Journal of the Royal Statistical Society
SERIES B (METHODOLOGICAL)
Vol. XVII, No. 1, 1955

PERMUTATION THEORY IN THE DERIVATION OF ROBUST CRITERIA AND THE STUDY OF DEPARTURES FROM ASSUMPTION*

By G. E. P. BOX and S. L. ANDERSEN

Imperial Chemical Industries, Dyestuffs Division, Blackley, Manchester, 9, and the Institute of Statistics, University of North Carolina.

[Read before the RESEARCH SECTION of the ROYAL STATISTICAL SOCIETY, November 3rd, 1954, Mr. E. C. FIELLER in the Chair]

SYNOPSIS

IN the practical circumstances in which statistical procedures are applied, little is usually known of the validity of assumptions such as the normality of the error distribution. Procedures are required which are "robust" (insensitive to changes in extraneous factors not under test) as well as powerful (sensitive to specific factors under test). Permutation theory, which provides one method for deriving robust criteria, is discussed and applied to the problem of comparing variances.

1. INTRODUCTION

IN this paper attention is confined to problems in hypothesis testing although we believe that much of this discussion applies in other fields.

To fulfil the needs of the experimenter, statistical criteria should

(1) be sensitive to change in the specific factors tested,

(2) be insensitive to changes, of a magnitude likely to occur in practice, in extraneous factors.

A test which satisfies the first requirement is said to be powerful and we shall typify a test which satisfies the second by calling it "robust".

In the derivation of parametric tests (for example by the methods of Neyman and Pearson) it is usual to assume a form of mathematical model involving some specific probability distribution and then carefully to select the form of criterion so as to satisfy the first requirement listed above. Because this procedure does not necessarily result in tests which satisfy the second requirement "non-parametric" tests have been devised in which the form of criterion is selected, usually somewhat arbitrarily, but in such a way that the assumptions which need to be made are of a less specific character. Thus parametric tests tend to satisfy the first requirement listed above (at least when the assumptions are true) but not necessarily the second, whilst non-parametric tests tend to satisfy the second requirement but not necessarily the first. For this reason much research has been conducted on (1) the robustness of various parametric tests and (2) the power of various non-parametric tests.

It is a disconcerting fact that while for one problem a set of apparently restrictive assumptions may lead to a remarkably robust criterion, yet for some other problem the same set of assumptions will give a criterion which is of little value unless the assumptions are very nearly justified. For example, both the analysis of variance test to compare k means and the L_1 test or Bartlett (M_1) test to compare k variances may be derived assuming normally distributed error terms. The very different effects on the null probability that accompany departures from normality are illustrate[illegible] Table 1.

* Sponsored by the Office of Ordnance Research, United States Army under contract DA-36-034-ORD-1177.

TABLE 1

Comparison of Effect of Kurtosis in Tests to Compare k *Means and Tests to Compare* k *Variances*

Departure from Normal Kurtosis	*Percentage Chance of Exceeding Nominal 5 per cent. point when Null Hypothesis True*			
	Comparison of k *means of 10 Observations (Analysis of Variance)**		*Comparison of* k *Variances (Bartlett Test* M_1*)†*	
$\gamma_2 = \beta_2 - 3$	$k = 2$	$k = 20$	$k = 2$	$k = 20$
2	4·74	4·90	16·6	71·8
0	5·00	5·00	5·0	5·0
−1	5·13	5·05	0·56	0·0004

* Taken from Gayen's (1950) values assuming parent Edgeworth population. Similar values are obtained from the method of Section 4.1 of this paper.

† Asymptotic values. Values of similar magnitude are obtained for any sample size for specific populations discussed by Box (1953*a*).

It is clear that the test on variances can be so misleading as to be almost valueless unless we can assert that in most situations met in practice the distribution is very close to the normal. The authors' belief is that such an assertion would certainly not be justified. Published data are comparatively meagre but the frequency distributions given in the older issues of *Biometrika* give little ground for supposing that distributions usually follow the normal law. For example, many of the curves in the Monier-Williams data on percentage butter-fat in milk quoted by Tocher (1928) while looking "reasonably normal" in fact show values of γ_2 between 1 and 2. Similarly of the eight sets of data supplied by Shewart to E. S. Pearson (1931) as typical of observations collected at Bell Telephone laboratories three showed values of γ_2 between 1 and 2.

Many writers including Pearson (1931), Geary (1947), Gayen (1950) and David and Johnson (1951*a*, *c*, 1952) have studied the analysis of variance criterion when the distribution is non-normal. It has been found to be remarkably insensitive to general non-normality.* It has also been shown (Welch, 1937*b*; David and Johnson, 1951*b*; Horsnell, 1953; Box 1954*a* and *b*) that in the commonly occurring case where the group sizes are equal this test is not very sensitive to variance inequalities from group to group. The analysis of variance test for equal group sizes can probably therefore be used with confidence in most practical situations and since we could not expect to obtain a criterion of greater power, unless the nature of the population sampled were specifically known, it may be regarded as fulfilling remarkably well both the requirements of power and robustness. This is perhaps a reason why this test has proved of such great practical utility.

It is clearly desirable that other tests should have properties as satisfactory as those of the analysis of variance test and the problem arises of what is to be done when a standard criterion is found to be unsatisfactorily sensitive to departures from assumption. It has been suggested that before using such a test we should employ one or more preliminary tests to "determine whether the assumptions are justified". This idea seems to us to be a mistaken one, for whether or not a departure from assumption is detected will depend upon the power of the preliminary test which will in turn depend upon the number of observations available. On the other hand, whether or not such a departure is of *importance* depends upon something quite different, namely the robustness of the main test. Thus in some circumstances (for example a preliminary test for equality of variances made before a test to compare means from equally sized groups) a detectable discrepancy might not be large enough to upset the main test, while in other circumstances (for example, a test of normality applied before a Bartlett test on variances) a discrepancy too small to be detected could very seriously upset the behaviour of the main test. It would seem that if this idea of preliminary testing were taken to its logical conclusion we ought to perform another test to check the assumptions made in the preliminary test and so on. We would thus be faced with an endless, and possibly circular, series of tests.

* "General" non-normality is meant to imply that the observations all have the same non-normal parent distribution with possibly different means. This would seem to provide a likely approximation to many experimental situations. Somewhat larger effects have been demonstrated (for example Gayen, 1950) when the distribution is different for observations in different groups.

What are really required are test criteria which 'can stand on their own feet', so that no preliminary testing is necessary. In situations where the standard criterion does not satisfy this requirement some alternative or modified criterion should be sought. One instance of such a modified criterion is that proposed by Cochran (1937) and by Welch (1937*b*, 1951) and James (1951) for the comparison of means in a one-way classification when the variances may differ. As we have noticed the divergencies occurring in the analysis of variance test due to inequality of variance are usually not very serious if the group sizes are equal. However, with unequal groups much larger effects can occur. The standard analysis of variance criterion for the comparison of k means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_t, \ldots, \bar{x}_k$ is of the form $\Sigma w_t(\bar{x}_t - \bar{x})^2$ where the weighting coefficient $w_t = n_t/s^2$ and $\bar{s}^2$ is the pooled estimate of variance. With this criterion the expected values of treatment and error mean squares have the ratio

$$b = 1 + \frac{k(N-1)}{(k-1)N}\left(\frac{\bar{\sigma}^2}{\dot{\sigma}^2} - 1\right)$$

where N is the total number of observations, $\bar{\sigma}^2$ is the unweighted mean variance and $\dot{\sigma}^2$ is the weighted mean variance, the weights being the numbers of degrees of freedom in the groups. When the group sizes are unequal, $\bar{\sigma}^2$ and $\dot{\sigma}^2$ will in general be different, b will differ from unity, and serious bias may be introduced into the comparison. To cure this deficiency weights $w_t = n_t/s_t^2$ are used in the modified criterion which is clearly more appropriate where the variances differ.

The exact null distribution of the modified criterion on the normal assumption has not been found but the approximation supplied by Welch (1951) is probably quite adequate in practice. Since the assumptions on which "exact" distributions are determined are seldom justified in practice, and since in any case the mind cannot appreciate small differences in probability, reasonable approximations to probability distributions are all that are really required.

The modified criterion would be expected to be insensitive to differences in group variances (and by analogy with the standard test) to departures from normality also. Its power has not been investigated but assuming that this is satisfactory it would seem to fill the need for a reliable test to compare means when the variances and sample sizes are different.

Modified tests of this kind are required for other situations. For example:

(1) It has been shown by Geary (1936) that the single-sided t test for the comparison of a sample mean with some hypothetical value is particularly sensitive to skewness in the parent population. Remedial measures have been proposed by Tukey (1948).

(2) It is almost certainly true that certain of the multivariate tests can be grossly misleading under practical conditions, but little has been done to determine the extent of these deficiencies or to remedy them.

(3) Some of the tests proposed for "Poisson variates" will almost certainly be upset in the commonly occurring case where the distribution deviates from the Poisson form.

(4) The tests to compare variances are extremely sensitive to non-normal kurtosis and alternative robust criteria are required.

2. Permutation Tests

A remarkable new class of tests which have since been called permutation tests (or randomisation tests) were introduced by Fisher (1935). After discussing the application of the "paired" t test to some experimental data of Darwin's, Fisher remarked:

> "It seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied."

He went on to show how the null hypothesis could be tested simply by counting how many of the mean differences obtained by rearranging the pairs exceeded the actual mean difference observed. He showed that for the particular set of data he examined, the null probability given by the permutation test and that given by the t test were almost identical.

In connection with a later application of the permutation principle to comparing means of unpaired data, Fisher (1936) said:

> "Actually, the statistician does not carry out this very simple and very tedious process but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

In a discussion of this type of test, E. S. Pearson (1937) emphasized that in a permutation test, as in any other, a choice of the criterion to be used still had to be made. For example, a two-sample permutation test of the type discussed by Fisher could be based on the differences in means, medians, mid-points or any other "position" statistics of the samples. He carried out a sampling experiment performed with a rectangular parent population in which the permutation test based on the differences of the mid-points detected departures from the null hypothesis more often than did the test using the sample means. He thus emphasized that if the permutation test was to be powerful, the choice of criterion would have to depend on the type of alternative hypothesis which the experimenter had in mind.

The points of view expressed by Fisher and Pearson are of course in no way contradictory. Fisher is concerned with the validity of the test of the null hypothesis, while Pearson is concerned with the power of the test when some alternative hypothesis is true.

Two alternative views of the nature of the inference in the permutation test can be taken. These differ in the conception of the population of samples from which the observed sample is supposed to have been drawn. On the first view our attention is confined only to that finite population of samples produced by rearrangement of observations of the experiment. We prefer to adopt the second view which is that the samples are regarded as being drawn from some hypothetical infinite population in the usual way.

It is of some interest to consider the permutation test from the point of view of Neyman-Pearson theory.

2.1. *Permutation Tests from the Point of View of Neyman-Pearson Theory*

In the Neyman-Pearson theory it is supposed that we wish to test some hypothesis, H_0, concerning the nature of the probability law governing N observations, $x_1, x_2, \ldots, x_N$. We have in mind some alternative hypothesis H_1 concerning the nature of this law. To make the test we select a region in the sample space, w, called the "critical region" which is such that, if the sample point is contained in w the null hypothesis will be rejected, and otherwise accepted. The criterion which defines the critical region w is chosen such that

(i) When H_0 is true the chance of rejecting H_0 will always be controlled at some level, α, called the risk of error of the first kind.

(ii) When H_1 is true the chance of rejecting H_0, called the power of the test, will be as large as possible.

Thus, if X_r is a vector of observations $(x_1, x_2, \ldots, x_N)$ and $p_0(X_r)$ and $p_1(X_r)$ are the probability laws for the null hypothesis and alternative hypothesis, we require that w is such that

$$\int_w p_0(X_r)\, dX_r = \alpha \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

$$\int_w p_1(X_r)\, dX_r \text{ is a maximum} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Now usually we do not know what form is appropriate for $p_0(X_r)$, and if we assume some specific function, such as the normal law, we run the risk of the test being valueless under practical circumstances when the real distribution is unknown. The object of the permutation procedure is to satisfy equation (1) with the minimum of assumption.

We can regard the sample X_r as a member of a set X containing the $N!$ samples $X_1, X_2 \ldots, X_r \ldots, X_{N!}$ which contain the same observations as X_r in all possible arrangements. The chance that we shall draw the sample X_r given that it belongs to the set X is

$$p_0(X_r/X) = \frac{p_0(X_r)}{\sum_{r=1}^{N!} p_0(X_r)} = \frac{p_0(X_r)}{p_0(X)} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Thus $p_0(X_r)$ can always be written

$$p_0(X_r) = p_0(X)\, p_0(X_r/X) \quad . \quad . \quad . \quad . \quad . \quad (4)$$

and equation (1) becomes

$$\int_w p_0(X)\, p_0(X_r/X)\, dX_r = \alpha \quad . \quad . \quad . \quad . \quad . \quad (5)$$

To satisfy this equation we now need only assume that $p_0(X_r)$ is some *symmetric* function of the observations $x_1, x_2, \ldots, x_N$. (This would be so if the observations were *independently* and *identically* distributed in any form whatever, or, if each observation were equally dependent on all the others, but not, for example, if the observations were not identically distributed or were serially correlated).

If then $p_0(X_r)$ is a symmetric function, we have $p_0(X_r) = p_0(X_q)$, $(r, q = 1, 2, \ldots N!)$ and

$$p_0(X_r/X) = (N!)^{-1} \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Provided, therefore, that α is such that an integer I exists* for which $I = \alpha N!$ we can ensure that equation (1) is exactly satisfied by arranging that I out of the $N!$ permutations of each set X are contained in w and $N! - I$ are outside w, when we have for the integral in (5)

$$\int_{\Delta} p_0(X) \sum_{r=1}^{I} p_0(X_r/X)\, dX = \frac{I}{N!} \int_{\Delta} p_0(X)\, dX = \alpha \quad . \quad . \quad . \quad (7)$$

where Δ is the entire region of the sample space.

2.2. *More than One Population*

When two or more treatments are compared within n blocks, each containing s observations, we need only assume that within any particular block the probability density function is symmetric. The nature of the functions can differ from block to block. If the vector of observations X_t^j in the j^{th} block is regarded as a member of a set X^j containing the $s!$ samples $X_1^j, X_2^j, \ldots X_t^j \ldots X_{s!}^j$ in all possible arrangements, then as before we can write the null density function in the form

$$p_{0j}(X_t^j) = p_{0j}(X^j)\, p_{0j}(X_t^j/X_j) \quad . \quad . \quad . \quad . \quad . \quad (8)$$

and we have to choose w to satisfy the equation

$$\int_{w} \prod_{j=1}^{n} p_{0j}(X^j) \prod_{j=1}^{n} p_{0j}(X_t^j/X^j) \prod_{j=1}^{n} dX_t^j = \alpha \quad . \quad . \quad . \quad . \quad (9)$$

If we take a vector from each of the n blocks to form a new vector X_r of ns elements

$$X_r = (X_t^1.\ X_u^2, \ldots, X_v^n)$$

where $r = 1, 2, \ldots (s!)^n$; $t, u, v, = 1, 2, \ldots s!$. Then as before we can write the integral in (9) as

$$\int_{w} p_0(X)\, p_0(X_r/X)\, d\,X_r \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

If the within-block distributions are symmetric functions then $p_0(X_r/X) = 1/(s!)^n$ and we can again construct a region w of size α by arranging that $I = \alpha(s!)^n$ out of the $(s!)^n$ within-block permutations of the $ns = N$ observations are contained in w. We have therefore

$$\int_{\Delta} p_0(X) \sum_{r=1}^{I} p_0(X_r/X)\, dX = \frac{I}{(s!)^n} \int_{\Delta} p_0(X)\, dX = \alpha \quad . \quad . \quad . \quad (11)$$

where Δ is the entire region of the sample space.

2.3. *Possible Procedures for Controlling the Value of* α

To maintain α at its nominal value when the null hypothesis is true we note three possible procedures which are, in order of reliability:

* When no such integer exists we cannot maintain the first kind of error exactly at the level α but may use the nearest value α' such that the integer I equals $\alpha' N!$. In practice this will usually be quite satisfactory.

(1) Construct a critical region of size α on the supposition that $p_0(X_r)$ follows some specific distribution function. If then the true distribution is not that assumed we shall be in error by a lesser or greater amount depending on the "robustness" of the criterion which we evolve. We might of course postulate a population sufficiently flexible to cover all the circumstances likely to be met in practice. However, such populations are usually difficult to define and to deal with mathematically and some more specific assumption such that the distribution follows the normal law, is usually made. The value of a test so derived will then depend on factors which are often unforeseeable and certainly its practical utility cannot be assumed.

(2) Assume only that $p_0(X_r)$ belongs to the class of symmetric distribution functions and construct a critical region based on permutation theory. We shall then be in error only if $p_0(X_r)$ does not belong to this very wide *class* of distributions (for example if the observations are serially correlated).

(3) Carry out a process of randomization (in cases where this is possible) so guaranteeing that the effective distribution $p_0(X_r)$ is symmetric and use a critical region based on permutation theory. We cannot now be in error so far as α is concerned unless further disturbances are introduced after the randomization has been performed.

2.4. *Power of the Permutation Test*

Now for each set of vectors X there are of course a very large number of ways in which we can choose the subset of I vectors which are to be included in the critical region. For the single-population test of Section 2.1, for instance, there are $(N!)/\{(N-I)!I!\}$ different ways of doing this.

To obtain a powerful test we should choose the particular subset of I samples so that

$$\int_A p_1(X) \sum_{r=1}^{I} p_1(X_r/X)\, dX \text{ is a maximum} \quad . \quad . \quad . \quad . \quad (12)$$

Now if, in fact, some region is better than another $p_1(X_r)$ cannot be a symmetric distribution. So that in all cases in which we are interested the value of this integral depends upon the form of $p_1(X_r)$. Thus, although we need not specify the particular form of the null distribution, we can only satisfy Neyman and Pearson's second condition if we are prepared to be specific about the class of probability density functions which we have in mind in our alternative hypothesis. If the alternative H_1 is so specified then as has been shown by Lehman and Stein (1949) it is a fairly simple matter to select a best critical region.

For example, suppose that the object of the permutation procedure was to test the hypothesis that each of two samples came from the same distribution against the alternative that they came from two different distributions, one of which had a larger location parameter than did the other.

If we assumed as the specific alternative hypothesis that the observations were drawn from normal populations having the same variance but different means $\mu_2 > \mu_1$ then

$$p_1(X_r) = \text{constant} \times \exp.\left\{-\frac{1}{2\sigma^2}\left(\sum_{\alpha=1}^{n_1}(x_\alpha-\mu_1)^2 + \sum_{\beta=n_1+1}^{n_1+n_2}(x_\beta-\mu_2)^2\right)\right\} \quad . \quad (13)$$

Clearly (12) will be satisfied only if $\sum_{r=1}^{I} p_1(X_r/X)$ is made a maximum for each set X of observations $x_1 \ldots x_{n_1}, x_{n_1+1} \ldots x_{n_1+n_2}$ and since $p_1(X) = \sum_{r=1}^{N!} p_1(X_r)$ is constant for all the samples in X, we have in fact to maximize $\sum_{r=1}^{I} p_1(X_r)$ for each set X.

Now if we consider the value of $p_1(X_r)$ for each of the $N!$ possible rearrangements of the observations we see that it takes its largest value for the $n_1!n_2!$ arrangements which are such that the smallest n_1 observations fall in the first sample and the largest n_2 in the second sample. Equivalently, these are the samples which have the largest difference in means. $p_1(X_r)$ takes its next largest value for the $n_1!n_2!$ samples with the next largest difference in means and so on.

Thus the assumption that the alternative distribution is normal leads to a test in which the I samples included in the critical region are those having the largest differences in group means.* If the distribution had been assumed to be rectangular, a more powerful test would be based on the comparison of mid-points as was found in Pearson's sampling experiment.

It will be noted that if the distribution were really normal then the permutation test would necessarily be less powerful than the normal theory test (i.e., the single-sided t test) since the latter is uniformly most powerful for the normal distribution. This difference in power represents the price we must pay to ensure greater robustness of the criterion in the practical situation where we do not know whether the distribution is normal or not. Some illustration of the amount of power lost (which seems to be remarkably small in the cases considered) is given later in the discussion of tests to compare variances.

2.5. *Rationale for Choice of Alternative Hypothesis*

On permutation theory then we need make no very restrictive assumption concerning the null distribution, but to obtain a "most powerful" criterion for some particular hypothesis we must define the precise alternative we have in mind. The actual test we make is a test of the null hypothesis so that lack of restrictive assumptions when the null hypothesis is true is all that we require to ensure the validity of the test procedure. So far as the practical choice of criterion is concerned, it seems we could argue as follows. Our feeling about the type of distribution likely to be encountered could usually best be expressed in terms of a distribution of possibilities rather than any one possibility. This mental "prior distribution of distributions" might be imagined to have some central value. We should be reluctant to treat this central value as if it were the only one that could occur so far as the null distribution was concerned since this might limit the validity of the test of the null hypothesis to cases where this central distribution really applied. On the other hand, once the validity of the test of the null hypothesis had been safeguarded by the use of permutation theory, it would seem natural to base our criterion on a statistic appropriate for what we supposed to be the central alternative distribution. Even though we might expect this distribution seldom, if ever, to be exactly realized, we would expect that the loss of power suffered in the long run for a series of tests on a series of varying distributions would be smallest for such a statistic.

Using this rationale, the choice in the above example between the difference in means and the difference in midpoints as the appropriate criterion would be based on whether the statistician's mental picture of the distribution of distributions likely to be met in practice in the particular circumstances was centred about the normal or about the rectangular distribution. In most cases the normal distribution would be chosen, though there could be experimental circumstances which would lead the statistician to choose some other distribution as the central one for experiments of a particular type. A test derived in this way would seem to satisfy as nearly as possible the two requirements of Section 1.

We have seen how we are led to a permutation test based on some particular function of the observations. We now consider the properties of such functions when the observations are permuted.

2.6. *Permutation Distribution*

Suppose $g(X_r)$ is a function of the N ordered observations, $x_1, x_2 \ldots x_N$. Then the probability $Pr\{(g(X_r) = g)/X\} = \sum_{g(X_r)=g} p(X_r/X)$ tabulated as a function of g is the "permutation distribution" of $g(X_r)$. Evaluation of the permutation distribution, or of such part of it as is necessary to determine the critical value of the statistic, is laborious. To make the permutation theory of practical value Pitman (1937*a*, *b*) and Welch (1937*a*, 1938) used an approximation to the permutation distribution based on the values of its moments.

* Although there are $N!$ possible arrangements of the sample, there are only $N!/n_1!\,n_2!$ arrangements which result in possibly different mean differences. Thus in practice α would have to be taken so that $I/n_1!\,n_2! = \alpha N!/n_1!\,n_2!$ was an integer.

2.7. *Permutation Moments*

The h^{th} permutation moment of $g(X_r)$ denoted by $\underset{P}{E}\{g(X_r)\}^h$ is the h^{th} moment of the permutation distribution of $g(X_r)$ and is defined by

$$\underset{P}{E}\{g(X_r)\}^h = \sum_{r=1}^{\Omega} \{g(X_r)\}^h \, p(X_r/X) \quad . \quad . \quad . \quad . \quad (14)$$

Where the summation is over all permissible arrangements Ω in number.

2.8. *The Parent Probability Density Function and the Permutation Distribution*

The probability density function of any statistic $g(X_r)$ can be regarded as a weighted aggregate of permutation distributions of $g(X_r)$; for

$$Pr\{g < g(X_r) < g + \delta g\} = \int_{g(X_r)=g}^{g(X_r)=g+\delta g} p(X_r)\, d\, X_r \quad . \quad . \quad . \quad . \quad (15)$$

$$= \int_{g(X_r)=g}^{g(X_r)=g+\delta g} p(X)\, p(X_r/X)\, dX = \int_{\Delta} p(X) \sum_{g(X_r)=g}^{g(X_r)=g+\delta g} p(X_r/X)\, d\, X \quad . \quad . \quad (16)$$

2.9. *Ordinary Moments and Permutation Moments*

As originally indicated by Welch (1937), the h^{th} overall moment of $g(X_r)$ denoted by $E\{g(X_r)\}^h$ may be evaluated by taking the expectation of the h^{th} permutation moment over all values of X; for

$$E\{g(X_r)\}^h = \int_{\Delta} \{g(X_r)\}^h \, p(X_r)\, d\, X_r \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

$$= \int_{\Delta} p(X) \sum_{r=1}^{\Omega} \{g(X_r)\}^h \, p(X_r/X)\, d\, X \quad . \quad . \quad . \quad . \quad (18)$$

$$= \int_{\Delta} p(X) \underset{P}{E}\{g(X_r)\}^h d\, X \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

The above theory may be employed to provide two useful results:

(1) Robust tests may be formulated by approximating to the permutation tests.

(2) The effect on standard test procedures of non-normality and certain other departures from assumption may be evaluated.

3. AN EXAMPLE

It is helpful to study in some detail the following simple example. Suppose that an experiment has been carried out in which n pairs of observations $x_{ti}(t = 1, 2;\ i = 1, 2, \ldots n)$ have been made. One observation within each pair has treatment A applied and the other has treatment B, as in Fisher's first example. This is the familiar situation encountered in the paired t test. It can equivalently be regarded as an example of a randomized block design having n blocks and $s = 2$ treatments with a total of $sn = N$ observations.

The null hypothesis is that within each pair the probability density is unchanged by interchanging the observations. Suppose the alternative hypothesis was that

$$x_{ti} = \alpha_t + \beta_i + z_{ti} \quad . \quad . \quad . \quad . \quad . \quad . \quad (20)$$

where α_t and β_t are treatment and block constants respectively and z_{ti} was a normally distributed random variable with mean zero and constant but unknown variance σ^2. Then using the type of argument indicated in Section 2.4, Lehman and Stein (1949) show that the best permutation critical region is that based on the difference between the sample means, $\bar{x}_1 - \bar{x}_2 = \bar{y}$. The

observed difference in means is referred to the permutation distribution of mean differences generated by all rearrangements within pairs. This is equivalent to the distribution obtained by associating all possible plus and minus signs with the individual differences $y_i = x_{1i} - x_{2i}$.

3.1. *Approximation to the Test*

The same critical region is obtained if we calculate the permutation distribution for the t statistic itself because $t = \{n(n-1)\}^{\frac{1}{2}}\,\bar{y}(\Sigma y^2 - n\bar{y}^2)^{-\frac{1}{2}}$ is a monotonic increasing function of $\bar{y}$. Equivalently we can use the analysis of variance criterion $F = t^2 = \{S_T/1\}/\{S_E/(n-1)\}$ where S_T and S_E are the treatment and error sums of squares respectively. Following Pitman and Welch, it is best to consider the form $W = S_E/(S_E + S_T)$ which in the present example is

$$W = \Sigma(y - \bar{y})^2/\Sigma y^2 = \{1 + t^2/(n-1)\}^{-1}$$

a monotonic decreasing function of t^2 in which the denominator Σy^2 remains constant in all permutations. The permutation moments and normal theory moments for W are as follows:

$$\underset{P}{E}(W) = \frac{n-1}{n} \qquad \underset{N}{E}(W) = \frac{n-1}{n} \quad . \quad . \quad . \quad (21)$$

$$\underset{P}{V}(W) = \frac{2(n-1)}{n^2(n+2)}\left(1 - \frac{b_2 - 3}{n-1}\right) \qquad \underset{N}{V}(W) = \frac{2(n-1)}{n^2(n+2)} \quad . \quad . \quad . \quad (22)$$

The mean of the permutation distribution of W is the same as that for normal theory. The variance differs from that of normal theory by the inclusion of a term of order n^{-1}, involving the sample value of the fourth moment ratio which is defined as

$$b_2 = (n+2)\,\Sigma y^4/(\Sigma y^2)^2 \quad . \quad . \quad . \quad . \quad . \quad (23)$$

so that $E(b_2) = 3$ when the y's are normal. It will be recalled that for normal theory, W follows a Beta distribution $Pr(W < W_0) = I_{w_0}(\frac{1}{2}\nu_2, \frac{1}{2}\nu_1)$ where ν_1 and ν_2, the degrees of freedom of the distribution, are in this case $\nu_1 = 1$ and $\nu_2 = n - 1$.

The permutation distribution of W is of course discontinuous. However, its value lies between zero and one and Pitman (1937) has shown that its third and fourth moments agree reasonably closely with those of the Beta distribution. It is therefore reasonable to approximate the permutation distribution by a Beta distribution, equating the first two moments of the two distributions.

For a Beta distribution with mean and variance μ_1 and μ_2 and degrees of freedom ν_1 and ν_2

$$\nu_1 = \frac{(1-\mu_1)}{\mu_1}\nu_2 \text{ and } \nu_2 = \frac{2\mu_1(\mu_1 - \mu_1^2 - \mu_2)}{\mu_2} \quad . \quad . \quad . \quad (24)$$

By substituting the values of the permutation moments for μ_1 and μ_2 a Beta distribution is obtained with modified degrees of freedom which approximates the permutation distribution Since ν_1/ν_2 involves μ_1 only which in this example, and in all the other examples we consider, is the same for permutation theory as for normal theory, both degrees of freedom in the approximation are multiplied by the same factor d.

In the present case

$$d = 1 + \frac{b_2 - 3}{n\{1 - b_2/(n+2)\}} \text{ or to order } n^{-1}\ d = 1 + \frac{b_2 - 3}{n} \quad . \quad . \quad (25)$$

We can now transform back to the t or F form and finally we have that, as an approximation for the permutation test, we should perform the usual t test or F test but, instead of employing 1 and $n-1$ degrees of freedom, we should use d and $d(n-1)$ degrees of freedom, where d is given by equation (25).* Thus a test is provided which, unlike the full permutation test, is readily carried out in practice.

We see that b_2 occurs in (25) only to order n^{-1} and that consequently for moderate or large values of n the effect of the modifying factor is negligible.

* The large number of approximate procedures which employ an F statistic with modified degrees of freedom would seem to justify a new table in which the significance points were tabulated at fractional values of ν_1 and ν_2 paying particular attention to the lower values.

3.2. *Comparison of Critical Regions*

Although in many practical cases the modified test would differ only slightly from the normal theory test, it is nevertheless instructive to consider in what ways the two tests differ. This can best be done by studying the relative shapes of the critical regions. It is only possible to compare the critical regions geometrically for n, the number of pairs of observations, as large as three. The permutation test would, of course, not be of any real value for so small a number of observations since the permutation distribution contains only eight distinct values and certainly if the modified test had to be justified only on the grounds of an approximation to the permutation test, it too would be of little value. However, we shall see later that the modified criterion can to some extent be justified independently of its approximation to the permutation test. In any case, the general tendencies shown for this case are preserved in a somewhat less extreme form when the sample size is larger. In Fig. 1 the regions are compared. It is supposed that the tests are single-sided and are for an increase in mean at the 5 per cent. level of significance. Both regions then lie entirely in the octant of the sample space shown in the diagram in which all the signs are positive. Since both tests are independent of scale the critical regions are necessarily conical. Sections of the cones are shown on the plane $y_1 + y_2 + y_3 = 30$ in which the mean $\bar{y}$ is equal to ten units.

The critical region for the normal theory t test lies within the cone having its apex at the origin and the circular section shown on the plane $\Sigma y = 30$. The critical region for the modified test (the approximate permutation test) lies within the cone having its apex at the origin and the propeller-like section shown on the plane $\Sigma y = 30$. The types of differences which are found between the tests are illustrated by the samples P and Q corresponding to the points (0, 15, 15) and (5, 5, 20). These samples are shown diagrammatically at the base of the figure. Both samples are of equal significance using the t test and fall outside the 5 per cent. critical region. With the modified test Q is significant and P not significant at the 5 per cent. point.

	Sample			t *Test*	*Approximate Permutation Test*
	y_1	y_2	y_3	(*per cent.*)	(*per cent.*)
P	0	15	15	9·2	12·0
Q	5	5	20	9·2	4·3

We see that the permutation test tends to select as "significant" samples like Q at the expense of those like P.

3.3. *An Alternative Derivation of the Modified Test*

The full permutation test ensures that a proportion is selected from each set of samples for which the observations $y_1, y_2, \ldots y_n$ are numerically the same but with possibly different signs attached. Those selected have the largest means. Now the form of modified test tells us that, as an approximation, we can ignore all properties of the sample except Σy, Σy^2 and Σy^4. Thus approximately the modified test selects a proportion α from each set of samples for which Σy^2 and Σy^4 are constant again choosing those samples with the largest means.

Now a test which is exactly of this type can be derived independently of permutation theory. First, following Neyman and Pearson (1933) let us sketch a rather more general derivation than that which is usually given for the ordinary t test. If the null distribution of the y's follows any spherical distribution, that is to say if*

$$p_0(y) = p_0(y_1, y_2, \ldots y_n) = f(\Sigma y^2) \qquad 0 < \Sigma y^2 < L \quad . \quad . \quad . \quad (26)$$

(where L may be infinite) then we can choose a region w for which, whatever be $f(\Sigma y^2)$

$$\int_w p_0(y) = \alpha \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

* This is the only assumption (Box, 1952, 1953*b*) that need be made in the derivation of those normal theory criteria, which are independent of scale such as the t and F tests, the tests of normality, and the Bartlett test.

by combining together sub-regions of size α on every region of the sample space for which Σy^2 is constant (that is to say on spherical shells centred at the origin). A best critical region may now be obtained by choosing each sub-region of size α to contain the maximum probability density when $E(y_i) = \eta > 0$. If the distribution of the y's about η is assumed to be normal or more

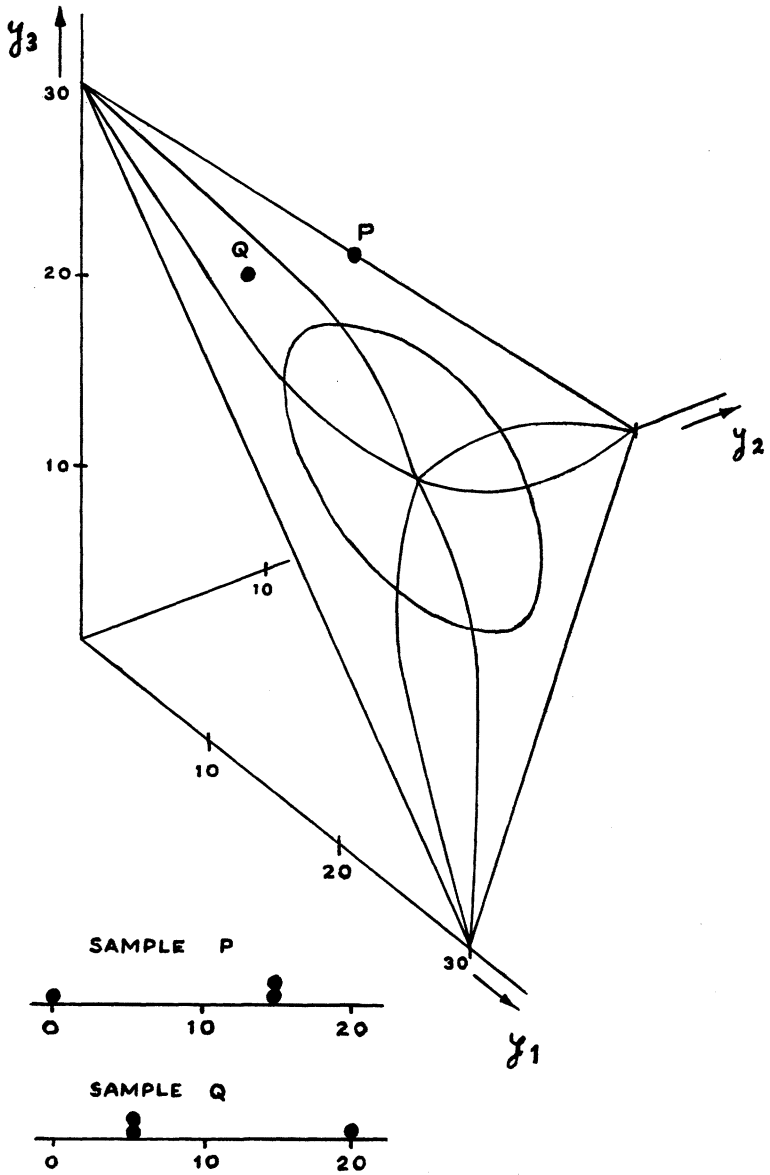

FIG. 1.—Comparison of critical regions for approximate permutation test and normal theory test.

generally spherical with $f(\Sigma y^2)$ a decreasing function of Σy^2 then (as can be seen geometrically and may be proved following the lines given by Neyman and Pearson for the normal distribution) the best critical region is built up by those sub-regions of the spherical shells for which $\bar{y}$ is largest (i.e., for which $\bar{y} >$ constant where the constant is suitably chosen so that the size of each sub-region is α).

If now, instead of supposing that the null distribution $p_0(y)$ is a function of Σy^2 only, we suppose that it is a function of Σy^2 and Σy^4, so that

$$p_0(y) = f(\Sigma y^2, \Sigma y^4) \quad . \quad . \quad . \quad . \quad . \quad . \quad (28)$$

then we can choose a critical region for which (27) is true by compounding sub-regions of size α on every region of the sample space for which Σy^2 is constant and Σy^4 is constant also. If the alternative hypothesis was the same as before, namely, that $E(y_i) = \eta > 0$ and the distribution of the y's about η was normal or followed a decreasing spherical distribution function, then we should obtain a best critical region by including in it those parts of the regions for which Σy^2 and Σy^4 were constant which gave the largest values of $\bar{y}$. (That is to say those parts for which $\bar{y} >$ constant, again choosing the constant so that a part of size α was taken on each sub-region.)

To picture the form of critical region more clearly imagine the case of three observations. Σy^2 is constant on a sphere. Suppose we draw on this sphere the lines of $\Sigma y^4 =$ constant. We now take a fraction α of each of these lines choosing the parts so that we include the largest possible values of $\bar{y}$. The region so obtained will be similar to that obtained for the approximate permutation test illustrated in Fig. 1.

3.4. *Use of the Theory to Determine the Effect of Departures from Assumptions for Standard Statistical Tests*

In equations (21) and (22) for the particular example of the "paired t test" the mean and variance of the permutation distribution of W and of the normal theory distribution of W were compared. As already noted we may obtain the overall (i.e., the ordinary) moments of any function of the observations by taking the expectation of the permutation moments for that function over all samples. In the present case we have, therefore, that the ordinary moments of the W criterion for the paired t test when the distribution is not necessarily normal are

$$E(W) = \frac{n-1}{n} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (29)$$

$$V(W) = \frac{2(n-1)}{n^2(n+2)}\left\{1 - \frac{E(b_2 - 3)}{n-1}\right\} \quad . \quad . \quad . \quad . \quad . \quad (30)$$

Using the Beta approximation as before we see that *for any parent population* which is such that the joint density function of the sample is symmetric in the observations the approximate probability in the paired t test will be given by referring the usual t or F criteria to tables with δ and $\delta(n-1)$ degrees of freedom where

$$\delta = 1 + \frac{E(b_2 - 3)}{n\{1 - E(b_2)/(n+2)\}} \text{ or to order } n^{-1} \quad \delta = 1 + \frac{E(b_2 - 3)}{n} \quad . \quad . \quad (31)$$

We are thus provided with an extremely simple and readily appreciated method for assessing the approximate effect of non-normality.

We are now approximating to the ordinary (continuous) distribution of W when the parent population may not be normal. We should expect the Beta approximation to provide an even better representation of the curve in this case than it does for the (discrete) permutation distribution. We shall see that the approximation does represent the permutation distribution remarkably well even under very extreme conditions and we shall therefore expect this approximation to be adequate in the present case. It may be noted that when the distribution is normal $E(b_2) = 3$ and the approximation is exact.

To obtain numerical results we need to calculate the value of $E(b_2)$ for non-normal parent distributions. Except when n is very large or the distribution is normal, $E(b_2)$ differs from the population value β_2. The sort of values to be expected can be seen from some sampling results of E. S. Pearson (1935). He found that in sampling from curves of the Pearson type the sample values of b_2 tended to be heavily biased. His definition of b_2 differs slightly from ours but the results indicate the general trend. Thus for a population with $\sqrt{\beta_1} = 0{\cdot}7$ and $\beta_2 = 3{\cdot}75$ the mean of six values for b_2 for samples of 50 was $2{\cdot}81$ while for a symmetric population with $\beta_2 = 7{\cdot}1$ the mean of ten values of b_2 for samples of 76 was $4{\cdot}61$.

We can proceed following Fisher (1928) by expanding the denominator of b_2 and taking expectations when we obtain to order n^{-2}

$$E(b_2 - 3) = \gamma_2' - n^{-1}(2\gamma_4' - 3\gamma_2'^2 + 11\gamma_2') + n^{-2}(3\gamma_6' - 16\gamma_4'\gamma_2' + 15\gamma_2'^3 + 38\gamma_4' - 3\gamma_2'^2 + 86\gamma_2') \quad (32)$$

where $\gamma'_{r-2} = \kappa_r(y)/\{\kappa_2(y)\}^{\frac{1}{2}r}$ are the standardized cumulants of the y's (*the differences* of the original observations). In terms of the standardized cumulants of the original observations x_{ti} we find

$$E(b_2 - 3) = \tfrac{1}{2}\gamma_2 - \tfrac{1}{4}n^{-1}(2\gamma_4 - 3\gamma_2^2 + 22\gamma_2) + \tfrac{1}{8}n^{-2}(3\gamma_6 - 16\gamma_4\gamma_2 + 15\gamma_2^3 + 76\gamma_4 - 6\gamma_2^2 + 344\gamma_2) \quad . \quad (33)$$

Calculations of this sort are greatly facilitated by the use of the tables of symmetric functions provided by David and Kendall (1949).

4. Tests to Compare Means

We have seen how approximate permutation theory as developed by Welch and Pitman may be employed in two distinct ways:

(1) to provide a test which is "robust" in the sense that the null probability is approximately correct provided only that the null distribution is a symmetric function of the observations (a requirement which may often be guaranteed by randomization);

(2) to provide an additional method for assessing the consequences of departure from assumptions in which the effects are shown in the convenient and readily comprehended form of a modification in degrees of freedom in the standard tests.

It is instructive to consider these two applications together and thus to study the nature of the "correction factors" which the permutation tests apply to the standard procedures. We do this first for the comparison of means using the results of Pitman and Welch and then, in the next section, apply the theory to tests on variances.

The correction factors d, by which the degrees of freedom must be multiplied in the one way classification analysis of variance test and in the randomized block tests, are set out below.

4.1. *One-way Classification Analysis of Variance*

4.1.1. *Approximate Permutation Test*

(a) *Equal Groups*

Suppose there are s groups of n observations and $sn = N$, then

$$d = 1 + \frac{N+1}{N-1}\frac{c_2}{N-c_2} \text{ or to order } N^{-1}\ d = 1 + \frac{c_2}{N} \quad . \quad . \quad . \quad (34)$$

(b) *Unequal Groups*

Suppose there are s groups with n_t observations in the t^{th} group and $\Sigma n_t = N$, also x_{ti} is the i^{th} observation in the t^{th} group ($i = 1, 2, \ldots n_t$; $t = 1, 2, \ldots s$). Then

$$d = 1 + \frac{N+1}{N-1}\frac{c_2}{(N^{-1}+A)^{-1} - c_2} \quad A = \frac{N+1}{2(s-1)(N-s)}\left(\frac{s^2}{N} - \Sigma\frac{1}{n_t}\right) \quad . \quad . \quad (35)$$

where

$$c_2 = k_4/k_2^2$$

and k_4 and k_2 are k statistics for the whole sample

$$k_4 = \{N(N+1)S_4 - 3(N-1)S_2^2\}/(N-1)(N-2)(N-3),$$

$$k_2 = S_2/(N-1) \text{ and } S_r = \sum_{t=1}^{s}\sum_{i=1}^{n_t}(x_{ti} - \bar{x})^r \quad . \quad . \quad . \quad . \quad . \quad (36)$$

4.1.2. *Effect of Non-normality on the Null Distribution*

The effect of general non-normality is represented by a modification of the degrees of freedom by the factor δ obtained by substituting $E(c_2)$ for c_2 in the formula for d. To order N^{-2}

$$E(c_2) = \gamma_2 - N^{-1}\{2\gamma_4 - 3\gamma_2^2 + 10\gamma_2 + 12\gamma_1^2\} + N^{-2}\{3\gamma_6 - 16\gamma_4\gamma_2 + 15\gamma_2^3 + 36\gamma_4 + 120\gamma_3\gamma_1 - 88\gamma_1^2\gamma_2 + 66\gamma_2 + 204\gamma_1^2\} \quad (37)$$

Since the modifying factor in the permutation test is of the approximate form $d = 1 + c_2/N$ we see that, for this case of comparison of means the normal theory test provides a close approximation to the permutation test for all but very small values of N. This fact justifies its use, as Fisher pointed out. It follows that the effect of non-normality, obtained by replacing c_2 by $E(c_2)$, is also small. The following table shows a number of calculated values for various levels of γ_1^2 and γ_2 (*a*) assuming the parent population is a Pearson curve and (b) assuming as does Gayen (1950) that the parent population follows the Edgeworth series

$$p(x) = \varphi(x) - \varphi^{(3)}(x)\,\gamma_1/6 + \varphi^{(4)}(x)\,\gamma_2/24 + \varphi^{(6)}(x)\,\gamma_1^2/72$$

where $\varphi(x)$ is the normal function and $\varphi^{(r)}(x)$ its r^{th} derivative. For the former distributions we can find the values of the higher moments and hence the higher cumulant ratios using Pearson's recurrence formula. For the Edgeworth series we have $\gamma_3 = 0$, $\gamma_4 = 0$, $\gamma_5 = -35\gamma_1\gamma_2$, $\gamma_6 = -35\gamma_2^2$.

TABLE 2

Effect of Departures from Normality on the Null Probability in the Analysis of Variance Test for 5 Groups of 5 Observations

Percentage Chance of Exceeding Nominal 5 per cent. point when Null Hypothesis is True

Measure of Skewness $\gamma_1^2 = \beta_1$		*Measure of Kurtosis* $\gamma_2 = \beta_2 - 3$				
		−1	−0·5	0	0·5	1
0·0	Pearson curve . .	5·29	5·13	5·00	4·85	*
	Edgeworth series .	5·26	5·12	5·00	4·88	4·77
	Gayen's result . .	5·24	5·12	5·00	4·88	4·76
0·5	Pearson curve . .	5·24	5·10	5·00	4·90	*
	Edgeworth series .	5·27	5·14	5·02	4·92	4·82
	Gayen's result . .	5·29	5·17	5·05	4·93	4·81
1·0	Pearson curve . .	5·16	4·99	4·89	4·84	4·79
	Edgeworth series .	5·31	5·17	5·05	4·96	4·87
	Gayen's result . .	5·34	5·22	5·10	4·98	4·86

* More terms would be needed in the asymptotic series to give reliable values for $E(c_2)$ in this region.

We note the good agreement obtained between the results of the present technique and the results of Gayen (who used an entirely different method) when the same type of parent distribution is assumed. In view of the known shortcomings of the Edgeworth series in regions not close to the normal curve it is also of some interest to compare the Edgeworth results with those obtained assuming a Pearson type curve.

4.2. *Randomized Blocks Test*

4.2.1. *Approximate Permutation Test*

Assuming s observations per block with n blocks and $sn = N$, with x_{ti}, the t^{th} observation in the i^{th} block ($i = 1, 2, \ldots n$; $t = 1, 2, \ldots s$) then

$$d = 1 + \frac{(ns - n + 2)\,V_2 - 2n}{n(s-1)(n - V_2)} \quad \text{or to order } n^{-1}\ d = 1 + \frac{V_2}{n} \quad . \quad . \quad (38)$$

where

$$V_2 = (n-1)^{-1} \sum_{i=1}^{n} (s_i^2 - \bar{s}^2)^2/(\bar{s}^2)^2$$

is the square of the coefficient of variation of the sample block variances,

$$s_i^2 = \sum_{t=1}^{s} (x_{ti} - \bar{x}_i)^2/(s-1)$$

is the sample variance in the i^{th} block and

$$\bar{s}^2 = \sum_{i=1}^{n} s_i^2/n.$$

4.2.2. *Effect of Departures from Assumption on the Null Distribution*

To determine the effects of departures from assumptions the factor δ modifying the degrees of freedom is obtained by substituting $E(V_2)$ for V_2 in the factor for d. If we assume normality and equal variances it is readily shown that $E(V_2) = 2n/(ns - n + 2)$ whence for this case δ is equal to 1 as it should be. We have not had to assume that the distribution within each block is the same so it is possible to determine the effects due to possibly different populations in the blocks We here consider two particular cases, (a) the effect of unequal block variances assuming normality and (b) the effect of non-normality assuming equal block variances, by evaluating approximately the values of $E(V_2)$ appropriate to these two assumptions:

(a) *Inequality of Block Variances Assuming Normality*

If $E(s_i^2) = \sigma_i^2$, $\sum_{i=1}^{n} \sigma_i^2/n = \bar{\sigma}^2$ and $C_r = n^{-1} \sum_{i=1}^{n} (\sigma_i^2)^r/(\bar{\sigma}^2)^r$ then by expanding the denominator of V_2 and taking expectations we have the following expression which is taken to terms as high as $\{n(s-1)\}^{-4}$ so that the examples in Table 3, in which n and s are small, may be studied.

$$E_1(V_2) = \frac{2n}{(ns-n+2)} + \frac{n(s+1)}{(n-1)(s-1)} \times \left(a_1 - \frac{2}{n(s-1)} a_2 + \frac{4}{n^2(s-1)^2} a_3 - \frac{8}{n^3(s-1)^3} a_4 + \frac{16}{n^4(s-1)^4} a_5 \quad \cdots \right) . \tag{39}$$

where

$$a_1 = C_2 - 1,$$
$$a_2 = 4C_3 - 3C_2^2 - 1$$
$$a_3 = 18C_4 - 32C_3C_2 + 15C_2^3 - 1$$
$$a_4 = 96C_5 - 210C_4C_2 - 80C_3^2 + 300C_2^2C_3 - 105C_2^4 - 1$$
$$a_5 = 600C_6 - 1584C_5C_2 - 1080C_4C_3 + 2520C_4C_2^2 - 3360C_3C_2^3 + 1960C_3^2C_2 + 945C_2^5 - 1.$$

(b) *Effect of Non-Normality Assuming Block Variances Constant*

With this assumption we find

$$E_2(V_2) = \frac{2n}{ns-n+2} + \frac{1}{s}\gamma_2 - \frac{1}{n}[12\gamma_2/s(s-1) - 3\gamma_2^2/s^2 + 2\gamma_4/s^2 + 8(s-2)\gamma_1^2/s(s-1)^2] + \frac{1}{n^2}[100\gamma_2/s(s-1)^2 - 6\gamma_2^2/s^2(s-1) + 40\gamma_4/s^2(s-1) + 160(s-2)\gamma_1^2/s(s-1)^3 + 3\gamma_6/s^3 + 96(s-2)\gamma_3\gamma_1/s^2(s-1)^2 - 16\gamma_4\gamma_2/s^3 - 64(s-2)\gamma_2\gamma_1^2/s^2(s-1)^2 + 15\gamma_2^3/s^3]. \quad . \quad . \quad . \quad . \quad . \quad . \tag{40}$$

The modifying factor δ contains $E(V_2)$ to order n^{-1} thus if the number of blocks was small

the effect of unequal block variances could be appreciable. On the other hand, assuming non-normality but equality of variance, the factor δ contains γ_2 only to order $(ns)^{-1}$. Thus the effect of general non-normality would be small unless both n and s were small.

As an example we may compare the results for inequality of variances in randomized blocks with exact values obtained using the theory of quadratic forms in multi-normally distributed variates (Box, 1954*b*).

TABLE 3

Approximate and Exact Probabilities of Exceeding Normal Theory 5 per cent. Point when Block Variances are Unequal, Assuming a Normal Population

Number of Treatments (s)	*Number of Blocks* (n)	*Block Variances*	*Percentage Chance of Exceeding 5 per cent. Point when Null Hypothesis is True*	
			Approx.	*Exact*
11	3	1, 2, 3	4·4	4·3
5	3	1, 2, 3	4·3	4·3
11	3	1, 1, 3	3·8	3·8
5	3	1, 1, 3	3·7	3·9
3	11	1, 1, . . . 1, 3	4·8	4·9

5. ROBUST TESTS FOR VARIANCES

In the tests to compare means studied in the last section the corrective factors were of order N^{-1} so that the normal theory tests were for these examples "non-parametric to order N^0". As the sample size was increased the sampling distribution of the criterion considered would thus ordinarily tend to its normal theory form whatever the parent distribution. It has been shown in an earlier paper (Box, 1953) that for tests on variances the corrective factors are of order N^0, and these tests depend directly upon the assumption of normality, therefore, for all sample sizes. It was also shown that this difference in behaviour arose because, whereas in tests to compare means we compare the variation among the means with an estimate of the variation obtained from internal evidence within the groups, in current tests to compare variances (F test on two independent samples, L_1 test, Bartlett test, Cochran's test, F(max) test, Wald's sequential test) we tacitly compare some measure of variation among the variances with a theoretical value which is correct only for the normal distribution. The variation among a set of variances depends upon the fourth moment just as the variation in a set of means depends upon the second moment so that what is needed is to "studentize" the variance tests for the fourth moment just as the tests on means are "studentized" for the second moment.

A simple way of doing this, which was discussed, involved the conversion of the test on variances to a test on means. The manner in which this had to be done was, however, somewhat arbitrary and we shall here investigate an alternative procedure based on the permutation theory discussed above.

5.1. *Tests to Compare Two Variances. Means Assumed Known*

The simplest case in which a test may be made for equality of variances is that of two groups with the mean of each group known. If a typical observation is denoted by x_{ti} ($i = 1, 2, \ldots n_t$; $t = 1, 2$) we can assume without loss of generality that $E(x_{ti}) = 0$. The normal theory criterion is then

$$F = n_2 \sum_{i=1}^{n_1} x_{1i}^2 \Big/ n_1 \sum_{j=1}^{n_2} x_{2j}^2 \quad . \quad . \quad . \quad . \quad . \quad (41)$$

An equivalent statistic for testing the same hypothesis is

$$W = \sum_{i=1}^{n_1} x_{1i}^2 \Big/ \left(\sum_{i=1}^{n_1} x_{1i}^2 + \sum_{j=1}^{n_2} x_{2j}^2 \right) \quad . \quad . \quad . \quad . \quad (42)$$

When the distribution is normal W follows the Beta distribution with degrees of freedom n_1 and

n_2. The first two moments of the permutation theory distribution are readily calculated and may be compared with those for the normal theory distribution

$$\underset{P}{E} = n_1/N \qquad \underset{N}{E} = n_1/N \quad . \quad . \quad . \quad . \quad (43)$$

$$\underset{P}{V} = \frac{2n_1n_2}{N^2(N+2)}\left\{1 + \frac{1}{2}\frac{N}{N-1}(b_2 - 3)\right\} \quad \underset{N}{V} = \frac{2n_1n_2}{N^2(N+2)} \quad . \quad . \quad . \quad (44)$$

where

$$b_2 = (N+2)\left(\sum_{i=1}^{n_1} x_{1i}^4 + \sum_{j=1}^{n_2} x_{2j}^4\right)\Big/\left(\sum_{i=1}^{n_1} x_{1i}^2 + \sum_{j=1}^{n_2} x_{2j}^2\right)^2 \quad . \quad . \quad . \quad (45)$$

Using the same argument as before we find that to carry out the approximate permutation test we should enter the F criterion in the usual tables but with dn_1 and dn_2 degrees of freedom where

$$d = \left[1 + \tfrac{1}{2}\left\{\frac{N+2}{N-1-(b_2-3)}\right\}(b_2 - 3)\right]^{-1} \text{ or to order } N^0,\ d = [1 + \tfrac{1}{2}(b_2 - 3)]^{-1} \quad (46)$$

As before, the approximate effect of non-normality is obtained by replacing b_2 by $E(b_2)$. We find that to order N^{-2}

$$\begin{aligned} E(b_2 - 3) = {} & \gamma_2 - N^{-1}(2\gamma_4 + 11\gamma_2 + 20\gamma_1^2 - 3\gamma_2^2) \\ & + N^{-2}(3\gamma_6 - 16\gamma_4\gamma_2 + 15\gamma_2^3 + 38\gamma_4 + 168\gamma_3\gamma_1 - 3\gamma_2^2 \\ & \qquad - 160\gamma_2\gamma_1^2 + 86\gamma_2 + 380\gamma_1^2) \end{aligned} \quad (47)$$

The result agrees to order N^0 with that previously given when it was shown by a different argument that the effect of non-normality on this test could be represented approximately by a modification in degrees of freedom by a factor $\delta = (1 + \frac{1}{2}\gamma_2)^{-1}$.

If we compare this result with that obtained from the test to compare means we have approximately for the modifying factors

comparison of means	comparison of 2 variances
$1 + \gamma_2/N$	$(1 + \frac{1}{2}\gamma_2)^{-1}$

. . . (48)

We note that in the modifying factor for a test on means γ_2 appears only to order N^{-1} whereas for the test on variances it appears to order N^0. Thus the sample size would not have to be very large before a discrepancy which seriously upset the variance test would be negligible for the test on means. We note also that the effects will be opposite in direction in the two tests; for example in sampling from a leptokurtic population the significance of the test on means would be slightly increased but that for the test on variances reduced.

Before the modified test on variances could be recommended as a useful procedure, two questions needed to be considered.

(1) How good is the moment approximation to the permutation test?

(2) How much power is lost by using the modified test when the distribution happens to be normal?

It may be remarked that since the divergencies of the permutation moments from those of normal theory are far larger for the tests on variances than for the tests on means the Beta function approximation would be expected to be much more heavily strained.

To shed light on these questions an extensive sampling experiment was conducted with the object of comparing the behaviour of the standard F test and the modified F test in regard to their robustness and power.

5.2. *Sampling Experiment to Compare Standard and Modified Procedures for Comparing Two Variances*

The power and robustness of the standard F test and the modified F test were investigated for the rectangular, normal and double-exponential parent distributions.

The empirical sampling procedure involved drawing 2000 samples of size 20 from each of these

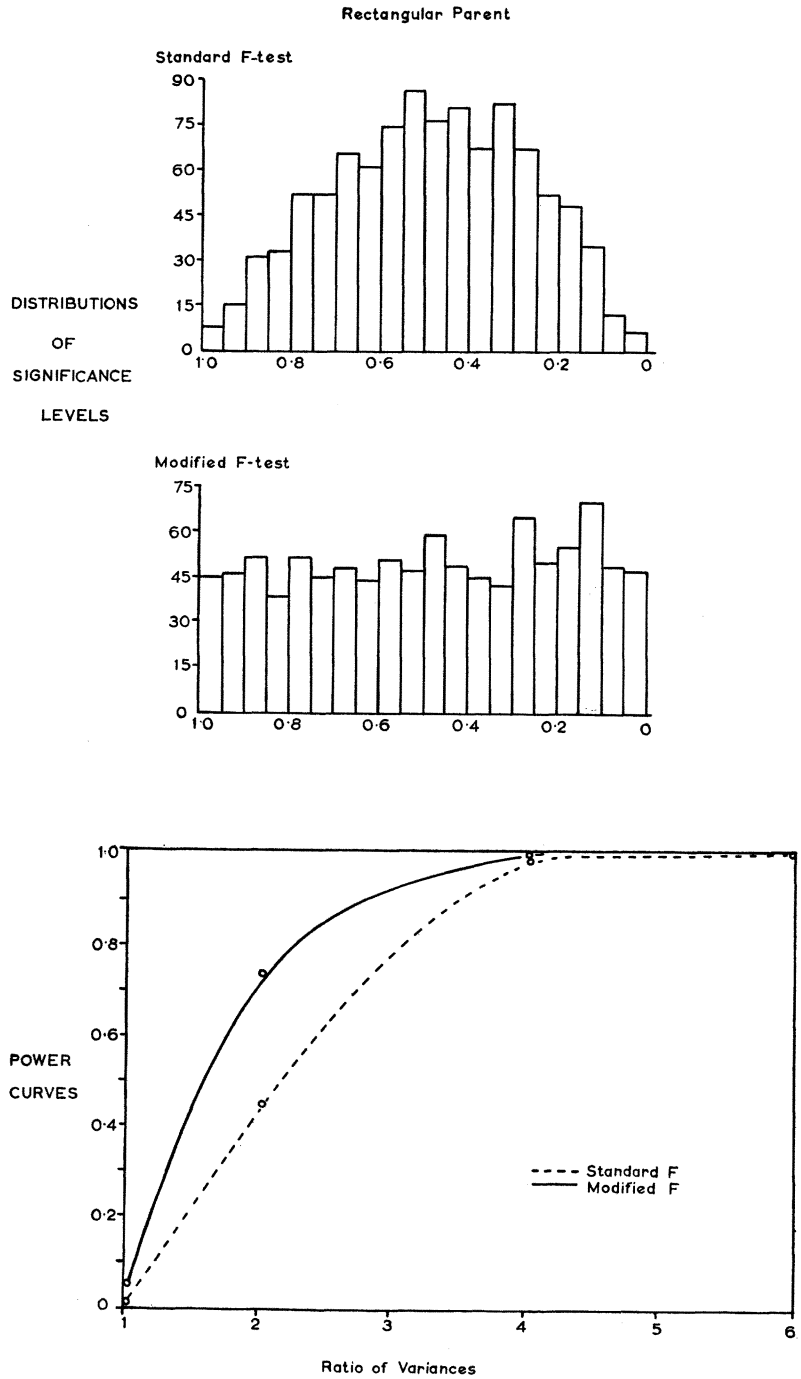

FIG. 2(a).—Behaviour of Standard F-test and Modified F-test for a 'Rectangular' Parent distribution.

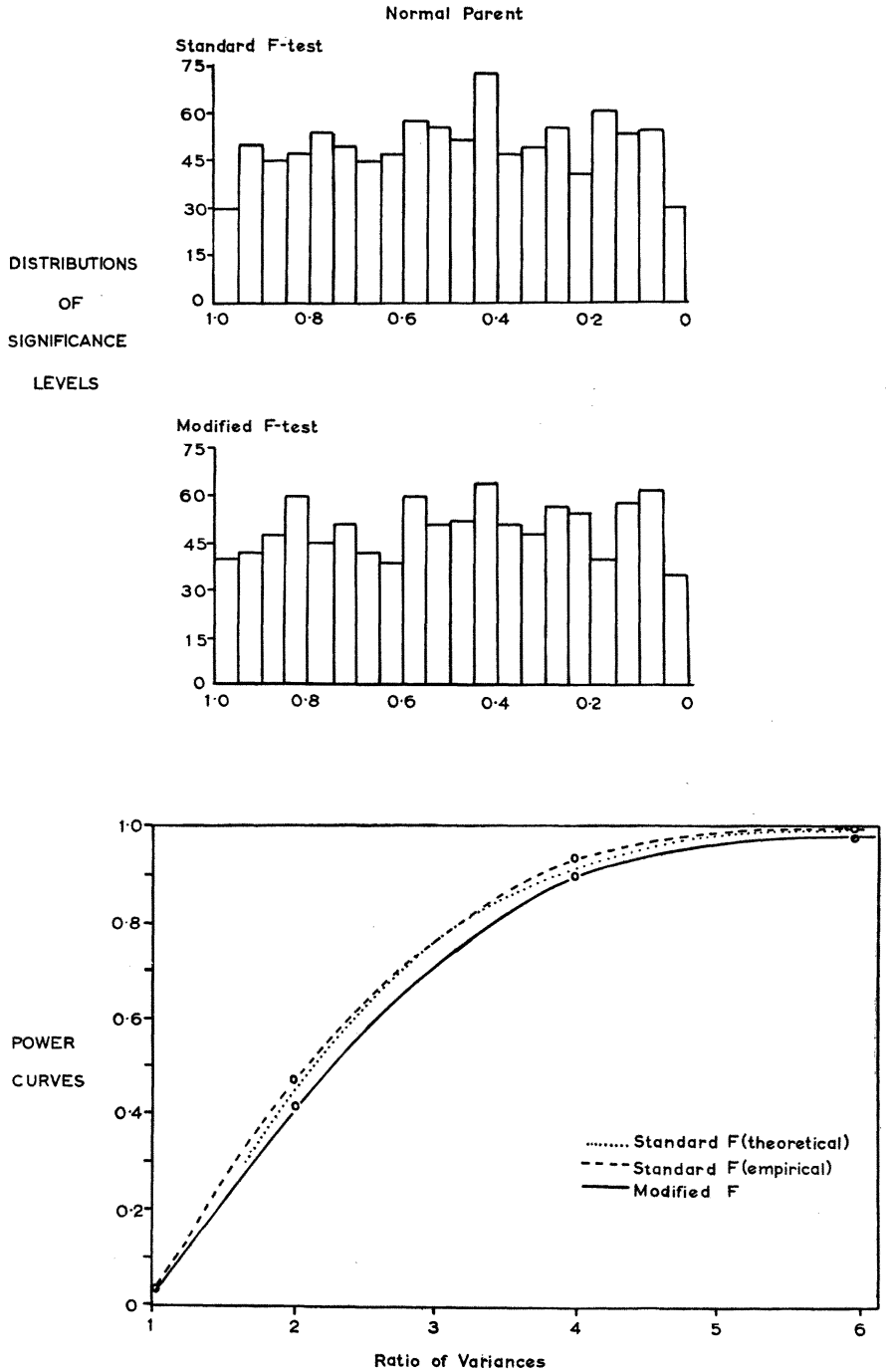

FIG. 2(b).—Behaviour of Standard F-test and Modified F-test for a 'Normal' Parent distribution.

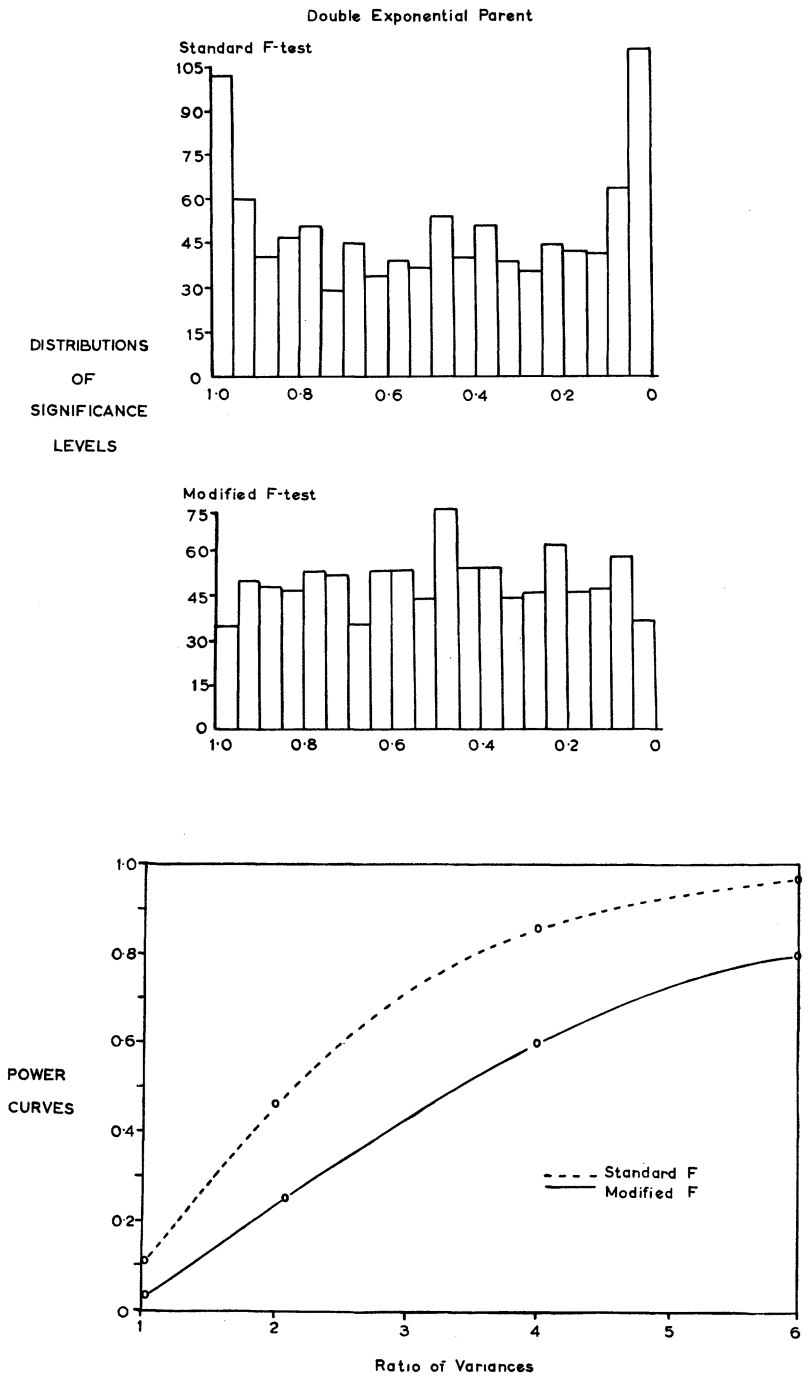

FIG. 2(c).—Behaviour of Standard F-test and Modified F-test for a 'Double Exponential' Parent distribution.

populations. The samples were paired to give 1000 values of the standard F statistic for each population. The appropriate probability associated with each of these F's was estimated by the standard and modified methods from a set of graphs prepared from Pearson's table of the incomplete Beta function.

In addition to the calculations above, which evaluate the behaviour of the criteria when the null hypothesis is true, the distributions were estimated for three alternative hypotheses:

$$H_1\colon\ 2\sigma_1^2 = \sigma_2^2 \qquad H_2\colon\ 4\sigma_1^2 = \sigma_2^2 \qquad H_3\colon\ 6\sigma_1^2 = \sigma_2^2$$

In drawing the original 40,000 observations, 100 values of the deviates of the density functions were calculated at the 0·5, 1·5, 2·5 . . . 99·5 percentile points for each of the three populations. The deviate together with its square and fourth power was punched on an IBM card for each of the 100 percentile points and for the three distributions, making a master deck of 100 cards for each distribution.

The sampling deck of 40,000 cards was then made up by putting on blank cards a two digit random number and a sequence number running from 1 to 40,000. These were then sorted into 100 groups according to the random number appearing on the card. On each card containing random number 00 was punched the 0·5 percentile value of the variate with its square, cube and fourth power; on the card with random number 01 was punched the 1·5 percentile value, etc. The 40,000 cards were then sorted into their original sequence by the sequence number to give the sampling deck from which 2000 samples of 20 were drawn.

The imperfection in the three populations caused by selecting only 100 values to approximate the continuous distributions may be noted. Comparisons of the theoretical and actual measures of kurtosis γ_2 are shown below.

Values of γ_2 in Parent Populations for Sampling Experiment:

	Rectangular	*Normal*	*Double Exponential*
"Theoretical" . .	−1·2000	0·0000	3·0000
Actual	−1·2002	−0·1660	1·7301

The deviations of these populations from their theoretical counterparts is of no serious consequence as we are concerned only to obtain populations representing three degrees of kurtosis centred about the normal.

If the tests behaved as we should wish, then when the null hypothesis was true, 5 per cent. of the samples would show a "level of significance" between 0 and 5 per cent., a further 5 per cent. between 5 and 10 per cent. and so on. Thus in making 1000 tests when the null hypothesis was true the expected number showing a "level of significance" within each 5 per cent. range would be 50. Fig. 2 shows the distributions of significance levels actually found. The observed distributions illustrate the failure of the standard F test and the improvement to be obtained by using the modified test. The χ^2 goodness of fit test may be used to compare the actual frequencies in the 20 groups with the theoretical expectation of 50 per group which would be appropriate if the test were unaffected by the non-normality of the parent universe. For the rectangular population the value of χ^2 is 254·2 ($P < 10^{-7}$ per cent.) for the standard test and 21·8 ($P = 29·0$ per cent.) for the modified test. The standard F test gives only 0·7 per cent. of the values below the 95 per cent. point and 0·6 per cent. of the values above the 5 per cent. point. The modified test corrects almost perfectly for the non-normality giving 4·5 per cent. of the values below the 95 per cent. point and 4·7 per cent. of the values above the 5 per cent. point.

For the "normal" parent distribution it is of some interest to note that the slight imperfection of the parent population and in particular its truncation is apparently enough to upset the standard test. The value of χ^2 for this population is 36·1 ($P = 1·1$ per cent.), for the standard test and 27·8 ($P = 8·6$ per cent.) for the modified test. In particular, the percentages below the 95 per cent. and above the 5 per cent. point are increased from 3·0 to 4·0 per cent. and from 3·0 to 4·5 per cent. respectively as a result of using the modified test.

For the double-exponential distribution the value of χ^2 is reduced from 166·2 ($P < 10^{-7}$ per cent.) to 33·0 ($P = 4·13$ per cent.). In this case the modified test appears to slightly over-correct,

the percentages below the 95 per cent. point and above the 5 per cent. point being reduced from 10·2 to 3·6 and 11·0 to 3·6 per cent. respectively.

Having demonstrated the relative insensitivity of the modified test to departures from normality the question arises as to how much power is lost by the modification of the standard test. Smooth curves drawn through the points obtained from the sampling experiments with populations corresponding to the alternatives H_1, H_2 and H_3 are shown in the lower part of Fig. 2.

It will be seen that, when the distribution was normal, close agreement was found between the empirical and theoretical curves for the standard F test and there appeared to be little loss of power with the modified test. For example, the probability of detecting a variance ratio of 3 : 1 was reduced from about 75 per cent. to about 71 per cent. by the use of the modified test. Since this difference in power is subject to a sampling error of 2 per cent. the 95 per cent. confidence range for the loss of power is 0–8 per cent.

The power curve for the standard F test with 18 and 18 degrees of freedom and $\alpha = 5$ per cent. coincides very closely with the modified F test curve for 20 and 20 degrees of freedom. It appears therefore that, in the case studied, a loss of power of about 10 per cent. occurs in the sense that 10 per cent. more observations are required with the modified test than with the standard test to obtain the same power.

The standard F test does not have a 5 per cent. intercept for the non-normal populations and consequently the comparison of power curves shown in Fig. 2 cannot be made directly. What can be noted is that the modified test is more powerful for the rectangular and less powerful for the double-exponential distribution.

5.3. *Tests to Compare* k *Variances*

The statistic most commonly used to test the equality of k variances when k is greater than 2 is the Bartlett statistic

$$M_1 = \nu \ln \bar{s}^2 - \sum_{t=1}^{k} \nu_t \ln s_t^2 \qquad . \quad . \quad . \quad . \quad . \quad (49)$$

where s_t^2 $(t = 1, \ldots, k)$ is an estimate of variance based on ν_t degrees of freedom, $\bar{s}^2$ is the average variance $\nu \bar{s}^2 = \Sigma \nu_t s_t^2$ and $\nu = \Sigma \nu_t$.

Bartlett showed that for a normal parent population M_1 is distributed very nearly as $(1 + A)$ χ^2_{k-1}, where χ^2_{k-1} follows the chi-square distribution with $k - 1$ degrees of freedom. The modifying factor $A = \{3(k - 1)\}^{-1} \{\Sigma \nu_t^{-1} - \nu^{-1}\}$ is of order ν^{-1} and is small for moderate and large samples.

5.3.1. *Means Assumed Known*

For two groups the Bartlett statistic and the "double-sided" F test are equivalent. Thus, if we assume the means are known (so that $n_t s_t^2 = \sum_{t=1}^{n_t} x^2_{ti}$ and $\nu_t = n_t$), the approximate permutation test could be alternatively carried out in the Bartlett form instead of in the F form. It is readily seen that to the same degree of approximation as before, M_1 would be distributed for this case $k = 2$, on permutation theory, as $d^{-1}(1 + Ad^{-1})\chi^2_{k-1}$. The constant A is of order ν^{-1} and since the test depends on d to order ν^0 there seems little point in using this refinement. We should therefore obtain a test (which, like the test on means, is "non-parametric" to order ν^0) by referring $\{1 + \frac{1}{2}(b_2 - 3)\}^{-1} M_1$ to the chi-squared distribution with $k - 1$ degrees of freedom.

One might expect that such a procedure could be generalized to more than two groups. As a confirmation of this we have evaluated the first two permutation moments of M_1 to order N^{-1}. The derivation is outlined below:

In this case where the means are assumed known and equal to zero,

$$n_t s_t^2 = \sum_{i=1}^{n_t} x^2_{ti}, \quad N\bar{s}^2 = \sum_{t=1}^{k} \sum_{i=1}^{n_t} x^2_{ti}, \quad \nu_t = n_t \text{ and } \nu = N.$$

If we write $z_t = (s_t^2 - \bar{s}^2)/\bar{s}^2$ the statistic M_1 may then be written

$$M_1 = -\sum_{t=1}^{k} n_t \ln(1 + z_t) \qquad . \quad . \quad . \quad . \quad . \quad (50)$$

which we expand to give

$$\sum_{t=1}^{k} n_t(z_t - \tfrac{1}{2} z_t^2 + \tfrac{1}{3} z_t^3 - \ \ldots\ ) \quad . \quad . \quad . \quad . \quad . \quad (51)$$

Now it will be noted that the denominator of z_t is a constant for all permutations of the observations. By straightforward but tedious algebra it is now possible to evaluate the permutation moments of each term in the expression and so to obtain the expected value of M_1. In a similar way $E(M_1^2)$ may be found and hence the variance of M_1. Proceeding in this way we find to order N^{-1}

$$\underset{P}{E}(M_1) = (k-1)\left[\tfrac{1}{2}(b_2' - 1) + N^{-1}\{\tfrac{1}{2}(b_2' - 1) + (NA - 3)(3b_2' - b_4' - 2) + \frac{9}{4}(3NA - 2)(b_2' - 1)^2\}\right] \quad . \quad (52)$$

where as before

$$A = \frac{1}{3(k-1)}\left(\Sigma \frac{1}{n_t} - \frac{1}{N}\right)$$

and is of order N^{-1}.

Also

$$\underset{P}{V}(M_1) = \tfrac{1}{2}(k-1)(b_2' - 1)^2$$

plus terms of order N^{-1} as follows:

$$\frac{1}{36N}\begin{array}{c|cccc} & N\Sigma\frac{1}{n_t} & k^2 & k & (1) \\ b_6 & +9 & -9 & -18 & +18 \\ b_4 & +72 & & -144 & +72 \\ b_4b_2 & -108 & +36 & +216 & -144 \\ b_2^3 & +168 & -36 & -288 & +156 \\ b_2^2 & -207 & +9 & +306 & -108 \\ b_2 & +72 & & -72 & \\ (1) & -6 & & & +6 \end{array} \quad . \quad . \quad . \quad (53)$$

where

$$b_r' = N^{\frac{1}{2}r} \sum_{t=1}^{k} \sum_{i=1}^{n_t} x_{ti}^{r+2} \Big/ \{\sum_{t=1}^{k} \sum_{i=1}^{n_t} x_{ti}^2\}^{\frac{r+2}{2}}.$$

The terms in N^{-1} involve the higher moment ratios b_4 and b_6. However, to order N^0 we have

$$\underset{P}{E}(M_1) = \{1 + \tfrac{1}{2}(b_2' - 3)\}(k-1) \qquad \underset{P}{V}(M_1) = 2\{1 + \tfrac{1}{2}(b_2' - 3)\}^2(k-1) \quad . \quad (54)$$

which are the first two moments of $\{1 + \frac{1}{2}(b_2' - 3)\}\chi^2$. This, together with the fact that it has been shown elsewhere that for a non-normal parent M_1 is distributed asymptotically as $(1 + \frac{1}{2}\gamma_2)\chi^2$, confirms that the statistic $M_1/\{1 + \frac{1}{2}(b_2' - 3)\}$ should supply a criterion of greater robustness for comparison of variances.

5.3.2. *Means Not Assumed Known*

If we assume group means known, we can derive the permutation moments of the W form of the F statistic and an expansion of the permutation moments of M_1. We are able to do this

because the denominators in the expansions to be considered contain the sum of squares of the N observations and are fixed for all permutations of the sample. If the means are not known then this quantity is replaced by the sum of the squares of deviations from the group means, which changes in different permutations, and we are faced with the problem of determining permutation moments for ratios, which greatly complicates the problem.

We have proceeded therefore by assuming a criterion which has a form analogous to that found for the case in which the means were assumed known.

$$M_1' = M_1/(1 + \tfrac{1}{2}c_2) \quad . \quad . \quad . \quad . \quad . \quad . \quad (55)$$

where in the case of equal groups

$$c_2 = k \Sigma k_{4t}/(\Sigma k_{2t})^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (56)$$

and k_{4t}, k_{2t} are k statistics for the t^{th} group.

It is possible to confirm by series expansion that, to order N^0, M_1' does in fact have mean and variance $k - 1$ and $2(k - 1)$.

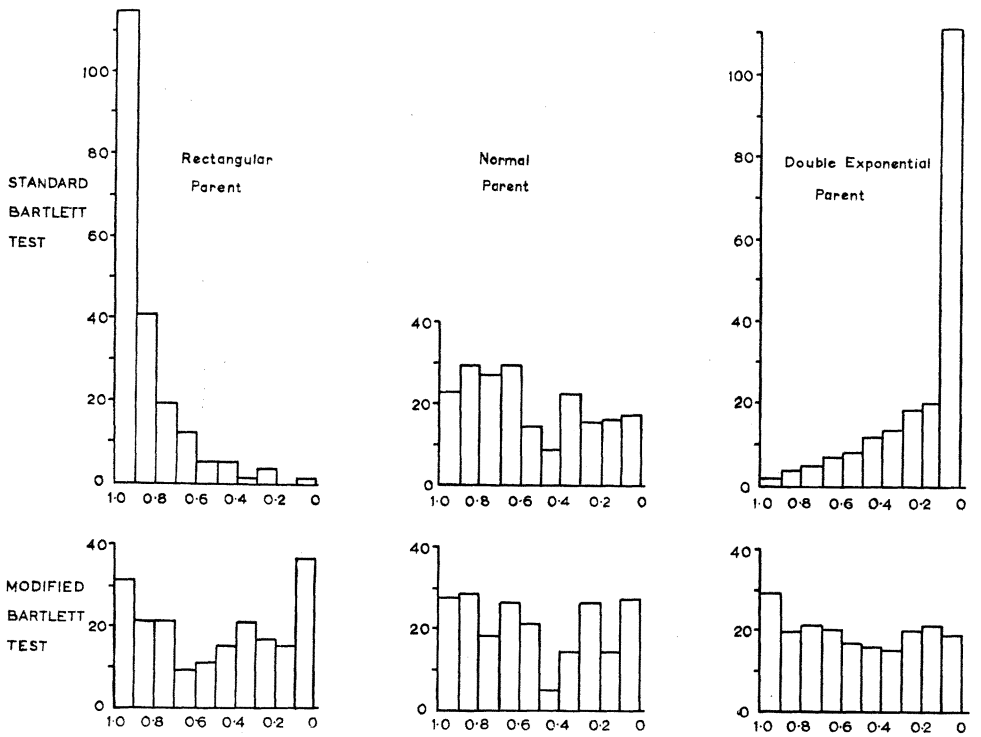

FIG. 3.—Behaviour of Standard and Modified Bartlett test for various Parent distributions.

5.4. *Sampling Experiment for* M_1'

2000 samples of 20 observations were drawn from each of the three populations described in Section 5.2. These were subdivided into 200 sets of 10 samples each and for each set of 10 samples both the standard M_1 and the modified M_1' were computed to obtain three sampling distributions of each statistic. The results of the sampling when the null hypothesis of equal variances was true are shown in Fig. 3. These verify the extreme sensitivity, already noted by Box (1953) and by Kempthorne and Barclay (1953), of the normal theory test under conditions of non-normality and the considerable improvement achieved by using the modified statistic M_1'.

The most serious aberration from ideal behaviour in the modified test is the occurrence of 37 cases as against the expected 20 in the upper decile when sampling from a rectangular population. However, some solace can be found in the fact that this population represents a more severe deviation from normality than is commonly encountered.

As before we may make a chi-squared test of the hypothesis that the underlying distribution of probabilities is rectangular and contains 20 samples in each decile grouping. The results of these computations show

Parent Distribution	*Test Criterion*	$\chi^2(9)$	*Probability (per cent.)*
Rectangular . . .	M_1	550·8	0·00001
	M_1'	36·1	0·004
Normal	M_1	21·2	1·170
	M_1'	22·9	0·738
Double-exponential . .	M_1	475·8	0·00001
	M_1'	6·7	66·9

While inspection of the diagrams for the rectangular parent population and the marked reduction in χ^2 indicates that the modified criterion is much less sensitive to non-normality than the unmodified criterion, yet the significance of 0·004 per cent. for the chi-square test indicates that M_1' is not behaving entirely as we should wish.

Again in the case of the normal population the results are not ideal for either test. In addition to the investigation of the properties of the test criteria in the null case, a further 2000 samples of 20 observations was drawn from normal populations with different variances to estimate the power with respect to the alternative hypothesis that

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1; \quad \sigma_5^2 = \sigma_6^2 = \sigma_7^2 = 1{\cdot}7; \quad \sigma_8^2 = \sigma_9^2 = \sigma_{10}^2 = 2{\cdot}6$$

For a significance level of 5 per cent. the probabilities of detecting heterogeneity of variance of the magnitude of H_1 for both the standard and modified statistics when sampling from the normal parent population are found to be for the standard test 81·5 per cent. and for the modified test 81·0 per cent., each value being subject to a sampling error of about 2·8 per cent.

6. Discussion and Summary

One of the simplest statistical procedures is the test of the hypothesis that the mean of a population is equal to μ when the standard deviation σ is known. If a sample of N observations $x_1, x_2, \ldots x_N$ is available the criterion usually chosen is $\sqrt{N}\,\bar{x}/\sigma$ which is referred to tables of the unit normal curve.

The validity of this test of the null hypothesis does not rest on the supposition that the observations are exactly normally distributed. We can appeal to the central limit theorem which tells us that for almost all parent distributions the criterion is distributed asymptotically in the form assumed; moreover that the convergence to this form is *rapid*. In particular, since $\gamma_r(\bar{x}) = \gamma_r(x)/N^{r/2}$, the coefficients of skewness and kurtosis for the mean are $\gamma_1(\bar{x}) = \gamma_1(x)/N^{\frac{1}{2}}$ and $\gamma_2(\bar{x}) = \gamma_2(x)/N$. Thus for all but extremely small sample sizes and "pathological" parent distributions the null test is approximately valid.

A very similar argument may be employed to establish the practical validity of the analysis of variance tests. It appears from permutation theory that the criterion tends to the normal theory form whatever the parent distribution and since the modifying factor in the degrees of freedom is approximately $\delta = 1 + E(\gamma_2)/N$ the convergence is rapid.

When this 'central limit' property is lacking the criterion is of much less practical utility and it seems necessary to seek alternative tests which have greater robustness. One way of doing this is by approximating to the appropriate permutation test. The form of permutation test statistic can be made to depend on the "most likely" alternative distribution.

In the test for equality of variances it is shown how a more robust test based on Bartlett's criterion, but employing the sample estimate of b_2, may be obtained. This form of criterion is

however somewhat complicated and other simpler criteria, for example a "studentized" form of Hartley's F(max) criterion, might be developed on similar lines.

We are greatly indebted to Professor Gertrude Cox who provided the opportunity for us to carry out this research, and to Dr. R. J. Hader who was responsible for the administration of the ordnance contract. Our thanks are also due to F. J. Verlinden, the computing staff at North Carolina State College and to Mr. E. Broughton for valuable assistance with the computations.

References

BARTLETT, M. S. (1934), *Proc. Camb. Phil. Soc.*, **30**, 164–169.
—— (1937), *Proc. Roy. Soc., A*, **160**, 268–282.
BOX, G. E. P. (1952), *Biometrika*, **39**, 49–57.
—— (1953*a*), *Biometrika*, **40**, 318–335.
—— (1953*b*), *Ann. Math. Statist.*, **24**, 687.
—— (1954*a*), *Ann. Math. Statist.*, **25**, 290–302.
—— (1954*b*), *Ann. Math. Statist.*, **25**, 484–498,
COCHRAN, W. G. (1937), *J.R. Statist. Soc. Suppl.*, **4**, 102–118.
DAVID, F. N., and JOHNSON, N. L. (1951*a*), *Biometrika*, **38**, 43–57.
—— —— (1951*b*), *Trobajos de Estadistica*, **2**, 179–188.
—— —— (1951*c*), *Ann. Math. Statist.*, **22**, 382–392.
DAVID, F. N., and KENDALL, M. G. (1949) *Biometrika*, **36**, 431–449.
FISHER, R. A. (1928), *Proc. London Math. Soc.*, Series 2, **30**, 199–238.
—— (1935), *The Design of Experiments*. Edinburgh: Oliver & Boyd.
—— (1936), *J.R. Anthrop. Soc.*, **66**, 57–63.
GAYEN, A. K. (1950), *Biometrika*, **37**, 236–255.
GEARY, R. C. (1936), *J.R. Statist. Soc. Suppl.*, **3**, 178–184.
—— (1947), *Biometrika*, **34**, 209–242.
HORSNELL, G. (1953), *Biometrika*, **38**, 128–186.
JAMES, G. S. (1951) *Biometrika*, **38**, 324–329.
KEMPTHORNE, O., and BARCLAY, W. D. (1953), *J. Amer. Stat. Soc.*, **48**, 610–614.
LEHMAN, E. L., and STEIN, C. (1949), *Ann. Math. Statist.*, **20**, 28–45.
NEYMAN, J., and PEARSON, E. S. (1933), *Phil. Trans. A.*, **231**, 289–337.
PEARSON, E. S. (1931), *Biometrika*, **23**, 114–133.
—— (1935), *Biometrika*, **27**, 333–352.
—— (1937), *Biometrika*, **29**, 53–64.
PITMAN, E. J. G. (1937*a*), *J.R. Statist. Soc. Suppl.*, **4**, 119–130.
—— (1937*b*), *Biometrika*, **29**, 322–335.
TOCHER, J. F. (1928), *Biometrika*, **20**, 105–244.
TUKEY, J. W. (1928), *Human Biology*, **20**, 205–214.
WELCH, B. L. (1937*a*), *Biometrika*, **29**, 21–52.
—— (1937*b*), *Biometrika*, **29**, 350–362.
—— (1938), *Biometrika*, **30**, 149–158.
—— (1951), *Biometrika*, **38**, 330–336.

DISCUSSION ON THE PAPER BY DR. BOX AND DR. ANDERSEN

Dr. B. L. WELCH: Dr. Box and Dr. Andersen have presented a paper concerned with topics which have often been discussed but which, nevertheless, have not recently received much consideration in the Research Section. With the paper as a whole I am in agreement. For instance, when comparing several variances in samples from non-normal populations, there is no doubt that we shall be wrong in assuming normality, because the standard error of any criterion used will not then be right even in the large sample sense. In the same way, when comparing means of samples of unequal sizes from populations with unequal variances, if we inject into the situation the assumption that the variances of the populations are equal, the standard error of our criterion will again be wrong to order n^0. The position is, however, somewhat different with the standard analysis of variance tests, at least in as far as they are employed to test null hypotheses in experiments where there has been randomization. There is more scope for expressions of personal opinion about the effects of departures from assumption in these cases and I shall accordingly confine my remarks to such tests.

Consider the simplest example—the E^2-test of homogeneity as it is usually carried out assuming normal theory. Now the normal theory variance of E^2 may be compared with the variance of E^2 calculated from permutation theory. The ratio of these variances differs from unity only

by a quantity of order N^{-1} where N is the number of individuals in all the groups combined, a number which will usually be large even if the separate groups are small. There seems little likelihood, therefore, that the normal theory E^2-test will prove misleading even when applied to non-normal data. In fact, by making use of a characteristic (0, 1) variable, the normal theory E^2-test can even be used to test the significance of the Lexis measure of dispersion when the numbers in the groups are as small as two—a situation with which the more usual χ^2-test does not cope adequately.

Consider next the usual normal-theory criterion used to assess the significance of treatments in a randomized block experiment. Here again the normal theory variance of the criterion may be compared with that deduced from permutation theory. The possibility that the ratio of these variances will differ appreciably from unity is again seen to be rather remote, although this is more likely to occur than in the case of the E^2-criterion which I have just mentioned. It can happen if the variances between plots within blocks differ very much from block to block, and where this is so one can improve upon the usual test of the null hypothesis by modifying the degrees of freedom in the very convenient way Dr. Box and Dr. Andersen suggest. But if there are one or two blocks obviously very much more variable than the others one might prefer to proceed more simply by ignoring altogether the more variable blocks and carrying out the usual analysis on the remainder.

As one proceeds to more and more complex experimental situations, consideration of the corresponding permutation theory indicates more and more possibilities of departure from the usual normal theory assumptions, and indeed it has been observed in uniformity trials that the permutation variance of the standard analysis of variance criterion in Latin Square experiments can on *individual* fields differ considerably from the normal theory variance of the same criterion. This has never seemed to me in itself conclusive, since the totality of plot yields in an individual experiment is usually to be regarded as in some sense a random sample from a larger population of sets of plots—however difficult it may be to give precision to the delineation of this population. The viewpoint of the present authors is, I gather, here much the same as mine. Empirical investigation on a large scale would be necessary to determine whether permutation theory and normal theory will often lead to differing conclusions when the wider view is taken. But the work which has been done so far on this problem does at least have the merit of emphasizing that the more complicated the experimental situation becomes, the less heavily one can lean on permutation theory to validate the usual normal theory tests and the more heavily one must lean on the mathematical assumptions made.

It is with very great pleasure that I propose the vote of thanks to the authors.

Dr. H. E. DANIELS: The authors are to be congratulated on their worthwhile study of an important practical problem. Sooner or later every statistician is confronted with data to which he hesitates to apply the standard tests, and in this paper an honest attempt is made to meet the difficulty by providing tests which avoid it at not too great a cost. I have only two observations to make, the first of a cautionary nature, the second perhaps more constructive.

It is pointed out in the paper that subsidiary tests of the assumptions underlying the main test may lead to quite a wrong decision as to whether or not the latter may be safely used. We are urged to avoid such preliminary explorations and to use whenever possible a robust test. Where the assumptions concern "extraneous" factors I would certainly agree, but it must be remembered that situations exist where a failure in the assumptions not merely affects the significance level and power of the test but renders pointless the very hypothesis being tested. Robustness is then irrelevant; the experimenter is obliged to do preliminary tests unless he has strong grounds for believing them unnecessary.

Consider, for example, the typical analysis of covariance situation. In comparing the adjusted treatment means it is assumed that neither the concomitant variable x nor the regression on it is affected by treatments. If the latter assumption is violated I suppose one can still regard the test as comparing treatment means adjusted to the average x for the whole experiment. With this more restricted point of view it would still be feasible to examine the test for robustness. But if x is also affected by the treatments, comparison of the adjusted means is an arbitrary procedure and a multivariate analysis becomes appropriate. In cases like this the question being asked is in fact conditional on the assumptions made.

My other point concerns permutation distributions. The authors have followed Pitman and Welch in approximating to the permutation distribution of a statistic lying between 0 and 1 by the corresponding beta distribution having the same mean and variance. Pitman showed that except in very unlikely samples the third and fourth moments also matched well. Nevertheless one is usually interested in the "tail" region where the relative error of the approximation might

still be appreciable, and, as Dr. Box has emphasized, the matter requires further study, particularly in regard to the test for variances.

There does exist another type of approximation obtained by the method of steepest descents familiar to students of statistical mechanics. This so-called saddlepoint approximation, though a good deal more elaborate to compute than the beta type, usually involves a relative error of $0\left(\frac{1}{n}\right)$ in the tails and could perhaps serve as a standard of comparison in place of the exact permutation distribution.

As a simple illustration let us consider the example of section 3, though the beta type approximation is known to be satisfactory in that case. The sample consists of n differences $y_i = x_{1i} - x_{2i}$. Let z_i take the values $\pm y_i$ with probability $\frac{1}{2}$ and consider the distribution of $Z = \Sigma z_i$, from which that of t or w can be derived. Its cumulant generating function is $K(T) = \Sigma \log \cosh Ty_i$ and the saddlepoint approximation to the distribution of Z turns out to have probability density

$$g(Z) = \frac{1}{\sqrt{2\pi K''(T)}} e^{K(T)-TZ}$$

where T is the unique real root of the equation

$$Z = K'(T) = \Sigma y_i \tanh Ty_i$$

(In the extreme case where $y_i = \pm 1$ the permutation distribution is binomial and

$$g(Z) = \frac{1}{\sqrt{2\pi}} \frac{n^{n+\frac{1}{2}}}{(n-Z)^{\frac{1}{2}(n-z+1)} (n+Z)^{\frac{1}{2}(n+z+1)}}$$

which is its Stirling approximation.) Numerical integration then yields the distribution function after renormalizing to make the total probability unity. In practice it is more convenient to work in terms of T which has probability density

$$h(T) = \sqrt{\frac{K''(T)}{2\pi}} e^{K(T)-TK'(T)}$$

and to interpolate at the required values of $Z = K'(T)$.

I tried this out on the following eight values for y_i extracted from Fisher and Yates's random numbers: 81, 91, 57, 7, 15, 35, 60, 9. Taking a series of probability levels for the normal t test the corresponding probabilities for the permutation distribution and the two approximations were computed as shown in the table:

Normal t	·90	·70	·50	·30	·10	·05	·02	·01	·001
Exact	·891	·672	·492	·259	·117	·055	·031	·008	—
Beta	·884	·680	·487	·271	·105	·055	·023	·012	·0015
Saddlepoint	·899	·697	·499	·278	·114	·063	·027	·010	—

For a sample size as small as eight the approximations fit reasonably well, but an interesting feature is that in both the t and the fitted beta cases non-zero probabilities occur beyond the range of the exact test which is determined by $|Z| \leqslant \Sigma |y_i|$. The beta approximation depends on the inequality $Z^2 < \Sigma y_i^2$, which is not attained except when every $|y_i|$ is the same. Possibly a beta approximation covering the correct range might be an improvement.

In the more interesting case of variances the saddlepoint approximation is unfortunately too elaborate to give here but I hope to continue the discussion elsewhere.

I have much pleasure in seconding the vote of thanks to Dr. Box and Dr. Andersen for this excellent paper.

The vote of thanks was put to the meeting and carried unanimously.

Dr. D. R. Cox: On a point of presentation, it might help to plot the power curves in Fig. 2 on log-probability paper. The curves for one test at different significance levels would then be, approximately, a set of parallel straight lines so that the curves for different tests at unequal significance levels can be compared easily in a rough and ready way.

I should like to say something about tests for dispersion based on a number of small samples, a matter discussed by Dr. Box in his earlier paper on robust tests. Suppose, for example, that we have a number k of small samples each of size n from each of two populations whose variations are to be compared, and that k is so small that the observed variation in dispersion within populations cannot be used. Then the question arises as to whether tests based on the sample range

are more robust than tests based on the sample r.m.s.; for normal populations and for $n \leqslant 5$ the difference in power between the tests is very small.

For general non-normality bias in the estimate derived from the range is irrelevant and what we have to consider is the coefficient of variation, as a function of β_2, of estimates derived from (*a*) the mean range and (*b*) the r.m.s. within samples. The coefficient of variation of range is, of course, not a function of β_2, but it is possible to estimate the general form of the relation and then to get results like the following.

Samples of four. Coefficient of variation of estimates from (*a*) range, (*b*) r.m.s. as a ratio of the coefficient of variation at $\beta_2 = 3$.

β_2	1	3	5	7	9
(*a*)	0·83	1·00	1·23	1·34	1·41
(*b*)	0·50	1·00	1·32	1·58	1·80

The tests based on range are therefore more robust than those based on r.m.s.

This is for a number of small samples. If the method is to be applied to larger samples these have first to be divided into subsamples. The best size of subsample is known to be seven or eight, although a moderate gain in efficiency can be achieved by using samples of about four or five and recovering some of the information in the variation of the subsample totals.

Mr. F. J. ANSCOMBE: Dr. Box and Dr. Andersen have given a most helpful account of what may be done in situations where we should like to apply the customary normal-law methods of analysis to a set of data but where we have no confidence that the assumptions underlying those methods are satisfied. The authors have shown how the customary methods may be modified so as to remain valid under very weak conditions, and have pointed out how great are the departures from the usual assumptions that can be tolerated if the customary methods are not modified in this way.

I should like to outline briefly some work done this last summer by Professor J. W. Tukey and me, which is complementary to the work of the present authors. Tukey and I have been considering the transformation of observations to improve the appropriateness of the usual normal-law methods. Ideally, if one could so arrange it, one would like the distribution of the observations to be normal, with constant variance, so that the usual procedures would be strictly valid. One would also like all regressions to be linear and all factors (in a factorial experiment) to have additive effects, i.e. no interactions, so that the interpretation of the observations would be as clear and easy to grasp as possible. An approach can sometimes be made towards this ideal by transforming the observations before analysis. But since we generally have no reliable theoretical information about the distribution of the observations or the structure of responses, we need some practicable method of investigating a set of observations, either in their original state or after a tentative transformation has been applied, to see how far these various requirements are met. It is convenient to begin by a graphical treatment, as follows. Corresponding to each observation y, calculate the fitted value Y and the residual $z = y - Y$. For example, in a two-way classification, with observations y_{ij}, we have, in the usual notation,

$$Y_{ij} = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot},$$

and then

$$z_{ij} = y_{ij} - Y_{ij}.$$

Now plot the z's against the Y's on a scatter diagram. Since

$$\Sigma z = \Sigma zY = 0,$$

the ordinates have zero mean and zero linear component of regression on Y. If there is non-additivity present of the sort that could be removed by a transformation of the observations, there will be a curved (quadratic) regression of the z's on the Y's. If the variance of the y's changes progressively with the mean, the variance of the z's will change progressively with Y and the points may have a wedge-shaped outline. If the distribution of the y's is not normal, the marginal distribution of the z's will be not normal (in general). So non-additivity, heteroscedasticity and non-normality can be investigated at once on the same diagram. It is possible to calculate criteria to test these effects, or estimate their magnitude; the criteria are based on the statistics

$$\text{(i) } \Sigma zY^2, \quad \text{(ii) } \Sigma z^2Y, \quad \text{(iii) } \Sigma z^3, \quad \text{(iv) } \Sigma z^4,$$

(i) being for non-additivity, (ii) for heteroscedasticity, and (iii) and (iv) for non-normality. The

complete formulae depend on the design of the experiment and are liable to be tedious to obtain, owing to the correlations between the z's.

The relevance of the present paper to these studies is that we require investigations of the sort that Dr. Box has carried out if we are to decide how far we can afford to sacrifice constancy of variance and normality for the sake of additivity.

Professor CHAPMAN: This study by Dr. Box and Dr. Andersen is indeed valuable from the point of view of evaluating some of the *ad hoc* procedures which have been adopted in statistics, such as the procedure of piling one test on another in order to try to improve the validity or efficiency of the final test; the present and previous papers by Dr. Box show that one does not necessarily gain anything by so doing. We should, however, not necessarily condemn all preliminary tests because in some cases they were unnecessary or have been used without adequate study. A study made by Dr. Paull (*Annals of Mathematical Statistics*, **21** (1950), 539), showed, for example, that a preliminary test could be used and a procedure derived which was better than standard procedures which did not involve a preliminary test. Therefore, in the long run, particularly in complicated situations, it may be necessary to use preliminary tests, but they must be used with care and only after adequate study has shown what is the overall behaviour of the whole series of tests.

Of course the applied statistician will frequently prefer to use a single robust test, in which connection the criteria that Dr. Box has introduced may well be extended into four criteria, using the familiar ideas of type 1 and type 2 error. First, the statistician looks for a test which satisfies the prescribed limits of type 1 error under the assumptions. He will also want to minimize the type 2 error under the basic assumption, i.e. to have a powerful test (this is requirement 1 in Dr. Box's paper). The statistician would also like a test which would keep the type 1 error fairly constant if the assumptions are not fulfilled. This is the idea of robustness. Still further, the statistician would like to have a test which keeps small the type 2 error if the assumptions fail; thus robustness has two aspects, and this particular point is of some interest in connection with analysis of variance tests.

If we consider the test for equality of means of two normal populations with common variances, then the analysis of variance test (which specializes in the case of two populations to the two sample t-test) is known to be most powerful among all unbiased tests. Recent papers by Dr. Mood and Professor Dixon (in the *Annals of Mathematical Statistics*, **25** (1954), 514, 610) on the power of distribution free tests show that the Wilcoxon test compares quite favourably with this test in this situation, i.e. where the assumptions underlying the t-test are valid. Not much is yet known of the comparison of the power of these tests when the assumptions are not fulfilled. Results of a study announced by Professor Lehmann at Stanford, California, in June 1953, comparing the asymptotic relative efficiency of these tests suggest however that for some non-normal distributions the Wilcoxon test may be much more powerful, for sufficiently large samples, than the two sample t-test.

This second aspect of robustness is in fact considered by Dr. Box and Dr. Andersen in the sampling study of the power of their proposed permutation test to compare variances. Moreover the sampling study suggests again that the permutation test may show to much greater advantage from the power point of view when the basic assumptions are not valid.

Professor BARTLETT: I am in sympathy with Dr. Box and Dr. Welch in their endeavour to ascertain how the results reached would affect practical work. Very many applied statisticians naturally want fairly simple rules, but there is the possible conflict between such simple rules and the assumptions on which these are based not always holding. As has already been mentioned, the analysis of variance technique in particular is robust in this way, in fact the actual technique of analysing variance and getting the various mean squares is, in itself, independent of any further testing, and the randomization in experimental work moreover makes these tests very fool-proof, at least in the case of equal numbers of observations for the different treatments. So that while it may be advisable occasionally to look at one's assumptions to see when they will break down, Dr. Box has confirmed in the case of analysis of variance that one does get this robustness. In other situations there may be the difficulty of not knowing quite where to stop in the search for robust tests. For example, when there is no randomization there will be the question what to do about independence. There is a whole field of problems in regard to which one is aware that the question of dependence has to be considered, and many previous analyses in which independence had been assumed are incorrect. One can consider, if necessary, making robust tests which allow for dependence, but in this broad field one can get into rather deep water if one attempts to make robustness cover everything, because there will then be a danger of the tests becoming not powerful

enough to be very useful. So that in some cases at least one may require to specify rather narrower assumptions than one would ideally wish from Dr. Box's standpoint.

To mention a small point in connection with the homogeneity of variances test, Dr. Box has not looked at this so much in the context of analysis of variance, in the sense of disentangling variances to see whether they are homogeneous. One may there have to be a little careful. It is well to bear in mind that if one takes linear combinations of observations the variances of such linear combinations can be the same as before (for an orthogonal transformation), but if the original observations are non-normal the linear combinations of them will be nearer normality. There is, of course, no paradox here; the fallacy would be to suppose that the resulting new observations were independent if the original observations were. This emphasizes the need to take care that the observations being studied are the original ones for which the assumption of independence may be reasonable.

Dr. F. YATES: I should like to give a word of warning concerning the approach to tests of significance adopted in this paper. It is very easy to devise different tests which, on the average, have similar properties, i.e. they behave satisfactorily when the null hypothesis is true and have approximately the same power of detecting departures from that hypothesis. Two such tests may, however, give very different results when applied to a given set of data. This situation leads to a good deal of contention amongst statisticians and much discredit of the science of statistics. The appalling position can easily arise in which one can get any answer one wants if only one goes around to a large enough number of statisticians.

This is itself an argument against lightheartedly elaborating a multiplicity of tests of significance to deal with data of the same type. But the tests given in the paper seem open to objection on other grounds also. It is stated in section 3.2 that the points P and Q shown in Fig. 1, which give the same significance level of 9·2 per cent. on the t-test, give significance levels of 12·0 per cent. and 4·3 per cent. respectively on the approximate permutation test. These differences are perhaps not very alarming, but the same argument will show that three observations having the values 9·8, 10·1 and 10·1 which, on the normal theory test, or indeed on any ordinary test of significance, will give a high level of significance ($t = 100$, giving a probability of about ·01 per cent.) will be judged non-significant on the approximate permutation test, while the values 9·9, 9·9 and 10·2 will be judged significant on the approximate permutation test. I, personally, would certainly find it difficult to convince an intelligent man that there is any profound difference between these two sets of observations.

I will not elaborate on these questions here, but should like to suggest that a good deal of the trouble has arisen through adopting the Neyman-Pearson attitude to tests of significance. This attitude, if I understand it aright, is a consequence of regarding tests of significance as criteria for definite action, i.e. essentially the decision function approach, whereas, in fact, tests of significance are used in scientific research as an aid to assessing the weight of evidence for or against a given hypothesis. Scientific research workers do not reject a hypothesis once and for all if a certain set of observations show a significant departure at the 5 per cent. point, any more than they accept it if the observations do not show any significant departure from the hypothesis. If scientific research proceeded on these lines it would in a very short time be in a state of inextricable confusion.

Incidentally, I should like to question the statement in section 2 that the points of view expressed by Fisher and Pearson are in no way contradictory. To suggest that Fisher, or scientists who use Fisherian tests, are not concerned with the power of tests of significance to detect relevant departures from the hypothesis being tested is a gross misrepresentation.

Mr. A. J. MAYNE: A possible new method of forming robust statistics is as follows. Consider a statistic S, which is a function of a sample drawn from a population with any distribution; S can be a univariate or multivariate statistic. Then polynomial transformations of S, with coefficients formed from the sample data, can usually be found, whose distributions can be made successive approximations to a standard normal distribution, or the distribution of S for a normal population. The r'^{th} approximation is usually a polynomial of degree r in S, with dominant term S for all except very small sample sizes. As r increases, the r'^{th} approximation in general yields a progressively more robust statistic.

Although little work has yet been done on this new method, and more investigations on its efficiency are needed, it has been applied to the t-statistic of §3.1 of Box and Andersen's paper. The null hypothesis considered for this example of application of the new method is that the parent distributions of both populations are the same, thus that the distribution of $y \equiv x_1 - x_2$ is a general *symmetric* univariate distribution. Let k_2 be the r'^{th} order *sample* k-statistic, formed

from the sample items which are the differences between the values of the observations in each pair, i.e. formed from $y_i \equiv x_{1i} - x_{2i}$ $(i = 1 \text{ to } n)$. Let

$$l_4 \equiv k_4 k_2^{-2}$$

Then

$$[1 - \tfrac{1}{4} l_4 n^{-1}] t + \frac{1}{12} l_4 n^{-1} t^3$$

is a statistic with improved fit to the distribution of t for normal x_1 and x_2 populations, and

$$[1 - \tfrac{1}{4}(1 + l_4) n^{-1}] t - \tfrac{1}{4}(1 - \tfrac{1}{3} l_4) n^{-1} t^3$$

gives an improved fit to the standard normal distribution.

The powers of the tests based on these transformed statistics have not yet been investigated, but it seems reasonable to suppose that, if the departure of the population from normality is not too large, they are comparable to the power for the original statistic when the population is normal, if the transformed statistics are fitted to the distribution of the original statistic for normal population.

Professor BARNARD: The strategic as opposed to detailed tactical advances in statistics are usually associated with the birth and christening of new ideas. "Efficiency", "sufficiency", "likelihood", "power", are some examples. "Robustness" is a new addition to the family, and Dr. Box is to be congratulated not only for giving birth to it, but also for such aptness in its christening. It can confidently be predicted that the idea will have a long and vigorous life.

However, the ideas we use in statistics form a somewhat unruly family, and the task of keeping order will not be made easier by the new arrival. I share Dr. Yates's concern lest we find ourselves in the position where two statistical tests, while applying with apparently equal appropriateness to answering apparently the same question on the basis of the same data, yet give different answers. It must not be forgotten that statistical inference owes its existence as a science to the recognition of the fact that there are good and bad ways of analysing data; and the existence of cases where the "good" answer is not determined tends to blur this recognition. While we must expect some difficult cases to occur, they should be rare.

From this point of view it is comforting to learn that the t-test and the F-test applied to the analysis of variance are robust tests. Thus we are able to say, if asked why we test the hypothesis that the observations are *normally* distributed with mean μ, that it would make little difference if our observations were not normally distributed, provided they are nearly so, and this latter we can usually justify by an appeal to the nature of the data (suitably transformed, perhaps). And these two tests seem to account for practically all the tests we make on measured data.

It is curious how difficult it is to get away from situations where the t- and F-tests apply. Just as attempts to generalize to multivariate tests leave me rather cool—I have yet to see a multivariate test not reducible to a t- or F-test which really seemed to mean something—so attempts to extend in the direction of other moments than the first seem more difficult than one might think. For example, we might be tempted, if challenged to give an instance where two wholly independent estimates of variance are to be compared, to quote the comparison of the accuracy of two methods of measurement, or of two guns; but here we can argue that what we are really interested in is not the mean square deviation from the sample mean, but the mean square deviation from the true value; and if we measure deviations from the true value (and perhaps make a transformation), the comparison reduces to a t-test.

Mr. EHRENBERG: I think that robust tests of significance are obviously desirable, but at the same time I should like to make a critical remark upon their use. There is, in these days, a tendency to treat variance heterogeneity and departures from normality as being of importance only in as far as they may affect tests of significance. In practice, however, it is surely necessary, before considering how to test the significance of the difference of means, say, to see whether such means are useful descriptive parameters; the data may, for example, be skew. This putting of the cart before the horse is suggested on p. 2, para. 3, for example, in which the authors seem to warn us against even bothering to look for such things as skewness, kurtosis, variance heterogeneity, and so on; they seem to warn us, that is, against seeing what the data are really like.

It may be that the authors' attitude can hardly be considered at fault within the framework adopted in their initial sentence, namely the restriction to the testing of *a priori* hypotheses. But that framework is extraordinarily narrow. To cling to a hypothesis which turns out to be technically inappropriate—for example one formulated in terms of mean values when some of the

data turn out to be skew—will, of course, prevent an understanding of the data. And practically any experiment, however well-considered in the first instance, will throw up many facts or implications which have not been considered beforehand. Fisher's famous dictum that every experiment may be said to exist to give the facts a chance to disprove the null hypothesis is stimulating, though hardly true. Every experiment may be said to exist, not to give the facts a chance to disprove the null hypothesis, but just to give the facts a chance. In view of Dr. Yates's remarks, I feel that an experimental scientist may say (if no statistician is involved) that every experiment exists merely to give the facts. The methodological notion of explicit hypothesis-making is nowadays rather allowed to master us instead of being used merely in as far as it can be helpful.

In using robust tests, as in any other application of random sampling theory, it is as well to ask oneself first how one would analyse the data if they were not random data at all. After all, tests of significance only become necessary and valid, as Fisher stressed in the quotation given at the beginning of section 2 of the paper, when there has been a physical act of randomization.

Professor BARTLETT: I think that Mr. Ehrenberg is somewhat exaggerating the possibility of allowing facts to speak for themselves. Always there is a compromise between theory and facts; facts of themselves are meaningless unless one is asking some kind of question. Certainly in some cases one allows the facts to suggest generalities which one then investigates further, but merely to search for apparent results from the facts can be taken too far.

Mr. EHRENBERG: I cannot help feeling that the test-book methods over-emphasize looking at only those facts about which one has made a hypohesis beforehand. Often one does not make *a priori* hypotheses, but has to see how existing facts fit in with other facts or theories.

The authors subsequently replied in writing as follows:—

We are happy that our paper provoked what to us at least was a most stimulating discussion and are grateful to all who took part. With most of the speakers we are in agreement, but a few points call for comment. As we have tried to emphasize, most of the standard normal theory tests to *compare means* are remarkably robust and admirably fulfil the needs of the experimenter. Our object in discussing permutation theory for *these* tests is to demonstrate this, and to consider more closely the behaviour of the permutation tests in those cases with which we are most familiar. Our object is not, as Dr. Yates supposes, to suggest alternative tests for these situations.

If the state of affairs which exists for tests to compare means was found for all other tests, then to introduce alternative procedures with the same general properties might be as pointless and harmful as Dr. Yates believes. Unfortunately the happy state of affairs found for tests on means is *not* a general one. For example, it has been shown that the current tests to compare several variances, which have for many years been recommended in the text-books, are grossly misleading for populations showing a degree of non-normality frequently encountered in practice but not readily detectable for sample sizes usually considered. There is reason to believe that a similar state of affairs may be found in other instances including those we mention. To ignore this situation for the sake of uniformity would produce the result that, presented with the same data, all statisticians arrived at the same *wrong* conclusions. This would lead to more discredit of the science of statistics than would an honest attempt to face the situation and remedy it. Incidentally, we see no particular reason why statisticians should be expected to give the same answer if they are asked different questions.

Concerning the behaviour of the approximations to Fisher's permutation test, our object in preparing Fig. 1 was to make the point that Dr. Yates has restated namely that for individual samples the normal theory and permutation test can give dissimilar results. The permutation test has real relevance only for significance levels corresponding with a probability level $\alpha > 1/2^n$. For such levels the difficulties mentioned by Dr. Yates do not of course arise, although quite large differences between the two sorts of test can be observed for particular samples. Unfortunately, it is not easy to illustrate diagrammatically the situation for $\alpha > 1/2^n$ and $n > 3$ and we feel that Fig. 1 is of some interest provided the reservations we made are taken into account. Incidentally, we find it difficult to detect the connection between the behaviour of this test and the Neyman-Pearson theory which plays no essential part in the test's derivation.

The basic difficulty is that whatever we assume may be "taken too literally". For example, on the one hand, if we assume normality, a test based on that assumption may fail unless that assumption is almost exactly true. On the other hand, a permutation test is derived on the supposition that we are prepared to believe almost anything about the population however unlikely and may not therefore give sensible results for certain extreme samples. In most practical situations the user of the test believes neither in exact normality nor in the possibility of extreme departures from it. The mental concept of a prior distribution of distributions is, however, difficult to

use as a basis for deriving a real test statistic, and it seems that in practice we must consider each on its merits and proceed, as in other sciences, by making the simplest assumptions first and elaborating these only if it transpires that to obtain a test of practical value it is necessary to do so. In this connection, it would be safest to behave as though the robustness of tests on means was a lucky accident. When new methods are suggested their authors should be required to show not only that their procedures have desirable properties when the assumptions are true, but also that they can be expected to work reasonably well with the degree of departure from assumption likely to be encountered with real data. This might have the secondary but desirable result of alleviating to some extent the present pressure on space in some of the statistical journals.

The difficulties that face us in attempting to reply to Dr. Yates's final point are similar to those which must be experienced by neutral countries who try to keep out of the "cold war". We suspected that if we included statements by Fisher and Pearson in the same paragraph, even though these were about different things, this could lead to a discussion generating more heat than light. We therefore were at some pains to explain that the points of view we quoted were complementary rather than contradictory. Alas, we are now not only accused of saying that Pearson and Fisher agree but in the very next sentence are charged with "gross misrepresentation" for saying (which we did not) that Fisher and Pearson do not agree about the importance of powerful tests. Against this double-edged weapon no defence is possible and we admit defeat.

We have confined consideration to hypothesis testing only for purposes of simplicity. We do not regard such tests as the be-all and end-all of statistics and we point out in the first sentence of our paper that similar considerations of robustness should be applied for other statistical procedures. We conceive scientific research as developing by dual processes which may be called synthesis and analysis. In the process of synthesis the hypothesis is built up. The appropriate data are then collected. The process of analysis now begins and out of it a new synthesis is generated and so on. The two processes used in alternation are applied not once but many times during the course of most real investigations. Of the two, we feel that process of synthesis (almost completely neglected by statisticians) is the most important to the ultimate success of real investigations. One of us has discussed these ideas in more detail in a paper shortly to appear in *Biometrics*. What we are concerned with in the present paper is that tests which may be applied at each stage of this process *should not mislead* the experimenter. On the question of the usefulness of tests other than those to compare means we agree with Mr. Ehrenberg rather than with Professor Barnard. However, our intention seems to have been misunderstood by the former speaker for, far from warning against the use of a test such as that for variance heterogeniety, we go to some trouble to try to formulate a procedure which will be useful for this purpose. We also point out that we consider that such tests should be made because we are interested in variance heterogeneity, normality, etc., in their own right and not as auxiliary features only of importance in relation to their effect on other tests. We believe with Mr. Ehrenberg that one must examine the data in various ways, particularly to decide what has to be done next. All that we would plead is that the tests used, whether they be few or many, shall be robust so that having applied them, we know what it is we have really tested.