PSYCHOLOGICAL BULLETIN
Vol. 57, No. 1, 1960

THE EFFECTS OF VIOLATIONS OF ASSUMPTIONS UNDERLYING THE *t* TEST[1]

C. ALAN BONEAU
Duke University

As psychologists who perform in a research capacity are well aware, psychological data too frequently have an exasperating tendency to manifest themselves in a form which violates one or more of the assumptions underlying the usual statistical tests of significance. Faced with the problem of analyzing such data, the researcher usually attempts to transform them in such a way that the assumptions are tenable, or he may look elsewhere for a statistical test. The latter alternative has become popular because of the proliferation of the so-called nonparametric or distribution-free methods. These techniques quite generally, however, couple their freedom from restricting assumptions with a disdain for much of the information contained within the data. For example, by classifying scores into groups above and below the median one ignores the fact that there are intracategory differences between the individual scores. As a result, tests which make no assumptions about the distribution from which one is sampling will tend not to reject the null hypothesis when it is actually false as often as will those tests which do make assumptions. This lack of power of the nonparametric tests is a decided handicap when, as is frequently the case in psychological research, a modicum of reinforcement in the form of an occasional significant result is required to maintain the research response.

Confronted with this discouraging prospect and a perhaps equally discouraging one of laboriously transforming data, performing related tests, and then perhaps having difficulty in interpreting results, the researcher is often tempted simply to ignore such considerations and go ahead and run a *t* test or analysis of variance. In most cases, he is deterred by the feeling that such a procedure will not solve the problem. If a significant result is forthcoming, is it due to differences between means, or is it due to the violation of assumptions? The latter possibility is usually sufficient to preclude the use of the *t* or *F* test.

It might be suspected that one could finesse the whole problem of untenable assumptions by better planning of the experiment or by a more judicious choice of variables, but this may not always be the case. Let us examine the assumptions more closely. It will be recalled that both the *t* test and the closely related *F* test of analysis of variance are predicated on sampling from a normal distribution. A second assumption required by the derivations is that the variances of the distributions from which the samples have been taken is the same (assumption of homogeneity of variance). Thirdly, it is necessary that scores used in

[1] This project was undertaken while the author was a Public Health Service Research Fellow of the National Institute of Mental Health at Duke University. The computations involved in this study were performed in the Duke University Digital Computing Laboratory which is supported in part by National Science Foundation Grant G-6694. The author wishes to express his appreciation to Thomas M. Gallie, Director of the Laboratory, for his cooperation and assistance.

the test exhibit independent errors. The third assumption is usually not restrictive since the researcher can readily conduct most psychological research so that this requirement is satisfied. The first two assumptions depend for their reasonableness in part upon the vagaries inherent in empirical data and the chance shape of the sampling distribution. Certain situations also arise frequently which tend to produce results having intrinsic non-normality or heterogeneity of variance. For example, early in a paired-associate learning task, before much learning has taken place, the modal number of responses for a group will be close to zero and any deviations will be in an upward direction. The distribution of responses will be skewed and will have a small variance. With a medium number of trials, scores will tend to be spread over the whole possible range with a mode at the center, a more nearly normal distribution than before, but with greater variance. When the task has been learned by most of the group, the distribution will be skewed downward and with smaller variance. In this particular case, one would probably more closely approximate normality and homogeneity in the data by using some other measure, perhaps number of trials for mastery. In many situations this option may not be present.

There is, however, evidence that the ordinary t and F tests are nearly immune to violation of assumptions or can easily be made so if precautions are taken (Pearson, 1931; Bartlett, 1935; Welch, 1937; Daniels, 1938; Quensel, 1947; Gayen, 1950a, 1950b; David & Johnson, 1951; Hornsnell, 1953; Box, 1954a, 1954b; Box & Anderson, 1955). Journeyman psychologists have been apprised of this possibility by Lindquist (1953) who summarizes the results of a study by Norton (1951). Norton's technique was to obtain samples of Fs by means of a random sampling procedure from distributions having the same mean but which violated the assumptions of normality and homogeneity of variance in predetermined fashions. As a measure of the effect of the violations, Norton determined the obtained percentage of sample Fs which exceeded the theoretical 5% and 1% values from the F tables for various conditions. If the null hypothesis is true, and if the assumptions are met, the theoretical values are F values which would be exceeded by chance exactly 5% or 1% of the time. The discrepancy between these expected percentages and the obtained percentages is one useful measure of the effects of the violations.

Norton's results may be summarized briefly as follows: (*a*) When the samples all came from the same population, the shape of the distribution had very little effect on the percentage of F ratios exceeding the theoretical limits. For example, for the 5% level, the percentages exceeding the theoretical limits were 7.83% for a leptokurtic population as one extreme discrepancy and 4.76% for an extremely skewed distribution as another. (*b*) For sampling from populations having the same shape but different variances, or having different shapes but the same variance, there was little effect on the empirical percentage exceeding theoretical limits, the average being between 6.5% and 7.0%. (*c*) For sampling from populations with different shapes and heterogeneous variances, a serious discrepancy between theoretical and obtained percentages occurred in some instances. On the basis of these results, Lindquist (1953 p. 86) concluded that "unless the heterogeneity of either form or vari-

ance is so extreme as to be readily apparent upon inspection of the data, the effect upon the F distribution will probably be negligible."

This conclusion has apparently had surprisingly little effect upon the statistical habits of research workers (or perhaps editors) as is evident from the increasing reliance upon the less powerful nonparametric techniques in published reports. The purpose of this paper is to expound further the invulnerability of the t test and its next of kin the F test to ordinary onslaughts stemming from violation of the assumptions of normality and homogeneity. In part, this will be done by reporting results of a study conducted by the author dealing with the effect on the t test of violation of assumptions. In addition, supporting evidence from a mathematical framework will be used to bolster the argument.

To temper any imputed dogmatism in the foregoing, it should be emphasized that there are certain restrictions which preclude an automatic utilization of the t and F tests without regard for assumptions even when these tests are otherwise applicable. It is apparent, for example, that the violation of the homogeneity of variance assumption is drastically disturbing to the distribution of t's and Fs if the sample sizes are not the same for all groups, a possibility which was not considered in the Norton study. It also seems clear that in cases of extreme violations, one must have a sample size large enough to allow the statistical effects of averaging to come into play. The need for such considerations will be made apparent in the ensuing discussion. There is abundant evidence, however, that both the t and the F tests are much less affected by extreme violations of the assumptions than has been generally realized.

A Sampling Experiment

At this point we will concern ourselves with the statement of the results of a random sampling study. The procedure is one of computing a large number of t values, each based upon samples drawn at random from distributions having specified characteristics, and constructing a frequency distribution of the obtained t's. The present study was performed on the IBM 650 Electronic Computer programmed to perform the necessary operations which can be summarized as follows: (a) the generation of a random number, (b) the transformation of the random number into a random deviate from the appropriate distribution, (c) the successive accumulation of the sums and sums of squares of the random deviates until the appropriate sample size is reached, (d) the computation of a t based upon the sums and sums of squares of the two samples, (e) the sorting of the t's on the basis of size and sign and the construction of a frequency distribution based upon the sorting operation. The complete sequence of operations was performed internally, the end result, the frequency distribution of 1000 t's, being punched out on IBM cards.

Comments on many of the above operations are relevant and will be made according to their order above.

(a) The random numbers consisted of 10 digits, the middle 10 digits of the product of the previously generated random number and of one of a sequence of 10 permutations of the 10 digits (0, 1, 2, $\cdots$, 9) placed as multipliers in the machine. To start the process it was necessary to place in the machine a 10 digit random number selected from a table of random numbers. The randomness of numbers generated in such a fashion was checked by sorting 5000 of them into 50 categories on the

basis of the first 2 digits. A χ^2 was computed to determine the fit of the obtained distribution to a theoretical one consisting of 100 scores in each of the 50 categories. The obtained χ^2 of 47.83 is extremely close to the 49.332 value, which is the theoretical median of the χ^2 distribution with 50 degrees of freedom.

(*b*) In order to obtain the random deviates (the individual random scores from the appropriate population), the random numbers obtained in the above fashion were considered to be numbers between 0 and 1 and interpreted as the cumulative probability for a particular score from the prescribed population. From a table entered in the machine, a random deviate having that probability was selected. This is identical with the procedure one uses in entering the ordinary z table to determine the score below which, say, 97.5% of the scores in the distribution lie. The obtained value, 1.96, is the deviate corresponding to that cumulative percentage. The distribution of such deviates from a normal population obtained by using a large random sample of probabilities is normally distributed. Similar tables can be constructed for other populations. The populations selected for this study were the normal, the exponential (J-shaped with a skew to the right) having a density function of $y=e^{-x}$, and the rectangular or uniform distribution. These distributions represent extremes of skewness and flatness to compare with the normal. The tables of deviates corresponding to each of the selected distributions were contained internally in the computer and were so arranged that the mean of each distribution was 0 and the variance 1. To verify these values, population means and variances based on samples of 5000 deviates from each of the three populations were estimated by the usual formulas. The results were for the normal distribution a sample mean of .0024 and a variance of 1.0118, for the exponential a mean of .0128 and a variance of 1.0475, and for the rectangular a mean of −.0115 and a variance of .9812. All of these results could quite easily have arisen from random sampling from distributions having the assumed characteristics. To change the size of the variance of the population, all deviates were multiplied when necessary by a constant, in this case, the number 2. The resulting distribution has a mean of 0 and variance of 4. The only variances used in this study were 1 and 4.

(*c*) The sample sizes selected were 5 and 15.

(*d*) The formula used for the computation of t was the following:

$$t=\frac{M_1-M_2}{\sqrt{\frac{\Sigma X_1^2-N_1M_1^2+\Sigma X_2^2-N_2M_2^2}{N_1+N_2-2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$

where M_1 and M_2 are the means of the first and second samples and N_1 and N_2 are the respective sample sizes. This expression, or an equivalent statement of it, is found in any statistics book and is undoubtedly employed in a preponderance of the research in which a t test involving nonrelated means is used. As pointed out in most statistic texts, this test is not appropriate when variances are different. Tests are available which are more or less legitimate under these conditions, but a certain amount of approximation is involved in them. It was felt, however, that

the ordinary t test might under some conditions be as good an approximation as the more complex forms of t tests and that a verification of this notion was desirable. In addition, the above formula makes use of a pooled estimate of variance for the error term and in this respect is similar to the F test of analysis of variance. Because of this fact, certain results can be generalized from the t to the F test.

To summarize, random samples were drawn from populations which were either normal, rectangular, or exponential with means equal to 0 and variances of 1 or 4. For several combinations of forms and variances, t tests of the significance of the difference between sample means were computed using combinations of the sample sizes 5 and 15. For each of these combinations, frequency distributions of sample t's were obtained on the IBM 650 Electronic Computer.

Results

The results of the sampling study will be presented in part as a series of frequency distributions in the form of bar graphs of the obtained distribution of t's for a particular condition. Upon these have been superimposed the theoretical t distribution curve for the appropriate degrees of freedom. This furnishes a rapid comparison of the extent to which the empirical distribution conforms to the theoretical.

First we shall consider those combinations possible when both of the samples are from normal distributions but variances and sample sizes may vary. Next will be considered the results of sampling from non-normal distributions, but both samples are from the same type of distribution. Finally we deal with the results of sampling from two different kinds of populations, for example, one sample from the normal distribution, and another from the exponential.

Potentially, a very large number of such combinations are possible. Limitations of the time available on the computer necessitated a paring down to a reasonable number. Although the computer is relatively fast when optimally programmed, it nevertheless required almost an hour, on the average, to complete a frequency distribution of 1000 t's. The combinations presented here are those which seemed most important at the time the study was made.

As a measure of the effect of violation of assumptions, the percentage of obtained t's which exceed the theoretical values delineating the middle 95% of the t distribution is used. For 8, 18, and 28 df which arise in the present study, the corresponding values are respectively ± 2.262, ± 2.101, and ± 2.048. If the assumptions are met, and if the null hypothesis of equality of means is true, 5% of the obtained t's should fall outside these limits. The difference between this nominal value and the actual value obtained by sampling should be a useful measure of the degree to which violation of assumptions changes the distribution of t scores. There is, of course, a random quality to the obtained percentage of t's falling outside the theoretical limits. Hence, the obtained value should be looked upon as an approximation to the true value which should lie nearby.

In the figures and in the text, the various combinations of population, variance, and sample size will be represented symbolically in the following form: $E(0, 1)5\text{-}N(0, 4)15$. Here the letters E, N, and R refer to the population from which the sample was drawn, E for exponential-

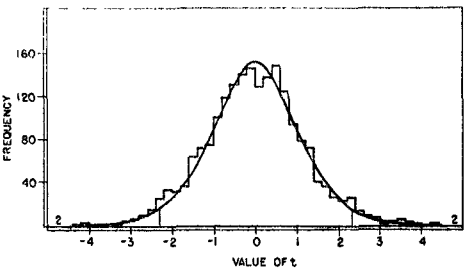

FIG. 1. Empirical distribution of t's from $N(0, 1)5$-$N(0, 1)5$ and theoretical distribution with 8 *df*.

N for normal, and R for rectangular. The first number in the parenthesis is the mean of the population distribution, in all cases zero, while the second number is the variance. The number following the parenthesis is the sample size for that particular sample. In the example above, the first sample is of Size 5 from an exponential distribution having a variance of 1. The second sample is from a normal distribution with variance of 4 and the sample size is 15.

Sampling from Normal Distributions

In order to justify the random sampling approach utilized in this study, and partly to confirm the faith placed in the tabled values of the mathematical statisticians, the initial comparisons are between the theoretical distributions and the obtained distributions with assumptions inviolate. Figures 1 and 2 exhibit the empirical distributions of t's when both samples are taken from the same normal distribution with zero mean and unit variance—designated $N(0, 1)$. In Fig. 1 both samples are of Size 5, while both are 15 in Fig. 2. The theoretical curves, one for 8 *df*, the other for 28, represent quite well the obtained distributions. Ordinates approximately two units from the mean of the theoretical distributions mark off the respective 5% limits for rejecting the null hypothesis. In Fig. 1,[2] 5.3% of the obtained t's fall outside these bounds, while in Fig. 2 only 4.0% of the sample t's are in excess. Since in both cases the expected value is exactly 5%, we must attribute the discrepancy to random sampling fluctuations. The size of these discrepancies should be useful measures in evaluating the discrepancies which will be encountered under other conditions of sampling. For examples of 2000 t's a discrepancy as large as 1% from the nominal 5% value evidently occurs frequently, and for this reason should not be considered as evidence to reject the theoretical distribution as an approximation to the empirical one.

As an initial departure from the simplest cases just presented, Fig. 3 compares theoretical and empirical distributions when samples are taken from the same $N(0, 1)$ population, but the first sample size is 5, the second is 15—that is, $N(0, 1)5$-$N(0, 1)15$. While this in no sense is a violation of the assumptions of the t test, it is interesting to note that again sampling fluctuations have produced an empirical distribution with 4.0% of the t's falling outside the nominal 5% limits.

[2] The numbers in the tails of some of the figures report the number of obtained t's falling outside the boundaries.

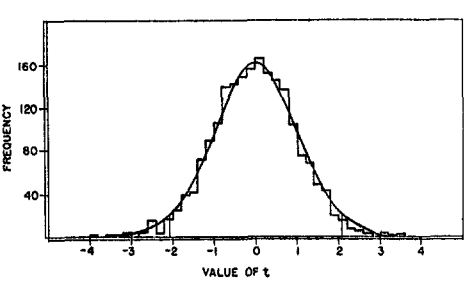

FIG. 2. Empirical distribution of t's from $N(0, 1)15$-$N(0, 1)15$ and theoretical distribution with 28 *df*.

The violation of the assumption of homogeneity of variance has effects as depicted in Fig. 4. Here the obtained distribution is based upon two samples of Size 5, one from $N(0, 1)$ and the other from $N(0, 4)$. The fit is again seen to be close between theoretical and empirical distributions, and 6.4% of the obtained t's exceed the theoretical 5% limits. By increasing the sample size to 15, a distribution results (not shown here) for which only 4.9% of the t's fall outside the nominal limits. It would seem that increasing the sample size produces a distribution which conforms rather closely to the t distribution. As will be seen later, this is a quite general result based upon mathematical considerations, the implications of which are important to the argument. For the moment it is evident that differences in variance at least in the ratio of 1 to 4 do not seriously affect the accuracy of probability statements made on the basis of the t test.

This last conclusion is true only so long as the size of both samples is the same. If the variances are different, with the present set of conditions there are two combinations of variance and sample size possible. In one case the first sample may be of Size 5 and drawn from the population with the smaller variance, while

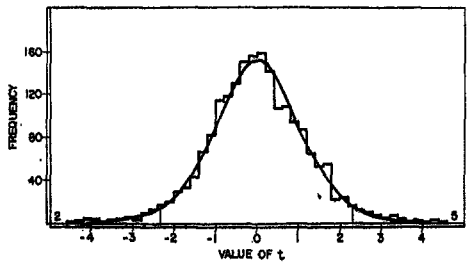

FIG. 4. Empirical distribution of t's from $N(0, 1)5$-$N(0, 4)5$ and theoretical distribution with 8 *df*.

the second sample of Size 15 is drawn from the population having the larger variance—$N(0, 1)5$-$N(0, 4)15$. In the second case the small sample size is coupled with the larger variance, the larger sample size with the smaller variance—$N(0, 4)5$-$N(0, 1)15$. The respective results of such sampling are presented in Fig. 5 and 6. The empirical distributions are clearly not approximated by the t distribution. For the distribution of Fig. 5, only 1% of the obtained t's exceed the nominal 5% values, while in Fig. 6, 16% of the t's fall outside those limits.

There are good mathematical reasons why a difference in sample size should produce such decided discrepancies when the variances are unequal. Recall that $\Sigma(X-M)^2/(N-1)$ is an estimate of the variance of the

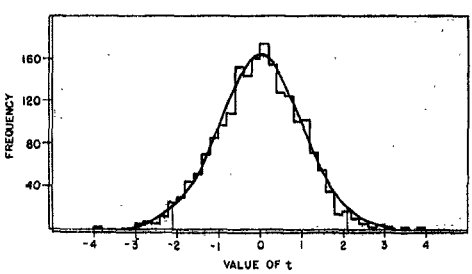

FIG. 3. Empirical distribution of t's from $N(0, 1)5$-$N(0, 1)15$ and theoretical distribution with 18 *df*.

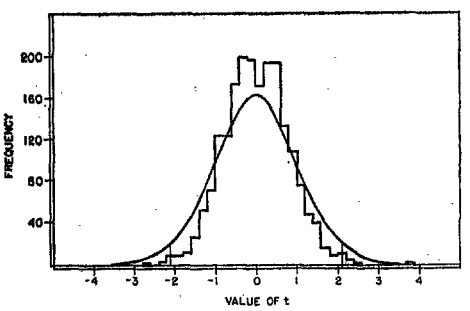

FIG. 5. Empirical distribution of t's from $N(0, 1)5$-$N(0, 4)15$ and theoretical distribution with 18 *df*.

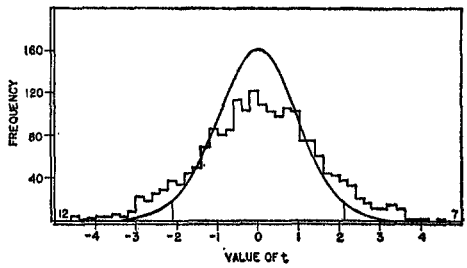

FIG. 6. Empirical distribution of t's from (N(0, 4)5-N(0, 1)15 and theoretical distribution with 18 *df*.

population from which the sample is drawn. Hence, $\Sigma(X-M)^2$ will in the long run be equal to $(N-1)\sigma^2$. The formula used in this study for computing t makes use of this fact and, in addition, under the assumption that the variances of the populations from which the two samples are drawn are equal, pools the sum of the squared deviations from the respective sample means to get a better estimate. That is $\Sigma(X_1-M_1)^2+\Sigma(X_2-M_2)^2$ is an estimate of $(N_1-1)\sigma_1^2+(N_2-1)\sigma_2^2$. If $\sigma_1^2=\sigma_2^2=\sigma^2$ (homogeneity of variance), then the sums estimate $(N_1+N_2-2)\sigma^2$. Hence,

$$\frac{\Sigma(X_1-M_1)^2+\Sigma(X_2-M_2)^2}{N_1+N_2-2} \quad [1]$$

is an estimate of σ^2. If $\sigma_1^2\neq\sigma_2^2$ the estimating procedure is patently illegitimate, the resulting value depending in a large measure upon the combination of sample size and variance used. For example, the case $N(0,1)5$-$N(0,4)15$ has $N_1=5$, $N_2=15$, $\sigma_1^2=1$, $\sigma_2^2=4$, and $N_1+N_2-2=18$. With these values, Formula 1 has an expected value of $[(4\cdot1)+(14\cdot4)]/18=3.33$. Using the appropriate values for the other situation, N(0, 4)5-N(0, 1)15, the result of formula 1 is $[(4\cdot4)+(14\cdot1)]/18=1.67$. This means that on the average, the denominator for the t test will be larger for the first case than for the second. If the sample differences between means were of the same magnitude for the two cases, obviously more "significant" t's would emerge when the denominator is smaller. It so happens that when this latter condition exists, the variance of the numerator also tends to be greater than in the other condition, a fact which accentuates the differences between the two empirical distributions.

Welch (1937) has shown mathematically that in the case of sample sizes of 5 and 15, a state which prevails here, the percentage of t's exceeding the nominal 5% value varies as a function of the ratio of the two population variances and can be as low as 0% and as high as 31.3%. If $N_1=N_2$ there is never much bias, except perhaps in the case in which the sample sizes are both 2. For $N_1=N_2=10$, the expected value of the percentage of t's exceeding the nominal 5% limits varies between 5% and 6.5% regardless of the difference between the variances. For larger sample sizes, the discrepancy tends to be even less.

Since the pooling procedure for estimating the population variance is used in ordinary analysis of variance techniques, it would seem that the combination of unequal variances and unequal sample sizes might play havoc with F test probability statements. That is, a combination of large variance and large sample size should tend to make the F test more conservative than the nominal value would lead one to expect, and, as with the t test, small variance and large sample size should produce a higher percentage of "significant" Fs than expected. These conclusions are based upon a very simple extension to more than two samples of the explanation for the behavior of the

t test probabilities with unequal sample sizes.

A more sophisticated mathematical handling of the problem by Box (1954a) reaches much the same conclusions for the simple-randomized analysis of variance. In a table in his article are given exact (i.e., mathematically determined) probabilities of exceeding the 5% point when variances are unequal. In this case, sampling is assumed to be from normal distributions. If the sample sizes are the same, the probability given for equal sample sizes range from 5.55% to 7.42%, for several combinations of variances, and numbers of samples. If, when variances are different, the samples are of different sizes, large discrepancies from the nominal values result. Combining large sample and large variance lessens the probability of obtaining a "significant" result to much less than 5%, just as we have seen for the t test. In a subsequent article, Box (1954b) presents some results from two-way analysis of variance. Since these designs generally have equal cell frequencies the results are not too far from expected. His figures all run within 2% of the 5% value expected if all assumptions were met.

It would seem then that both empirically and mathematically there can be demonstrated only a minor effect on the validity of probability statements caused by heterogeneity of variance, provided the sizes of the samples are the same. This applies to the F as well as the t test. If however, the sample sizes are different, major errors in interpretation may result if normal curve thinking is used.

Sampling from Identical Non-Normal Distributions: (Equal Variances)

Let us now proceed to violate the other main assumption, that of normality of distribution from which sampling takes place. At this time we will consider the t distributions arising when both samples are taken from the same non-normal distribution. The distributions shown here, and all subsequent ones, are based upon only 1000 t's, and hence will exhibit somewhat more column to column fluctuation than the preceding distributions.

Figure 7 compares the theoretical t distribution and the empirical distribution obtained from two samples of Size 5 from the exponential distribution—$E(0, 1)5$-$E(0, 1)5$. The fit is fairly close, but the proportion of cases in the tails seems less for the empirical distribution than for the theoretical. By count, 3.1% of the obtained t's exceed the nominal 5% values—that is, the test in this case seems slightly conservative. If both sample sizes are raised to 15 (distribution not shown here), the corresponding percentage of obtained t's is 4.0%. While this is probably not an appreciably better fit than for samples of Size 5, we shall see later that there are theoretical reasons to suspect that increasing the sample size should better the approximation of the empirical curve by the theoretical no matter what the parent population may be.

If both samples are of Size 5 from the same rectangular distribution—$R(0, 1)5$-$R(0, 1)5$—the result is as

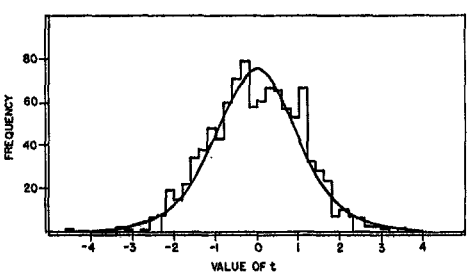

FIG. 7. Empirical distribution of t's from $E(0, 1)5$-$E(0, 1)5$ and theoretical distribution with 8 *df*.

depicted in Fig. 8. The fit of theoretical curve to empirical data here is as good as any thus far observed. The percentage of obtained t's exceeding the 5% values is 5.1% in this particular case. For the case in which the sample sizes are both 15 (not shown here), the fit is equally good, with 5.0% of the cases falling outside of the nominal 5% bounds.

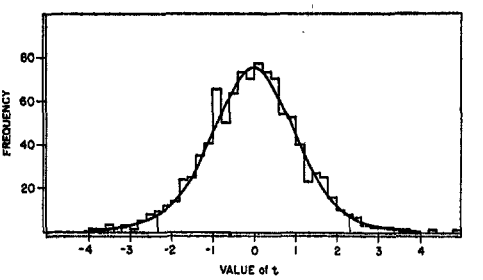

FIG. 8. Empirical distribution of t's from $R(0, 1)5$-$R(0, 1)5$ and theoretical distribution with 8 *df*.

Sampling from Non-Normal Distributions: (Unequal Variances)

We may assume that if the variances are unequal, and at the same time the sample sizes are different, the resulting distributions from non-normal populations will be affected in the same way as the distributions derived from normal populations, and for the same reasons. These cases will not be considered.

If sampling is in sizes of 5 from two exponential distributions, one with a variance of 1, and the other of 4, a skewed distribution of obtained t's emerges (not shown here). We shall discover that a skewed distribution of t's generally arises when the sampling is from distributions which are different in degree of skewness or asymmetry. (For an explanation, see discussion of $E(0, 1)5$-$N(0, 1)5$ below.) Apparently, the effect of increasing the variance of the exponential distribution as in the present case—$E(0, 1)5$-$E(0, 4)5$—is to make the negative sample means arising from the distribution with larger variance even more negative than those from the distribution with smaller variance. In terms of percentage exceeding the nominal 5% limits for this case, the value is 8.3%, of which 7.6% comes from the skewed tail. This combination of variances and distribution was not tested with larger samples, but we shall see when comparing exponential and normal distributions that an increase in the sample size decreases the skew of the obtained t distribution there. Theoretically, this decrease should occur in almost all cases, including the present one.

The result is much less complicated if, while variances are different, the sampling is from symmetrical rectangular distributions—$R(0, 1)$-$R(0, 4)5$. For this small sample situation, (not illustrated), there occurs a distribution of obtained t's having 7.1% of the values exceeding the nominal 5% points. This is roughly the same magnitude as the corresponding discrepancy from normal distributions. For the normal, it will be recalled that an increase of the sample sizes to 15 decreased the obtained percentage to 4.9%. There is no reason to believe that increasing the size of the rectangular samples would not have the same effect. However, time did not permit the determinaiton of this distribution.

Sampling from Two Different Distributions

By drawing the first sample from a distribution having one shape, and by drawing the second from a distribution having another shape (other than shape differences arising from heterogeneity of variance), yet another way has been found to do violence to the integrity of the as-

sumptions underlying the t test. Perhaps the least violent of these happenings is that in which at least one of the populations is normal.

When one sample is from the exponential distribution and the other from the normal, the interesting result shown in Fig. 9 occurs. This is the small sample case—$E(0, 1)5$-$N(0,1)5$. It will be recalled that for skewed distributions the mean and median are at different points. In the exponential distributions, for example, the mean is at the 63rd centile. If samples from the exponential distribution are small, there will be a tendency for the sample mean to be less than the population mean, obviously since nearly two thirds of the scores are below that mean. Since the population mean of the present distributions is 0, the result will be a preponderance of negative sample means for small samples. If the other sample is taken from a symmetrical distribution, which would tend to produce as many positive as negative sample means, the resulting distribution of obtained t's would not balance about its zero point, an imbalance exacerbated by small samples. In Fig. 9, 7.1% of the obtained cases fall outside the 5% limits, with most, 5.6%, lying in the skewed tail. The effect of increasing the sample size to 15 is to normalize the distribution considerably; the resulting curve, Fig. 10, is

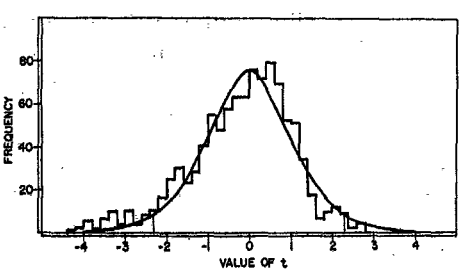

FIG. 9. Empirical distribution of t's from $E(0, 1)5$-$N(0, 1)5$ and theoretical distribution with 8 *df*.

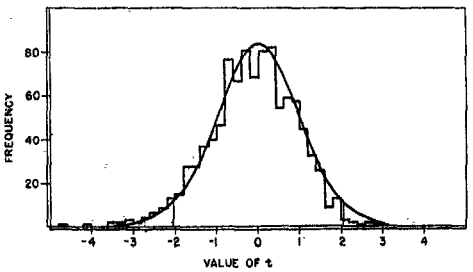

FIG. 10. Empirical distribution of t's from $E(0, 1)15$-$N(0, 1)15$ and theoretical distribution with 28 *df*.

fairly well approximated by the t distribution. One of the tails, however, does contain a disproportionate share of the cases, 4.2% to 0.9% for the other tail, or a total of 5.1% falling outside the nominal 5% limits. Nevertheless, the degree to which the theoretical and empirical distributions coincide under these conditions is striking. It seems likely that if both samples were each of Size 25, the resulting sample distribution of t's would be virtually indistinguishable from the t distribution for 48 *df*, or the next best thing, the normal curve itself. To test this hypothesis, an additional empirical t distribution based on sample sizes of 25 from these same exponential and normal populations was obtained (not shown here). The results nicely confirm the presumption. Comparison with the usual 5% values reveals 4.6% of the empirical t's surpassing them. Whereas with the smaller samples the ratio of t's in the skewed tail to those in the other tail is roughly 80:20, the corresponding ratio for the larger sample case is 59:41. Clearly, the increase in sample sizes has tended to normalize the distribution of t's.

For these conditions, involving rather drastic violation of the mathematical assumptions of the test, the t test has been observed to fare well with an adequate sample size. Such

a state of affairs is to be expected theoretically. By invoking a few theorems of mathematical statistics it can be shown that if one samples from any two populations for which the Central Limit Theorem holds, (almost any population that a psychologist might be confronted with), no matter what the variances may be, the use of equal sample sizes insures that the resulting distribution of t's will approach normality as a limit. It would appear from the present results that the approach to normality is rather rapid, since samples of sizes of 15 are generally sufficient to undo most of the damage inflicted by violation of assumptions. Only in extreme cases, such as the last which involves distributions differing in skew, would it seem that slightly larger sizes are prescribed. Thus it would appear that the t test is functionally a distribution-free test, providing the sample sizes are sufficiently large (say, 30, for extreme violations) and equal.

The distributions arising when sampling is from the normal and the rectangular distributions—$N(0, 1)5$-$R(0, 1)5$ and $N(0, 1)15$-$R(0, 1)15$—would further tend to substantiate this claim. The respective percentages exceeding the 5% nominal values are 5.6% and 4.6% from the empirical distributions for these cases, the distributions of t's being symmetrical and close to the theoretical (not shown).

The only other combination examined in the sampling study is the uninteresting case of exponential and rectangular distributions. This distribution (not shown) is again skewed with the effect of increase of sample size from 5 to 15 to cut down the skew and to decrease the percentage of cases falling outside the theoretical 5% values from 6.4% to 5.6%. For those cases falling outside the nominal 5% values, the ratio is 79:21 for the smaller samples. This is changed to 69:31 for the sample size of 15. Here again it would seem that larger sample sizes would be required to insure the validity of probability statements utilizing the t distribution as a model.

The results of the total study are summarized in Table 1 which gives for each combination of population, variance, and sample size (a) the percentage of obtained t's falling outside the nominal 5% probability limits of the ordinary t distribution, and (b) the percentage of obtained t's falling outside the 1% limits. The combinations are represented symbolically as before. The table is divided into two parts, the first part presenting information on the empirical distributions which are intrinsically symmetrical. The second part is based upon the intrinsically nonsymmetrical distributions, additional information in this section of the table being the percentage of obtained t's falling in the larger of the tails. The percentage for the smaller tail may be obtained by subtraction of the percentage in the larger tail from the total.

Certain implications of the table should be discussed. In the Norton study, more severe distortions sometimes occurred with significance levels of 1% and .1% than appeared with the 5% level. The inclusion in Table 1 of the percentages of obtained t's falling outside the nominal 1% values makes possible the comparison of the 1% and 5% results. The 1% values seem to be approximately what would be expected considering that sampling fluctuations are occurring. It was not felt feasible to determine the results for the .1% level since with only 1000 or 2000 cases the number of obtained t's falling outside the prescribed limits was negligible in most cases. It is pos-

sible, however, that the distortions in the apparent level of significance are more drastic for the smaller α values.

All the results and discussion have been limited thus far to the two-tailed t test. With notable exceptions, the conclusions we have reached can be applied directly to the one-tailed t test as well. The exceptions involve those distributions which are intrinsically *asymmetric* (see Table 1). In these distributions a preponderance of the obtained t's fall in one tail. Depending upon the particular tail involved in the one-tailed test the use of t should produce too many or too few significant results when sampling is from a combination of populations from which an asymmetric t distribution is expected. It seems impossible to make any simple statements about the behavior of the tails in the general case of asymmetric t distribution except to say that such distributions are expected whenever the skew of the two parent populations is different. The experimenter must determine for each particular instance the direction of skew of the expected distribution and act accordingly. Table 1 gives for the intrinsically asymmetric distributions the total percentage of obtained t's falling outside the theoretical 5% and 1% limits and the percentage in the larger tail. From these values can be assessed the approximate magnitude of the bias incurred when a one-tailed test is used in specific situations.

TABLE 1

OBTAINED PERCENTAGES OF CASES FALLING OUTSIDE THE APPROPRIATE TABLED t VALUES FOR THE 5% AND 1% LEVEL OF SIGNIFICANCE

Symmetric Distributions	Obtained Percentage at 5% level	Obtained Percentage at 1% level
$N(0,1)5$-$N(0,1)5$	5.3	0.9
$N(0,1)15$-$N(0,1)15$	4.0	0.8
$N(0,1)5$-$N(0,1)15$	4.0	0.6
$N(0,1)5$-$N(0,4)5$	6.4	1.8
$N(0,1)15$-N$(0,4)15$	4.9	1.1
$N(0,1)5$-$N(0,4)15$	1.0	0.1
$N(0,4)5$-$N(0,1)15$	16.0	6.0
$E(0,1)5$-$E(0,1)5$	3.1	0.3
$E(0,1)15$-$E(0,1)15$	4.0	0.4
$R(0,1)5$-$R(0,1)5$	5.1	1.0
$R(0,1)15$-$R(0,1)15$	5.0	1.5
$R(0,1)5$–R$(0,4)5$	7.1	1.9
$N(0,1)5$-$R(0,1)5$	5.6	1.0
$N(0,1)15$-$R(0,1)15$	5.6	1.1

Asymmetric Distributions	Obtained Percentage at 5% level		Obtained Percentage at 1% level	
	Total	Larger Tail	Total	Larger Tail
$E(0,1)5$-$N(0,1)5$	7.1	5.6	1.9	1.9
$E(0,1)15$-$N(0,1)15$	5.1	4.2	1.4	1.2
$E(0,1)25$-$N(0,1)25$	4.6	2.7	1.3	1.1
$E(0,1)5$-$R(0,1)5$	6.4	5.0	3.3	2.5
$E(0,1)15$-$R(0,1)15$	5.6	3.9	1.6	1.2
$E(0,1)5$-$E(0,4)5$	8.3	7.6	1.7	1.7

DISCUSSION AND CONCLUSIONS

Having violated a number of assumptions underlying the t test, and finding that, by and large, such violations produce a minimal effect on the distribution of t's, we must conclude that the t test is a remarkably *robust* test in the technical sense of the word. This term was introduced by Box (1953) to characterize statistical tests which are only inconsequentially affected by a violation of the underlying assumptions. Every statistical test is in part a test of the assumptions upon which it is based. For example, the null hypothesis of a particular test may be concerned with sample means. If, however, the assumptions underlying the test are not met, the result may be "significant" even though the population means are the same. If the statistical test is relatively insensitive to violations of the assumptions other than the null hypothesis, and, hence, if probability statements refer pri-

marily to the null hypotheses, it is said to be robust. The t and F tests apparently possess this quality to a high degree.

In this particular context, an important example of a test lacking robustness is Bartlett's test for homogeneity of variance (Bartlett, 1937). Box (1953) has shown that this test is extremely sensitive to non-normality and will under some conditions be prone to yield "significant" results even if variances are equal. For example, Box tables a number of exact probabilities of exceeding the 5% normal theory significance level in the Bartlett test for various levels of λ_2, the kurtosis parameter, for different quantities of variances being compared. As an extreme case, if $\lambda_2 = 2$ (i.e., a peaked distribution) with 30 variances being tested, the probability of rejecting the hypothesis at the nominal .05 level is actually .849. If $\lambda_2 = -1$ (i.e., a flat distribution), the probability is .00001. Note that in both these cases, all variances are actually equal. Box, realizing that in the case of equal sample sizes the analysis of variance is affected surprisingly little by heterogeneous variance and non-normality, concludes that the use of the nonrobust Bartlett test to "make the preliminary test on variances is rather like putting out to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" Apparently, as reported in this same article, other commonly used tests for evaluating homogeneity are subject to the same weakness.

We may conclude that for a large number of different situations confronting the researcher, the use of the ordinary t test and its associated table will result in probability statements which are accurate to a high degree, even though the assumptions of homogeneity of variance and normality of the underlying distributions are untenable. This large number of situations has the following general characteristics: (*a*) the two sample sizes are equal or nearly so, (*b*) the assumed underlying population distributions are of the same shape or nearly so. (If the distributions are skewed they should have nearly the same variance.) If these conditions are met, then no matter what the variance differences may be, samples of as small as five will produce results for which the true probability of rejecting the null hypothesis at the .05 level will more than likely be within .03 of that level. If the sample size is as large as 15, the true probabilities are quite likely within .01 of the nominal value. That is to say, the percentage of times the null hypothesis will be rejected when it is actually true will tend to be between 4% and 6% when the nominal value is 5%.

If the sample sizes are unequal, one is in no difficulty provided the variances are compensatingly equal. A combination of unequal sample sizes and unequal variances, however, automatically produces inaccurate probability statements which can be quite different from the nominal values. One must in this case resort to different testing procedures, such as those by Cochran and Cox (1950), Satterthwaite (1946), and Welch (1947). The Welch procedure is interesting since it has been extended by Welch (1951) to cover the simple randomized analysis of variance which suffers the same defect as the t test when confronted with both unequal variance and unequal sample sizes. The Fisher-Behrens procedure suggested by many psychologically oriented statistical textbooks has had its validity questioned (Bartlett, 1936) and, hence, is ignored by

some statisticians (e.g., Anderson & Bancroft, 1952, p. 82).

If the two underlying populations are not the same shape, there seems to be little difficulty if the distributions are both symmetrical. If they differ in skew, however, the distribution of obtained t's has a tendency itself to be skewed, having a greater percentage of obtained t's falling outside of one limit than the other. This may tend to bias probability statements. Increasing the sample size has the effect of removing the skew, and, due to the Central Limit Theorem and others, the normal distribution is approached by this maneuver. By the time the sample sizes reach 25 or 30, the approach should be close enough that one can, in effect, ignore the effects of violations of assumptions except for extremes. Since this is so, the t test is seen to be functionally nonparametric or distribution-free. It also retains its power in some situations (David & Johnson, 1951). There is, unfortunately, no guarantee that the t and F tests are uniformly most powerful tests. It is possible, even probable, that certain of the distribution-free methods are more powerful than the t and F tests when sampling is from some unspecified distributions or combination of distributions. At present, little can be said to clarify the situation. Much more research in this area needs to be done.

Since the t and F tests of analysis of variance are intimately related, it can be shown that many of the statements referring to the t test can be generalized quite readily to the F test. In particular, the necessity for equal sample sizes, if variances are unequal, is important for the same reasons in the F test of analysis of variance as in the t test. A number of the cited articles have demonstrated both mathematically and by means of sampling studies that most of the statements we have made do apply to the F test. It is suggested that psychological researchers feel free to utilize these powerful techniques where applicable in a wider variety of situations, the present emphasis on the nonparametric methods notwithstanding.

REFERENCES

ANDERSON, R. L., & BANCROFT, T. A. *Statistical theory in research.* New York: McGraw-Hill, 1952.

BARTLETT, M. S. The effect of non-normality on the t-distribution. *Proc. Camb. Phil. Soc.*, 1935, **31**, 223–231.

BARTLETT, M. S. The information available in small samples. *Proc. Camb. Phil. Soc.*, 1936, **32**, 560–566.

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proc. Roy. Soc. (London)*, 1937, **160**, 268–282.

BOX, G. E. P. Non-normality and tests on variances. *Biometrika*, 1953, **40**, 318–335.

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. *Ann. of math. Statist.*, 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. *Ann. of math. Statist.*, 1954, **25**, 484–498. (b)

BOX, G. E. P., & ANDERSEN, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. Roy. Statist. Soc. (Series B)*, 1955, **17**, 1–34.

COCHRAN, W. G., & COX, G. M. *Experimental designs.* New York: Wiley, 1950.

DANIELS, H. E. The effect of departures from ideal conditions other than non-normality on the t and z tests of significance. *Proc. Camb. Phil. Soc.*, 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.

GAYEN, A. K. The distribution of the vari-

ance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, 1950, **37**, 236–255. (a)

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

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

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

NORTON, D. W. An empirical investigation of some effects of non-normality and heterogeneity on the *F*-distribution. Unpublished doctoral dissertation, State Univer. of Iowa, 1952.

PEARSON, E. S. The analysis of variance in the case of non-normal variation. *Biometrika*, 1931, **23**, 114–133.

QUENSEL, C. E. The validity of the *Z*-criterion when the variates are taken from different normal populations. *Skand. Aktuarietids*, 1947, **30**, 44–55.

SATTERTHWAITE, F. E. An approximate distribution of estimates of variance components. *Biomet. Bull.*, 1946, **2**, 110–114.

WELCH, B. L. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 1937, **29**, 350–362.

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.

(January 28, 1959)