts can be adapted (by means of using $\bar{\sigma}^2_{\cdot}$ for σ^2 in the ϕ formula) to approximate the actual power of the F-test when the group variances are heterogeneous. Even though investigations of this question are basically different from the other studies reviewed in this paper, they are reviewed here because historically they were suggested by studies of robustness of the ANOVA; they are often confused with such studies; and they are of practical importance (the practical import being that they are relevant to the question of how to approximate power when variances are heterogeneous).

Horsnell (1953) published the first investigation of the power of the F-test under heterogeneous variances. Horsnell approximated the power of the F-test by fitting the Edgeworth series* to the first four moments of a transformation of the F-ratio under conditions of unequal means and variances. Normal distributions within groups were assumed, and both equal and unequal n's for $J = 4$ groups were investigated. The approximate power curves were compared with the power of the F-test calculated from the Pearson-Hartley charts with $\bar{\sigma}^2_{\cdot}$ substituted for the common σ^2. Horsnell's results appear in Figure 6.

In Figure 6, the variances for the four groups are in the ratio 1:1:1:3; the population means for the last three groups are equal and the first mean differs from the other three (the degree of difference being reflected in the value of ϕ); the level of significance is .05. The broken-line curve in Figure 6 is the power of the F-test calculated with $\bar{\sigma}^2_{\cdot}$. The three remaining curves are based on various distributions of unequal n's. (Horsnell also presented results for the situation in which the one divergent population mean lies in the fourth group, with the largest variance, and all other parameters are the same. These results do not differ in any important respects from those in Figure 6.) One sees clearly in Figure 6 that *for equal n's of ten per group, there is a close correspondence between the actual power and the "theoretical" power calculated with* $\bar{\sigma}^2_{\cdot}$ in place of the non-existent common σ^2. For

*See footnote to Table 4 in Horsnell (1953).

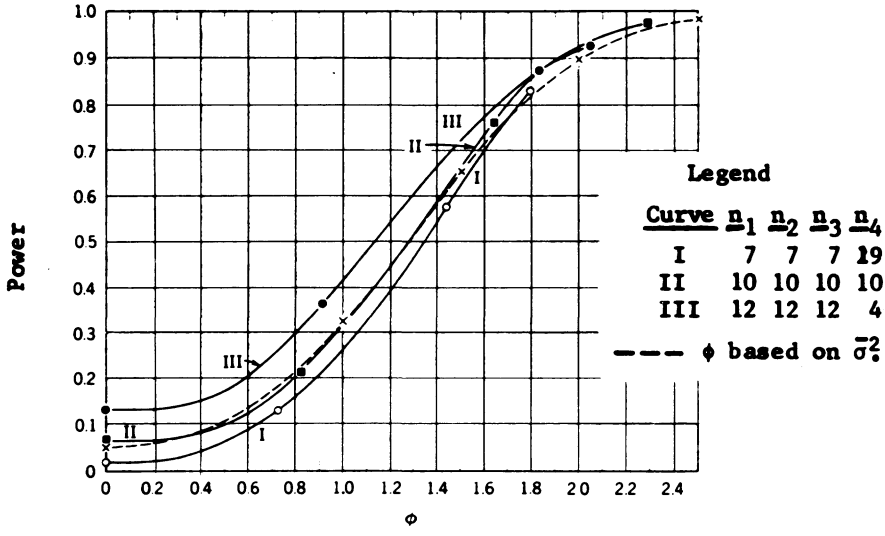


FIG. 6: Approximation of the power of the F-test under unequal variances ($\alpha = .05$). (Data due to Horsnell, 1953.)

moderate values of the power ($.10 < 1 - \beta < .60$) and unequal n's, the power calculated using $\bar{\sigma}^2$ is a poor approximation to the actual power of the test.

Donaldson's (1968) empirical study of the effect of non-normality on power included a brief investigation of power with heterogeneous variances. The data in Table 14 are based on Donaldson's results. Donaldson studied only the condition of equal n's. Values of the power were determined empirically (10,000 replications) for various values of numbers of groups, group sizes, α, and ϕ; samples were drawn from three types of distribution: normal, exponential, and lognormal (see Figure 5). Where available, we have supplied (in parentheses) the "theoretical" normal-theory power calculated with $\bar{\sigma}^2$ substituted for σ^2 in ϕ. Donaldson's findings for sampling from normal distributions are comparable to Horsnell's findings. They reveal close correspondence between the actual empirical power and the "theoretical" power calculated via $\bar{\sigma}^2$. It is clear, however, that with the particular non-normal distributions investigated and in those instances in which comparison with normal-theory "theoretical" power is possible, approximation of actual power *via* $\bar{\sigma}^2$ may be substantially in error when in fact the populations are not normally distributed.

Lunney (1968, 1969, 1970) apparently inadvertently conducted a simulation study which paralleled the work of Horsnell and Donaldson. Lunney extended his investigation of the effect of using a dichotomous variable on the level of significance (discussed above) to the effect of a dichotomous variable on the power of the test. However,

TABLE 14

Actual Power (Empirically Determined) Under Conditions of Heterogeneous Variances and Non-normality (Data based on Donaldson, 1968)

Parameters						*(Empirical) Power Values*									
						$\alpha = .10$			$\alpha = .05$				$\alpha = .01$		
J	n	$\Delta\mu$*	ϕ	$\bar{\sigma}^2_{\cdot}$**	$\frac{\sigma^2 max}{\sigma^2 min}$	*Norm.*	*Exp.*	*Lognor.*	*Norm.*		*Exp.*	*Lognor.*	*Norm.*		*Exp.*	*Lognor.*
2	16	2.5	0.44	128.25	1.56	.156	.157	.165	.090		.084	.086	.024		.015	.017
		5.0	0.79	162.50	2.25	.289	.287	.324	.186		.176	.200	.063		.042	.054
		10.0	1.27	250.00	—	.548	.582	.653	.414	(.40)***	.415	.498	.185		.153	.201
		15.0	1.58	362.50	—	.706	.789	.846	.586	(.58)	.643	.726	.317		.297	.405
		20.0	1.79	500.00	—	.802	.891	.935	.692	(.70)	.786	.856	.432		.443	.572
		25.0	1.94	662.50	12.25	.846	.948	.972	.760	(.76)	.872	.924	.512	(.50)	.566	.703
4	4	2.5	0.40	196.88	3.06	.143	.115	.123	.080		.056	.063	.020		.014	.014
		5.0	0.61	337.50	—	.209	.162	.205	.122		.089	.112	.032		.022	.029
		10.0	0.82	750.00	16.00	.300	.251	.330	.188		.141	.201	.057		.039	.058
		15.0	0.92	1337.50	—	.351	.313	.420	.228		.181	.265	.074		.053	.085
		25.0	1.01	3037.50	72.25	.405	.388	.523	.274	(.30)	.235	.348	.045	(sic)	.073	.123
4	16	1.25	0.47	142.97	1.89	.177	.166	.173	.104		.092	.096	.027		.022	.024
		2.50	0.80	196.88	3.06	.338	.318	.354	.221		.204	.225	.079		.064	.073
		3.75	1.04	261.72	—	.492	.485	.546	.362		.338	.342	.162		.127	.158
		5.00	1.22	337.50	—	.621	.633	.697	.484	(.48)	.475	.549	.247	(.25)	.211	.266
		7.50	1.45	521.88	—	.773	.827	.874	.664	(.65)	.698	.761	.406	(.40)	.379	.480
		10.00	1.63	750.00	—	.857	.921	.941	.762	(.76)	.827	.876	.527	(.53)	.531	.636
		12.50	1.75	1021.88	22.56	.897	.963	.970	.822	(.84)	.902	.925	.611	(.62)	.641	.737

* $\Delta\mu$ is the difference between successive means (e.g., if $\mu_1 = 10$ and $\Delta\mu = 2.5$, then $\mu_2 = 12.5$).

** $\bar{\sigma}^2_{\cdot} = (\sigma_1^2 + \ldots + \sigma_J^2)/J$.

*** Theoretical, normal-theory power assuming all J variances equal to $\bar{\sigma}^2_{\cdot}$; obtained from the Pearson-Hartley charts.

in setting the binomial parameters ($P_1, \ldots, P_J$) at unequal values to investigate power, one necessarily creates heterogeneous group variances (viz., $P_1 (1 - P_1), \ldots, P_J (1 - P_J)$). Thus Lunney compounded extreme non-normality (dichotomous data) with heterogeneous variances; but more importantly, with the group variances heterogeneous, the "theoretical" power is not defined. Without acknowledging the earlier leads of Horsnell and Donaldson, Lunney (1968, p. 56) averaged the group variances in his calculation of ϕ. Thus, his methodology was identical to that used by Horsnell and Donaldson.

Some of Lunney's data are reported in Table 15. He selected values of $P_1, \ldots, P_J$, J, n and α so as to create "theoretical" power values of either .80 or .60 (with power calculations based on the average group variance, of course). With the parameters set, 1,000 replications were performed and empirical estimates of actual power values obtained. For example, in the first line of Table 15, it is reported that 13 simulations (various sets of values of the P_j) were conducted with three groups and eleven cases per group with $\alpha = .01$ and the P_j chosen to give a "theoretical" power of .80. The 13 empirical estimates of actual power averaged .764 with a standard deviation of .051.

Unfortunately, Lunney made some incorrect calculations of the "theoretical" power values in the 1968 dissertation; the summary table in the 1969 paper and his conclusions on power in the 1970 publication (p. 266) are based on these erroneous data. Several of the .80 "theoretical" power values in Appendix 3 of the dissertation should be .60. The correct figures seem to indicate far closer agreement between the "theoretical" and actual power values than Lunney reported.

TABLE 15

*Comparison of Actual (Empirical) and Nominal Values of Power in a One-factor ANOVA on a Binomial Variable (1,000 replications)**

J	n	α	*Nominal power*	*Actual Power* Averaged *over simulations for varying values of binomial parameters*	*Standard deviation of actual power values over simulations*	*Number of simulations for various values of binomial parameters*
3	11	.01	.80	.764	.051	13
		.05	.80	.801	.048	12
3	11	.01	.60	.607	.034	21
		.05	.60	.602	.022	13
3	21	.01	.80	.782	.020	12
		.05	.80	.846	.050	6
3	21	.01	.60	.616	.027	4
		.05	.60	.594	.022	8

* Data based on Lunney (1969).

Summary of Effects of Violation of Assumptions of Fixed-effects ANOVA

We have attempted with Table 16 to summarize the major conclusions that can be drawn from the ANOVA robustness literature. The language in which the conclusions are stated is somewhat more categorical than it should be, perhaps. There appear to be exceptions, even in the tables reported above, to the general conclusions contained in Table 16; however, we are skeptical of the exceptions because they could possibly be accounted for by sampling errors in simulation studies or unreported variation in unknown parameters (for example, higher moments or the distributions sampled) and because we anticipate mathematical regularity in the effect of violating a particular ANOVA assumption *ceterus paribus.*

Clearly there are boundary conditions on the conclusions in Table 16. There must surely be some breaking point at which a distribution is so pathologically skewed that nominal levels of significance and power are seriously misleading, for example. Thus, while holding the general conclusions of Table 16 in mind, the prudent data analyst would nonetheless attempt to estimate the skewness, kurtosis, and variances of the populations he has sampled and reference the tables of data presented above in the event that any of these values is extreme.

Analysis of Covariance

The robustness of the analysis of variance to violations of some of its assumptions has led to posing similar questions for its more complex kin, the analysis of covariance (ANCOVA). This approach, a wedding of the analysis of variance and regression analysis, was introduced by R. A. Fisher in 1932 in his classic work, *Statistical Methods for Research Workers.* The analysis of covariance provides a method of achieving increased precision over the corresponding analysis of variance by employing statistical control of the sources of variation not directly controlled by the experimenter. This is accomplished by reducing the observed variation in the dependent variable by an amount which can be accounted for by a concomitant variable or variables. However, the increased precision over the analysis of variance is obtained at the cost of satisfying an additional set of assumptions. An excellent overview of the ANCOVA technique has been presented by Elashoff (1969).

The following assumptions are made in a simple one-way fixed-effects ANCOVA:

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \overline{x}_{..}) + e_{ij} \tag{14}$$

$$e_{ij} \sim NID(0, \sigma^2) \tag{15}$$

$$\Sigma \alpha_j = 0 \tag{16}$$

TABLE 16

Summary of Consequences of Violation of Assumptions of the Fixed-effects ANOVA

	Equal n's		*Unequal n's*	
Type of Violation	*Effect on α*	*Effect on Power*	*Effect on α*	*Effect on Power*
Non-independence of errors	Non-independence of errors seriously affects both the level of significance and power of the F-test regardless whether n's are equal or unequal.			
Non-normality: Skewness	Skewed populations have very little effect on either the level of significance or the power of the fixed-effects model F-test; distortions of nominal significance levels of power values are rarely greater than a few hundredths. (However, skewed populations can seriously affect the level of significance and power of *directional*—or "one-tailed"—tests.)			
Kurtosis	Actual α is less than nominal α when populations are leptokurtic (i.e., $\beta_2>3$). Actual α exceeds nominal α for platykurtic populations. (Effects are slight.)	Actual power is less than nominal power when populations are platykurtic. Actual power exceeds nominal power when populations are leptokurtic. Effects can be substantial for small n.	Actual α is less than nominal α when populations are leptokurtic (i.e., $\beta_2>3$). Actual α exceeds nominal α for platykurtic populations. (Effects are slight.)	Actual power is less than nominal power when populations are platykurtic. Actual power exceeds nominal power when populations are leptokurtic. Effects can be substantial for small n's.
Heterogeneous Variances	Very slight effect on α, which is seldom distorted by more than a few hundredths. Actual α seems always to be slightly increased over the nominal α.	(No theoretical power value exists when variances are heterogeneous.)	α may be seriously affected. Actual α exceeds nominal α when smaller samples are drawn from more variable populations; actual α is less than nominal α when smaller samples are drawn from less variable populations.	(No theoretical power value exists when variances are heterogeneous.)
Combined non-normality and heterogeneous variances	Non-normality and heterogeneous variances appear to combine additively ("non-interactively") to affect either level of significance or power. (For example, the depressing effect on α of leptokurtosis could be expected to be counteracted by the elevating effect on α of having drawn smaller samples from the more variable, leptokurtic populations.)			

$$\beta_1 = \beta_2 = \beta_3 \ldots = \beta \tag{17}$$

x is a fixed mathematical variable, not a stochastic variable. It is measured without error. (18)

The fixed-effects ANCOVA model is parallel to the corresponding model for the analysis of variance, but the values of μ, α_j, and e_{ij} will be different for a given experiment because of the inclusion of a new term in the model. It is the addition of the term $\beta(x_{ij} - \bar{x}_{\cdot\cdot})$, which permits an increase in the sensitivity of tests of significance based on this model. The coefficient, β, is the slope of the linear regression of the dependent variable; y, on the concomitant variable; x; x_{ij} is the value of the concomitant variable for the ith individual of the jth treatment group; and $\bar{x}_{\cdot\cdot}$ is the overall mean of the concomitant variable. The regression term removes from the total variation of y that portion which can be predicted from x. Two results of this process of special interest to researchers in the behavioral sciences are (1) that there is increased precision in the randomized experiment and (2) that bias due to differences in the concomitant variable is removed or reduced.

It is the inclusion of the regression term, $\beta(x_{ij} - \bar{x}_{\cdot\cdot})$ in the model which makes additional assumptions necessary. These include assumptions that derive from a statement of the model, those required to derive quantities with the appropriate expected values, and logical assumptions required to make the results interpretable.

The first assumption, that of additivity, has the same meaning as in the fixed-effects ANOVA model and should be of little concern for the researcher.

The second assumption consists of three components: normality, homogeneity of variance, and independence of error. The formation of a statistic whose distribution on replications approximates the F-distribution requires that the e_{ij}'s are normally and independently distributed with a common variance, σ^2, for all j. These are the familiar assumptions of ANOVA and are required for the same reason. It should be noted that in ANCOVA, the e_{ij}'s are residuals about within-group regression lines with estimated common slope and σ_e^2 is a variance error of estimate.

Cochran (1957) and Winer (1962) indicated that the robustness of the analysis of variance to violations of the assumptions of normality and homogeneous variances carry over into the analysis of covariance. Although neither cited evidence for their conclusions, there is some work that tends to confirm them. Box and Anderson (1962) reach the somewhat surprising conclusion that the sensitivity of the F-test in ANCOVA to departures from normality in the dependent variable depends upon the degree of "non-normality" in the covariate. The quotation marks are used because the covariate values may not be random variables but may in fact be values selected by the experimenter to approximate a normal distribution. Atiqullah (1964), using

an analytic approach somewhat different from Box and Watson, arrived at similar conclusions. Atiqullah also showed that if x is a random variable and is normally distributed, non-normality in the dependent variable has little effect on the F-test.

These results indicate that the analysis of covariance, in the balanced layout, is robust with respect to non-normality in the dependent variable when the concomitant variable approximates normality. When the concomitant variable is a random variable and is not normally distributed, the result is to increase the sensitivity of the F-test to non-normality in the dependent variable. It should be noted that Atiqullah considered only distributions with non-normality reflected in the kurtosis. The effects of the more frequently encountered skewed distributions were not investigated.

There is little evidence, beyond that produced in connection with the analysis of variance, to indicate the effect of heterogeneous variances on the analysis of covariance. Apparently most researchers are willing to accept the statements of Cochran and Winer cited above. It was suggested by Potthoff (1965) that this generalization may be gratuitous. He found that the sensitivity of the covariance analysis to unequal variances in the dependent variable depends on the ratio, $\frac{n_1 \sigma_{x_1}}{n_2 \sigma_{x_2}}$. The more this ratio departs from 1, the greater the sensitivity of the test. It is remarkable that, here, as in the case of non-normality, the robustness of the test is determined by the characteristics of the covariate. Unfortunately, most analytic studies of robustness do not tell us the extent of error in the final probability statements even when they tell us that the analysis is biased. Thus, even though they are less elegant and less general, sampling studies have the advantage of indicating the degree of error in probability statements.

There is no reason to believe that the importance of independence of errors is reduced in the analysis of covariance. It seems that meeting the independence assumptions, unlike many others, is essential for actual probability levels to be reasonably near nominal levels. Unfortunately, there is no entirely adequate measure of the independence of errors within groups. The issue is primarily a logical one and must be met by designing the experiment to assure the desired independence.

The third assumption ($\Sigma \alpha_j = 0$) should be of no concern to the researcher as in the ANOVA fixed-effects model.

Homogeneity of Regression

Examination of the model reveals that β is assumed to be constant for all treatment groups. This is the familiar homogeneity of regression assumption. Interpretations of significant main effects in the presence of heterogeneous regression slopes in the analysis of covariance is similar to making main effect interpretations in the presence of interactions in the factorial analysis of variance set up. In the latter, however, the

possibility of making meaningful statements about the independent variables is greater than in the former. This suggests that if heterogeneous regression slopes are suspected, the concomitant variable might better be used as a blocking variable in the standard factorial analysis of variance. At other times, the investigator may have no *a priori* knowledge of treatment slope interactions and, in this case, knowledge of the effects of possible violations on probability statements is important.

Atiqullah (1964) used analytic methods to investigate the robustness of the F-test of the analysis of covariance to the violation of the assumption of parallel regression slopes. He considered two models, the usual one,

$$L_1 : y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x}_{..}) + e_{ij} \tag{19}$$

and an alternative model reflecting heterogeneous regression slopes,

$$L_2 : y_{ij} = \mu + \alpha_j + \beta_j (x_{ij} - \bar{x}_{..}) + e_{ij}. \tag{20}$$

The investigation was centered on the effects of employing the F-test based on model L_1 when in fact model L_2 holds.

If the following notation is adopted:

$$W_{jj} = \sum_{i}^{n} (x_{ij} - \bar{x}_{.j})^2 , \; W_2 = \sum_{j}^{J} W_{jj}, \tag{21}$$

$$\hat{\beta} = \sum_{i}\sum_{j} y_{ij}(x_{ij} - \bar{x}_{.j})/W_2 , \tag{22}$$

then (using model L_1)

$$\hat{\alpha}_j - \hat{\alpha}_k = \bar{y}_{.j} - \bar{y}_{.k} - \beta(\bar{x}_{.j} - \bar{x}_{.k}) \text{ for } j \neq k \tag{23}$$

$$E(\hat{\alpha}_j - \hat{\alpha}_k) = \alpha_j - \alpha_k, \tag{24}$$

and

$$\text{Var}(\hat{\alpha}_j - \hat{\alpha}_k) = \sigma^2 \left[\frac{2}{J} + \frac{(\bar{x}_{.j} - \bar{x}_{.k})^2}{W_2}\right]. \tag{25}$$

These are the usual covariance results given model L_1 for the expectation and variance of the contrast of two means. In an experiment involving only two groups, the comparable results given model L_2 are

$$E(\hat{\alpha}_1 - \hat{\alpha}_2 \mid L_2) = \alpha_1 - \alpha_2 - (\tfrac{1}{2} W_2) (\bar{x}_{.1} - \bar{x}_{.2}) (W_{11} - W_{22}) (\beta_1 - \beta_2) \tag{26}$$

and

$$\mathrm{Var}(\hat{\alpha}_1 - \hat{\alpha}_2 \mid L_2) = \sigma^2\left[\frac{2}{J} + \frac{(\bar{x}_{.1} - \bar{x}_{.2})^2}{W_2}\right]. \tag{27}$$

It can be seen that the expectations are different in general but are equivalent if the means of the concomitant variable are equal or if the within group sums of squares are equal. However, in the case of the contrast of two means in an experiment with more than two groups, the expectation of the treatment contrast is unbiased only when the means of all groups are the same.

The results of Atiqullah's investigation into the omnibus F-test are limited since they employ asymptotic results requiring the number of treatment groups to approach infinity. He showed, however, that even under asymptotic conditions, the analysis assuming model L_1 is not unbiased under conditions of model L_2. Atiqullah's work is of considerable theoretical interest, but assumptions required for his analysis are such that little light is shed on the practical effects of the bias in typical research applications.

Peckham (1968) investigated the same phenomena using Monte Carlo techniques. He systematically varied population regression slopes for different combinations of number of treatment groups and number within each treatment group. The values of the covariate were chosen to be nearly "normally distributed" and were fixed over all replications of the experiment, that is, fixed for the production of each empirical sampling distribution of the F-statistic. Peckham found that the empirical sampling distribution of the F-statistic differed little from the theoretical sampling distribution unless the departure from homogeneous slopes was extreme. As the degree of heterogeneity increased, the analysis became more conservative with respect to making a Type I error. The results for the first phase of Peckham's study are presented in Table 17.

A second phase of the same investigation indicated that the robustness held for the quasi-experiment in which the treatment groups differed with respect to covariate means. The results for the second phase of Peckham's study are presented in Table 18. One should guard against overgeneralizing from the results of this empirical investigation for two reasons: (1) The variance of X was held constant as β was varied, resulting in unequal σ_e^2 across groups; it is possible, therefore that the two assumption violations had compensating effects. And, (2) since the covariate was fixed over replications, the results of Peckham's study may not hold for experiments in which the covariate is a random variable. Nevertheless, it appears that one is not very likely to make Type I errors due to heterogeneity of regression slopes alone. Further

TABLE 17

Comparison of Actual and Nominal Levels of Significance in Simulations of "True" Experiments

		Actual significance levels corresponding to nominal levels of											
		.200			*.100*			*.050*			*.010*		
J	*Regression coefficients*	n: *5*	*10*	*20*	*5*	*10*	*20*	*5*	*10*	*20*	*5*	*10*	*20*
2	.5, .5	.204	.199	.197	.101	.094	.102	.053	.052	.049	.009	.013	.011
2	.4, .6	.192	.191	.187	.093	.096	.089	.043	.050	.045	.009	.010	.009
2	.3, .7	.179	.203	.195	.088	.091	.097	.042	.045	.051	.008	.011	.009
2	.2, .8	.167	.170	.171	.081	.076	.076	.038	.039	.038	.006	.006	.008
2	.1, .9	.143	.141	.133	.062	.055	.055	.029	.029	.027	.008	.004	.002
3	.5, .5, .5	.189	.205	.209	.100	.092	.103	.048	.045	.049	.008	.008	.006
3	.4, .5, .6	.191	.200	.181	.100	.097	.096	.046	.050	.049	.010	.009	.011
3	.3, .5, .7	.195	.192	.202	.098	.098	.102	.049	.047	.048	.010	.009	.008
3	.2, .5, .8	.167	.169	.170	.083	.079	.086	.038	.032	.043	.007	.006	.010
3	.1, .5, .9	.139	.148	.149	.062	.071	.070	.028	.035	.031	.006	.006	.006
5	.5, .5, .5, .5, .5	.197	.208	.201	.105	.105	.103	.054	.049	.049	.010	.012	.009
5	.4, .4, .5, .6, .6	.189	.197	.202	.092	.094	.101	.045	.048	.048	.009	.007	.007
5	.3, .4, .5, .6, .7	.199	.174	.177	.090	.091	.084	.046	.044	.049	.007	.006	.008
5	.2, .4, .5, .6, .8	.178	.173	.170	.085	.086	.076	.041	.038	.041	.008	.007	.005
5	.1, .3, .5, .7, .9	.154	.148	.165	.078	.073	.083	.035	.041	.044	.005	.010	.011
5	.1, .5, .5, .5, .9	.168			.082			.038			.008		
5	.1, .1, .5, .9, .9	.124			.055			.026			.007		
5	.1, .5, .5, .5, .5	.196			.097			.048			.009		
5	.1, .7, .7, .7, .7	.180			.084			.041			.009		
5	.1, .9, .9, .9, .9	.129			.067			.037			.013		
5	.1, .1, .9, .9, .9	.111			.052			.032			.007		

TABLE 18

Comparison of Actual and Nominal Levels of Significance in Simulations of Quasi-Experiments

J	Regression coefficients	$(\bar{X}_{.1} - \bar{X}_{.2})$	*Actual significance levels corresponding to nominal levels of*							
			.200		.100		.050		.010	
		n:	5	10	5	10	5	10	5	10
2	.5, .5	.25	.190	.207	.095	.101	.043	.053	.012	.011
2	.4, .6	.25	.213	.190	.107	.090	.057	.044	.014	.012
2	.3, .7	.25	.188	.193	.087	.094	.043	.047	.007	.010
2	.2, .8	.25	.167	.183	.082	.084	.038	.039	.008	.008
2	.1, .9	.25	.145	.154	.066	.066	.031	.027	.006	.005
2	.5, .5	.50	.198	.188	.104	.098	.048	.053	.009	.011
2	.4, .6	.50	.199	.201	.105	.100	.051	.048	.011	.008
2	.3, .7	.50	.188	.191	.086	.092	.043	.042	.006	.007
2	.2, .8	.50	.180	.167	.080	.078	.037	.035	.006	.006
2	.1, .9	.50	.156	.145	.075	.064	.034	.032	.007	.005
2	.5, .5	1.00	.204	.207	.099	.103	.041	.050	.008	.008
2	.4, .6	1.00	.196	.200	.092	.099	.044	.053	.010	.007
2	.3, .7	1.00	.176	.196	.087	.094	.041	.046	.007	.008
2	.2, .8	1.00	.176	.171	.083	.088	.044	.039	.009	.008
2	.1, .9	1.00	.144	.133	.063	.065	.032	.033	.007	.007

study is needed to determine the effects of unequal slopes on the power of the analysis of covariance. This could very well be the crucial issue.

Covariate Fixed and Measured Without Error

It may not be immediately clear from inspection of the ANCOVA model (14) that the error term (e_{ij}) represents the unknown component of y_{ij} only. It does not reflect error in measuring x_{ij}, the covariate. The inclusion of the covariate as part of the design matrix implies that inferences based on results of an experiment can be made only over replications in which the elements of the design matrix are identical, that is, the x values are constant.

There are four possible combinations of the two conditions stated in the heading. The covariate values may be: (1) fixed and measured without error, (2) fixed and measured with error, (3) random and measured without error, (4) random and measured with error.

The first condition meets the assumption of the analysis and need not be discussed here. The fourth condition will be considered next. Brownlee (1965, p. 32) pointed out that if the dependent variable y is measured with error e and x is measured with error d, then in terms of deviation scores.

$$y = \alpha + \beta x + (e - \beta d). \tag{28}$$

He then showed that

$$cov(x, e - \beta d) \neq 0, \tag{29}$$

which violates the assumption that the random error is independent of the other terms of the model. This analytic result, however, does not reveal the practical effects of violating the assumption.

Lord (1960) used examples to show how errors of measurement in the concomitant variable can produce misleading results in the standard analysis of covariance procedures. He developed a large sample technique to increase precision in this situation. His development is extended to only two groups and requires two measurements on the concomitant variable for each individual. It is clear that if these conditions prevail or can be obtained, the experimenter would be well advised to follow the procedure outlined by Lord.

Porter (1967) obtained the empirical sampling distribution of the F-statistic using estimated true scores of the concomitant variable rather than the observed score as the covariate. He found that there was a reasonably good fit of the empirical distribution to the theoretical distribution when the correlation between the dependent variable and the concomitant variable was not too large (less than .9) and the reliability of the concomitant variable was between .5 and .9. It appears that, if estimates of the reliability of the concomitant variables are available and lie within the range indicated, the results of Porter's study suggest that his procedure gives results comparable to the more cumbersome approach given by Lord. In addition, Porter's Monte Carlo studies indicate that his procedures generalize to the many-sample case.

Cochran (1968) found that if x, the true value of the covariate, and d, the error component, are independently and normally distributed, then the expected values of the contrast of treatment means will be unbiased if

$$\mu_{x_1} = \mu_{x_2}.$$

Unfortunately, errors of measurement in the covariate may either obscure true differences or create the illusion of differences where they don't exist. Lord (1960) presented two examples with contrasting results. The first illustrated an experiment in which there was a false conclusion that differences exist on the dependent variable which cannot be attributed to the covariate. In the second example, a real difference was not detected by the standard covariance analysis.

Kahneman (1965) pointed out that errors of measurement result in an increase in Type I errors over the corresponding analysis when the covariate is perfectly reliable. He concluded that the analysis of covariance "is suspect of undercorrection whenever prior analysis suggests that real differences among groups exist on the [covariate] unless this variable is identified with very high reliability."

The conclusion derived from these investigations is that, as errors of measurement in x increase, the analysis becomes more like the corresponding analysis of variance. Thus, (1) the increase in precision afforded by the ANCOVA is attenuated and (2) there is less reason to assume that "the groups have been statistically equated" on the concomitant variable. It should be noted, however, that to the extent there is reliability in x, the ANCOVA represents an "improvement" over the corresponding ANOVA.

The second condition (covariate fixed, but measured with error) might prevail if the covariate values were selected by the experimenter but there was error in measurement from one observation to another. Berkson (1950) and Scheffé (1959, pp. 213-216) showed that if the values of the concomitant variable are preselected and gauges or meters are brought to this preselected value so that the expected value of the meter reading is the true value of the concomitant variable, then the desired independence of errors is achieved and the standard regression model is applicable and one is justified in making inferences over the preselected values. This procedure obviously has application in industry and some laboratory sciences but very little in education or the social sciences. Presumably, though, because of the unreliability of x, the effect would be to lower the precision of the analysis for the reasons cited above.

Apparently the third condition (the covariate is a random variable and is measured without error) has not been studied explicitly although the work of Atiqullah (1964) and Cochran (1968) indicated that there is little effect on the F-test when the covariate is a random variable, provided the other assumptions are met.

Additional references: Box and Watson (1962); Porter (1967).

Suggestions for Robustness Simulation Studies

Many studies like those reported here leave several open questions or—at worst—fail to resolve some important questions satisfactorily because of minor flaws in execution or reporting. We offer the following suggestions for improvement of robustness simulation studies:

1. Computer programs should be thoroughly documented. Complete references to library subroutines can be expected to appear in

published reports of the study, and listings or original programs should be available to inquirers.

2. Psuedo-random number generators should be thoroughly tested for performance characteristics. Significance tests should be performed of the goodness-of-fit of the output of such generators to the appropriate theoretical distributions. (See Chen, 1971.)
3. Distributions sampled in a simulation should be described as completely as possible. Preferably these distributions will be described by a mathematical function with all parameters specified. Failing this, at least the first moment about the origin and second, third, and fourth moments about the mean should be reported.
4. "Baseline checks" of the entire simulation procedure should be performed and reported. These test runs are carried out under conditions in which all assumptions of the statistical model under investigation are satisfied; hence, actual and theoretical probabilities should be equal within sampling error.
5. Ideally the complete cumulative (relative) frequency distribution of the test statistic under investigation should be reported for each simulation. Failing this, actual versus nominal comparison should be tabulated at the following theoretical percentiles: 0.5, 1.0, 2.5, 5.0, 10.0, 20.0, 30.0, 40.0, 50.0, 60.0, 70.0, 80.0, 90.0, 95.0, 97.5, 99.0, and 99.5.

   Comparisons with extreme nominal percentiles, for example, 99.95 and greater are unnecessary and potentially misleading and should be discouraged.
6. Inferential tests of the hypothesis of exact correspondence between actual and theoretical distributions of test statistics are unnecessary (because the null hypothesis is *a priori* almost certainly false) and potentially misleading (because they would tend to reject because of lack of fit in the central regions of the distribution which could be irrelevant to the use of extreme percentiles for example, 95, 97.5, 99.5 in actual inferential applications of the test statistic).

   However, some indication of the sampling error in the determination of the percentiles of the actual distribution of the test statistic would be helpful.

Glossary

Actual power -- the true probability of rejecting the null hypothesis (when a certain alternative hypothesis is true) calculated from a knowledge of the manner in which the assumptions underlying the statistical test are violated. To be contrasted with the "nominal power" which is the power calculated by the experimenter in the belief that all assumptions are satisfied.

Actual probability of Type-I error — the true probability of a Type-I error calculated from a knowledge of the manner in which certain assumptions underlying the statistical test are violated. To be contrasted with the "nominal probability of a Type-I error" which is the α set by the experimenter in the belief that all assumptions are satisfied.

α — the probability of a Type-I error (see Type-I error).

ϕ — a function of the non-centrality parameter (v.q.), δ^2, descriptive of the degree of falsity of the null hypothesis:

$$\phi^2 = \delta^2 / J.$$

Exponential distribution — the variable X is said to be *exponentially* distributed if the probability density function of X is $p(X) = \lambda e^{-\lambda x}$ for $X > 0$ and $p(X) = 0$ for $X \leqslant 0$.

Homoscedasticity (literally "equal spread")—the assumption of homoscedasticity in ANOVA is that all groups have equal variance.

Kurtosis — a term used to describe the "peakedness" of a unimodal distribution. One measure of kurtosis is $\beta_2 = E(x - \mu)^4 / \sigma^4$, which equals 3 for the normal distribution, is less than 3 for relatively flat distributions, and is greater than 3 for relatively peaked distributions.

Leptokurtic (see kurtosis) — a term used to describe a "peaked" distribution. ("Lepto" means *thin* or *slender.*) A distribution is said to be "leptokurtic" when $\beta_2 > 3$.

Lognormal distribution — the variable X is said to have a *lognormal distribution* if the probability density function of X is

$$p(X) = (X\sigma\sqrt{2\pi})^{-1} e^{-(\log X - \mu)/(2\sigma^2)}$$

Nominal power — a concept defined when considering the effects of violation of the assumptions of a statistical test. If an experimenter believes that the assumptions underlying his test are met and designs things so that the power of his test against a certain alternative is P (even though in reality the assumptions are not met and the power against that alternative is something other than P), the nominal ("in name only") power is P.

Nominal probability of a Type-I error (nominal significance level) — a concept defined when considering the effects of violation of the assumptions of a statistical test. If an experimenter believes the assumptions underlying a test are met and tests hypotheses at the α level (even though the assumptions are not met and he actually has

some probability other than α of committing a Type-I error), the nominal ("in name only") probability of a Type-I error is α.

Non-centrality parameter — a measure, δ^2, descriptive of the differences among a set of J population means:

$$\delta^2 = n\sum_{1}^{J} (\mu_j - \bar{\mu}.)^2 / \sigma^2 .$$

Platykurtic — a term used to describe a "flat" distribution. A distribution is said to be "platykurtic" if β_2, the kurtosis measure, is less than 3.

Power — the probability of rejecting the null hypothesis given that a particular hypothesis alternative to the null is true. The power of a test is its probability of leading to a Type II error for a particular alternative hypothesis.

Rectangular distribution — the variable X is said to be *rectangularly distributed* if the probability density function of X is $p(X) = 1/(b - a)$ for $a \leqslant X \leqslant b$ and $p(X) = 0$ elsewhere.

Robustness — property (rather vaguely defined) of some statistical tests. A "robust" statistical test preserves the validity of the probability statements applied to it even though the assumptions upon which it is based are violated. The fixed-effects ANOVA F-test is said to be "robust" with respect to heterogeneous variances when n's are equal, for example.

Skewness — the property of symmetry or asymmetry of a distribution. Skewness is usually measured by $E(X - \mu)^3 / \sigma^3$ denoted by $\sqrt{\beta_1}$.

Type-I error — the error of concluding, on the basis of a statistical test, that the null hypothesis is false when it is actually true.

Type-II error — the error of concluding, on the basis of a statistical test, that the null hypothesis is true when it is actually false.

REFERENCES

Atiqullah, M. The estimation of residual variance in quadratically balanced least-squares problems and the robustness of the F-test. *Biometrika*, 1962, 49, 83-91.

Atiqullah, M. The robustness of the covariance analysis of a one-way classification. *Biometrika*, 1964, 51, 365-372.

Atiqullah, M. On the robustness of analysis of variance. *Bulletin of the Institute of Statistical Research and Training*, 1967, 1, 77-81.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 1966, 26, 291-309.

Bartlett, M. S. The effect of non-normality of the *t* distribution. *Proceedings of the Cambridge Philosophical Society*, 1935, **31**, 223-231.

Bartlett, M. S. The use of transformations. *Biometrics*, 1947, **3**, 39-57.

Behrens, W. U. Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landweise Jahresbuch*, 1929, **68**, 807-837.

Berkson, J. Are there two regressions? *Journal of the American Statistical Association*, 1950, **45**, 164-180.

Bhattacharjee, G. P. Non-normality and heterogeneity in two-sample *t* test. *Annals of the Institute of Statistical Mathematics*, 1968, **20**, 239-254.

Boneau, C. A. The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 1960, **57**, 49-64.

Boneau, C. A. A comparison of the power of the *U* and *t* tests. *Psychological Review*, 1962, **69**, 246-256.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 1954, **25**, 290-302. (a)

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, II: Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 1954, **25**, 484-498. (b)

Box, G. E. P., & Anderson, S. L. Robust tests for variances and effect of non-normality and variance heterogeneity on standard tests. *Technical Report* No. 7, Ordinance Project No. TB 2-0001(832), Dept. of Army Project No. 599-01-004, 1962.

Box, G. E. P., & Cox, D. R. An analysis of transformation. *Journal of Royal Statistical Society* (B), 1964, **26**, 211-252.

Box, G. E. P., & Watson, G. S. Robustness to non-normality of regression tests. *Biometrika*, 1962, **49**, 93-106.

Bradley, J. V. Studies in research methodology. IV. A sampling study of the central limit theorem and the robustness of one-sample parametric tests. Behavioral Science Lab., Wright-Patterson Air Force Base, Ohio, March 1963.

Bradley, J. V. Studies in research methodology: VII. The central limit effect for two dozen populations and its correlation with population moments. Behavioral Science Lab., Wright-Patterson Air Force Base, Ohio, December, 1966.

Bradley, R. A. Corrections for non-normality in the use of the two-sample *t*- and *F*-tests at high significance levels. *Annals of Mathematical Statistics*, 1952, **23**, 103-113.

Chen, E. H. A random normal number generator for 32-bit word computers. *Journal of the American Statistical Association*, 1971, **66**, 400-403.

Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, **3**, 22-38.

Cochran, W. G. The comparison of percentages in matched samples. *Biometrika*, 1950, **37**, 256-266.

Cochran, W. G. Analysis of covariance: Its nature and uses. *Biometrics*, 1957, **13**, 261-281.

Cochran, W. G. Errors of measurement in statistics. *Technometrics*, 1968, **10**, 637-666.

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika*, 1967, **32**, 339-353.

Cox, D. R. *Planning of experiments*. New York: Wiley, 1959.

Curtiss, J. H. On transformations used in the analysis of variance. *Annals of Mathematical Statistics*, 1943, **14**, 107-122.

Daniels, H. E. The effect of departures from ideal conditions other than non-normality on the *t* and *z* tests of significance. *Proceedings of the Cambridge Philosophical Society*, 1938, **34**, 321-328.

David, F. N., & Johnson, N. L. The effect of non-normality on the power function of the *F*-test in the analysis of variance. *Biometrika*, 1951, **38**, 43-57.

Donaldson, T. S. Robustness of the *F*-test to errors of both kinds and the correlation between the numerator and denominator of the *F*-ratio. *Journal of the American Statistical Association*, 1968, **63**, 660-676.

Elashoff, J. D. Analysis of covariance: A delicate instrument. *American Educational Research Journal*, 1969, **6**, 383-401.

Fisher, R. A. *Statistical methods for research workers.* Edinburgh: Oliver and Boyd, Ltd., 1932.

Fisher, R. A. The fiducial argument in statistical inference. *Annals of Eugenics*, 1935, **5**, 391-398.

Gaito, J. Non-parametric methods in psychological research. *Psychological Reports*, 1959, 115-125.

Games, P. A., & Lucas, P. A. Power of the analysis of variance of independent groups on nonnormal and normally transformed data. *Educational and Psychological Measurement*, 1966, **26**, 311-327.

Garrett, H. E., & Zubin, J. The analysis of variance in psychological research. *Psychological Bulletin*, 1943, **40**, 233-267.

Gayen, A. K. The distribution of "student's" *t* in random samples of any size drawn from non-normal universes. *Biometrika*, 1949, **36**, 353-369.

Gayen, A. K. Significance of difference between the means of two non-normal samples. *Biometrika*, 1950, **37**, 399-408. (a)

Gayen, A. K. the distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, 1950, **37**, 236-255. (b)

Geary, R. C. The distribution of "student's" ratio for non-normal samples. *Journal of the Royal Statistical Society*, (b), 1936, **3**, 178-184.

Glass, G. V., & Stanley, J. C. *Statistical methods in education and psychology.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Gourlay, N. *F*-test bias for experimental designs in educational research. *Psychometrika*, 1955, **20**, 227-248.

Grant, D. A. On the analysis of variance in psychological research. *Psychological Bulletin*, 1944, **41**, 158-166.

Greenhouse, S. W., & Geisser, S. On methods in the analysis of profile data. *Psychometrika*, 1959, **24**, 95-112.

Gronow, D. G. C. Test for the significance of the difference between means in two normal populations having unequal variances. *Biometrika*, 1951, **38**, 252-256.

Hack, H. R. B. An empirical investigation into the distribution of the *F*-ratio in samples from two nonnormal populations. *Biometrika*, 1958, **45**, 260-265.

Hawkridge, D. G. Designs for evaluative studies. In American Institutes for Research, Evaluative Research, Palo Alto, Calif.: AIR, 1970. Pp. 24-47.

Heerman, E. F., & Braskamp, L. *Readings in statistics for the behavioral sciences.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Hey, G. B. A new method of experimental sampling illustrated on certain non-normal populations. *Biometrika*, 1938, **30**, 68-80.

Horsnell, G. The effect of unequal-group variances on the F-test for the homogeneity of group means. *Biometrika*, 1953, **40**, 128-136.

Hsu, P. L. Contribution to the theory of "student's" *t*-test as applied to the problem of two samples. *Statistical Research Memoirs*, 1938, **2**, 1-24.

Hsu, T. C., & Feldt, L. S. The effect of limitations on the number of criterion score values on the significance level of the *F*-test. *American Educational Research Journal*, 1969, **6**, 515-527.

Kahneman, D. Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin*, 1965, **64**, 326-329.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics, Vol. I.* London: Griffin & Co., 1963.

Kohr, R. L. A comparison of statistical procedures for testing $\mu_1 = \mu_2$ with unequal *n*'s and variances. Doctoral dissertation, Pennsylvania State University, 1970.

Lana, R. E., & Lubin, A. The effect of correlation on the repeated measures design. *Educational and Psychological Measurement*, 1963, **23**, 729-739.

Levene, H. Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics*. Stanford: Stanford University Press, 1960. Pp. 278-292.

Lindquist, E. F. *Design and analysis of experiments in education and psychology*. Boston: Houghton Mifflin, 1953.

Lord, F. M. A theory of test scores. *Psychometric Monograph No. 7*, 1952.

Lord, F. M. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 1960, **55**, 307-321.

Lunney, G. H. A Monte Carlo investigation of basic analysis of variance models when the dependent variable is a Bernoulli variable. *Dissertation Abstracts*, Ann Arbor, Michigan (order no. 68-17, 696), 1968.

Lunney, G. H. Using analysis of variance with a dichotomous variable: An empirical study. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1969.

Lunney, G. H. Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, 1970, **7**, 263-269.

Mandeville, G. K. A Monte Carlo investigation of the adequacy of standard analysis of variance test procedures for dependent binary variates. Ph.D. thesis, Minneapolis: University of Minnesota, 1969.

Mueller, C. G. Numerical transformations in the analysis of experimental data. *Psychological Bulletin*, 1949, **46**, 198-223.

Neave, H. R., & Granger, C. W. J. A Monte Carlo study comparing various two-sample tests for differences in mean. *Technometrics*, 1968, **10**, 509-522.

Nelder, J. A. Response to a paper by G. E. P. Box and D. R. Cox. *Journal of the Royal Statistical Society*, 1964, **26**, 244.

Pearson, E. S. The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika*, 1929, **21**, 259-286.

Pearson, E. S. The analysis of variance in cases of non-normal variation. *Biometrika*, 1931, **23**, 114-133.

Pearson, E. S. Note on Mr. Srivastava's paper on the power function of Student's test. *Biometrika*, 1958, **45**, 429-430.

Pearson, E. S., & Hartley, H. O. (Eds.) *Biometrika Tables for Statisticians*, Vol. I, (3rd ed.). Cambridge, England: Cambridge University Press, 1966.

Pearson, K. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 1897, **60**, 489-498.

Peckham, P. D. An investigation of the effects of non-homogeneity of regression slopes upon the *F*-test of analysis of covariance. Laboratory of Educational Research, Report No. 16, University of Colorado, Boulder, Colorado, 1968.

Pitman, E. J. G. Significance test which may be applied to samples from any populations. III: The analysis of variance test. *Biometrika*, 1937, **29**, 322-335.

Porter, A. C. The effects of using fallible variables in the analysis of covariance, Ph.D. thesis, Madison, Wis.: University of Wisconsin, 1967.

Potthoff, R. F. Some Scheffé-type tests for some Behrens-Fisher type regression problems. *Journal of the American Statistical Association*, 1965, **60**, 1163-1190.

Pratt, J. W. Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 1964, **59**, 665-680.

Quesnel, C. E. The validity of the z-criterion when the variates are taken from different normal populations. *Skandanavian Aktuarietids*, 1947, **30**, 44-55.

Rider, P. R. On the distribution of the ratio of mean to standard deviation in small samples from non-normal populations. *Biometrika*, 1929, **21**, 124-143.

Sawrey, W. L. A distinction between exact and approximate nonparametric methods. *Psychometrika*, 1958, **23**, 174.
Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.
Schlesselman, J. J. Data transformation in two-way analysis of variance. Princeton, N.J.: Educational Testing Service, 1971.
Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.
Srivastava, A. B. L. Effect of non-normality on the power function of the t-test. *Biometrika*, 1958, **45**, 421-429.
Srivastava, A. B. L. Effect of non-normality on the power of the analysis of variance test. *Biometrika*, 1959, **46**, 114-122.
Steel, R. G. D., & Torrie, J. H. *Principles and procedures of statistics.* New York: McGraw-Hill, 1960.
"Student." Errors of routine analysis. *Biometrika*, 1927, **19**, 151-164.
Tukey, J. W. One degree of freedom for non-additivity. *Biometrika*, 1949, **5**, 232-242.
van der Vaart, H. R. On the robustness of Wilcoxon's two-sample test. In H. de Jonge (Ed.), *Quantative methods in pharmacology.* New York: Interscience, 1961. Pp. 140-158.
Welch, B. L. The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 1947, **34**, 28-35.
Welch, B. L. On the comparison of several mean values: An alternative approach. *Biometrika*, 1951, **38**, 330-336.
Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962; 2nd edition, 1971.
Young, R. K., & Veldman, D. J. Heterogeneity and skewness in analysis of variance. *Perceptual and Motor Skills*, 1963, **16**, 588. (a)
Young, R. K., & Veldman, D. J. Heterogeneity and skewness in analysis of variance. Document No. 7498, ADI Auxiliary Publication Office, Photoduplication Service, Library of Congress, Washington, D. C. 1963(b)

AUTHOR

GENE V GLASS, *Address:* Laboratory of Educational Research, University of Colorado; Boulder, Colorado 80302. *Title:* Professor of Education. *Degrees:* B.A., University of Nebraska; M.S. and Ph.D., University of Wisconsin. *Specialization:* Research and Evaluation Methodology.