ROBUST TESTS FOR VARIANCES AND
EFFECT OF NON NORMALITY AND VARIANCE
HETEROGENEITY ON STANDARD TESTS

Prepared Under Office of Ordnance Research
Contract No. DA-36-034-ORD-1177 (RD)

by

G. E. P. Box
S. L. Andersen

Institute of Statistics
Mimeo Series No. 101
April, 1954
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter I</th>
<th>GENERAL REVIEW OF LITERATURE ON NON NORMALITY AND PERMUTATION TESTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1.1</td>
<td>Effects of Non Normality 5</td>
</tr>
<tr>
<td>Section 1.2</td>
<td>Nature and Use of Permutation Tests 8</td>
</tr>
<tr>
<td>Section 1.3</td>
<td>Rejection of Preliminary Tests and Introduction of Robustness 13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter II</th>
<th>PERMUTATION TESTS FROM THE POINT OF VIEW OF THE NEYMAN-PEARSON THEORY OF TESTING STATISTICAL HYPOTHESES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 2.1</td>
<td>Introduction 16</td>
</tr>
<tr>
<td>Section 2.2</td>
<td>Permutation Tests 18</td>
</tr>
<tr>
<td>Section 2.3</td>
<td>Permutation Tests with More than One Population 19</td>
</tr>
<tr>
<td>Section 2.4</td>
<td>Form of the Permutation Test to be Used 21</td>
</tr>
<tr>
<td>Section 2.5</td>
<td>Permutation Moments 24</td>
</tr>
<tr>
<td>Section 2.6</td>
<td>Overall Moments from Permutation Moments 24</td>
</tr>
<tr>
<td>Section 2.7</td>
<td>Example of a Permutation Test 25</td>
</tr>
<tr>
<td>Section 2.8</td>
<td>Approximation to The Permutation Test 30</td>
</tr>
<tr>
<td>Section 2.9</td>
<td>The Critical Region for The Approximate Permutation Test 33</td>
</tr>
<tr>
<td>Section 2.10</td>
<td>Use of the Permutation Test to Estimate the Approximate Effect of Non Normality 38</td>
</tr>
<tr>
<td>Section 2.11</td>
<td>Summary 41</td>
</tr>
</tbody>
</table>
Chapter III  MODIFIED t TESTS AND ANALYSIS OF VARIANCE

Section 3.1  Randomized Blocks  43
Section 3.2  Paired t Test  44
Section 3.3  Analysis of Variance (one-way classification)  44

Chapter IV  EFFECT OF NON NORMALITY AND UNEQUAL VARIANCES ON t TESTS AND ANALYSIS OF VARIANCE

Section 4.1  Formulae for $\delta$  47
Section 4.2  Tabled Values of Type I Error where Failure of Assumptions Occur in Testing Means  51
Section 4.3  Summary  60

Chapter V  TESTS ON VARIANCES, MEANS ASSUMED KNOWN

Section 5.1  Test to Compare Two Variances, Means Known  62
Section 5.2  Sampling Experiment to Study Two Procedures for Comparing Two Variances  65
Section 5.3  Test to Compare k Variances, Means Known  73

Chapter VI  TESTS ON VARIANCES, MEANS ASSUMED UNKNOWN

Section 6.1  Theoretical Justification  79
Section 6.2  Empirical Sampling Experiment  81
Section 6.3  Summary of Robust Tests for Variances  87

Chapter VII  SUMMARY AND CONCLUSIONS  88
BIBLIOGRAPHY

APPENDIX A  NOTES ON GENERAL TECHNIQUE OF FITTING A BETA DISTRIBUTION BY THE FIRST TWO MOMENTS

APPENDIX B  CALCULATION OF FACTOR, d, FOR ROBUST F TESTS

APPENDIX C  CALCULATION OF FACTOR, e, FOR EVALUATING THE EFFECT OF ASSUMPTION FAILURES

APPENDIX D  METHOD OF CONSTRUCTION OF CRITICAL REGION FOR THE APPROXIMATE PERMUTATION TEST

APPENDIX E  DERIVATION OF ROBUST F TEST FOR COMPARING TWO VARIANCES, MEANS KNOWN

APPENDIX F  DERIVATION OF MODIFIED BARTLETT STATISTIC, M', MEANS UNKNOWN

APPENDIX G  DERIVATION OF MODIFIED BARTLETT STATISTIC, MEANS UNKNOWN
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Example of a Permutation Test</td>
<td>25</td>
</tr>
<tr>
<td>2.</td>
<td>Type I Error for Paired t Test</td>
<td>40</td>
</tr>
<tr>
<td>3.</td>
<td>Comparison of Exact and Approximate Permutation Type</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>I Error for Randomized Blocks with Unequal Block Variances</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Type I Error for Randomized Blocks with Unequal Treatment Variances</td>
<td>54</td>
</tr>
<tr>
<td>5.</td>
<td>Type I Error for Randomized Blocks with Non-Normal Parent Population</td>
<td>55</td>
</tr>
<tr>
<td>6.</td>
<td>Type I Error for the Paired t Test</td>
<td>56</td>
</tr>
<tr>
<td>7.</td>
<td>Type I Error for Comparing k Means in The Analysis of Variance, Non-Normal Parent Population</td>
<td>57</td>
</tr>
<tr>
<td>8.</td>
<td>Type I Error for Comparing Three Means in The Analysis of Variance, Heterogeneous Error Variances</td>
<td>58</td>
</tr>
<tr>
<td>9.</td>
<td>Type I Error for Comparing Two Means, Non-Normal Parent Population</td>
<td>59</td>
</tr>
<tr>
<td>10.</td>
<td>Type I Error for One Sided t Test</td>
<td>60</td>
</tr>
<tr>
<td>11.</td>
<td>Parent Populations for Sampling Experiment, Values of $\beta_2$</td>
<td>67</td>
</tr>
<tr>
<td>12.</td>
<td>Sampling Results Summary for Comparison of F and Modified F Test</td>
<td>72</td>
</tr>
<tr>
<td>13.</td>
<td>Empirical Comparison of The Robustness of the Two Criteria, $M$ and $M^*$</td>
<td>85</td>
</tr>
<tr>
<td>14.</td>
<td>Power of The Test Criteria, $M$ and $M^*$</td>
<td>87</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1.</td>
<td>Permutation Distribution of The Sample Mean</td>
<td>27</td>
</tr>
<tr>
<td>2.</td>
<td>Permutation Distribution of The Statistic, W</td>
<td>27</td>
</tr>
<tr>
<td>3a.</td>
<td>Normal Theory and Approximate-Permutation Critical Regions in Three Dimensions</td>
<td>36</td>
</tr>
<tr>
<td>3b.</td>
<td>Comparison of Normal-Theory and Approximate-Permutation Critical Regions, W_{10} (Two Dimensional Cross Section)</td>
<td>37</td>
</tr>
<tr>
<td>4.</td>
<td>Sampling Results to Compare F tests, Rectangular Parent Population</td>
<td>68</td>
</tr>
<tr>
<td>5.</td>
<td>Sampling Results to Compare F tests, Normal Parent Population</td>
<td>69</td>
</tr>
<tr>
<td>6.</td>
<td>Sampling Results to Compare F tests, Double Exponential Parent Population</td>
<td>70</td>
</tr>
<tr>
<td>7.</td>
<td>Sampling Power Curves, Normal Distribution</td>
<td>74</td>
</tr>
<tr>
<td>8.</td>
<td>Sampling Power Curves, Rectangular Distribution</td>
<td>75</td>
</tr>
<tr>
<td>9.</td>
<td>Sampling Power Curves, Double Exponential Distribution</td>
<td>76</td>
</tr>
<tr>
<td>10.</td>
<td>Empirical Sampling Distributions to Compare M and M', Double Exponential Distribution</td>
<td>82</td>
</tr>
<tr>
<td>11.</td>
<td>Empirical Sampling Distributions to Compare M and M', Normal Distribution</td>
<td>83</td>
</tr>
<tr>
<td>12.</td>
<td>Empirical Sampling Distributions to Compare M and M', Rectangular Distribution</td>
<td>84</td>
</tr>
<tr>
<td>13.</td>
<td>Graphical Construction of Approximate-Permutation Test Critical Region</td>
<td>102</td>
</tr>
</tbody>
</table>
INTRODUCTION

The problem considered in this dissertation arises in connection with the use of standard parametric tests of significance. In practical experimental situations the assumption of normality and other common assumptions such as homoscedasticity and independence of errors are never perfectly satisfied. Yet the inductive inferences made depend directly on these unrealistic assumptions, on the basis that they are often approximately true or that the derivation of the test criteria is greatly simplified by these assumptions. This reasoning, in itself, offers little consolation to the research worker whose decisions depend largely on the statistical analysis which conceivably may be in gross error due to the failure of the standard assumptions to hold in the physical system with which he is working.

Specifically, the nature of the induction in hypothesis testing is concerned with the "acceptance" or rejection of some predetermined null hypothesis in such a way that we have a known probability of making the error of rejecting the null hypothesis when it is actually true. This error, known as the type I error, $\alpha$, can be determined exactly if the assumptions underlying the test criterion hold. The usefulness of the test procedure in practical situations does not require that the actual type I error equal the value calculated by the standard test, but rather requires that the actual type I error be approximately equal to the nominal value. Test procedures, for which the actual and nominal value of $\alpha$ do not differ greatly for the type and magnitude of assumption failures found in practice, have been called "robust", and may be used
with confidence by the research worker. Those tests for which nominal and actual values of \( \alpha \) differ greatly in practical situations are of little use to the experimenter.

Therefore, the general approach to the problem of making tests of significance in the real world, which is not "ideal", is to:

1. Investigate the robustness of standard statistical tests to failures of assumptions which occur in practical situations.

2. Give a stamp of approval to those tests for which nominal and actual type I error agree fairly closely (e.g., actual \( \alpha = .03 - .07 \) for nominal \( \alpha = .05 \)).

3. Develop new test criteria for those situations where the existing standard tests are extremely sensitive to the underlying assumptions.

Previous research has shown that, for the univariate case, tests for comparing means are generally robust, while tests for comparing variances are not. The primary problem of this paper is the development of a more robust test for comparing \( k \) variances than the existing standard test due to M. S. Bartlett.

The form of the new test criterion is suggested by first considering the permutation test based on Bartlett's statistic,

\[
M = N \ln (s^2) - \sum_{t=1}^{k} n_t \ln (s_t^2)
\]

where: \( s_t^2 \) = sample variance in \( t \)-th group
\[ s^2 = \frac{1}{N} \sum_{t=1}^{k} n_t s_t^2 \]

\[ n_t = \text{number of observations in each group} \]

\[ k = \text{number of groups} \]

\[ N = \text{total number of observations in all k groups} \]

for the situation where the means are known for all k groups. An approximate-permutation test can then be obtained by fitting a Beta distribution to the first two moments of the permutation distribution.

Having derived this test criterion, the analogous test for the more common case where the means are unknown is suggested. Having obtained the statistic by this heuristic principle, it can be justified by showing that to order \( N^0 \), its mean and variance agree with the chi-square distribution to which it is referred.

An empirical sampling experiment has been performed to demonstrate the greater robustness of the modified criterion. The relative power of the standard Bartlett test and the modified test have been compared for normal-parent populations to insure that the "price" of robustifying the test has not been too great.

As an auxiliary result of dealing with moments of permutation distributions of various test criteria, it has been possible to calculate the deviation of the actual type I error from the nominal type I error for a number of other standard tests. For some tests whose robustness has been previously evaluated by rather complex methods, this technique affords an easily evaluated result which may be compared with
the former results. For other tests, tables of the correct type I error have been prepared for the first time.

To give a perspective of the effect of non-normality and variance heterogeneity on a large number of standard tests, excerpts have been taken from several published tables of the same type.
Chapter I

GENERAL REVIEW OF LITERATURE ON NON NORMALITY AND PERMUTATION TESTS

1.1 Effects of Non Normality

Statistical tests are often based on the assumption of normality of the parent population. This is justified on various grounds.

(1) Some distributions met with in practice appear to be not unlike the normal distribution.

(2) While some system of unit error might or might not be normal, the appropriate linear function of these unit errors could be expected to approach normality.

(3) The normality assumption often makes an otherwise insoluble problem mathematically tractable.

(4) In some instances the tests have been shown to be relatively insensitive to moderate departures from normality.

For example, the normal distribution is justified in a test comparing two means by the central limit theorem, when the variance may be substantially assumed known from a large amount of previously accumulated data.

That the third justification for using normal-theory tests in practice is untenable hardly needs comment.

In refutation of the first justification, it may be pointed out that Karl Pearson developed a comprehensive system of frequency curves based on the first four moments to describe a wide range of naturally occurring phenomena. Of the many non-normal data which have been
published, and which illustrate the original motivation for developing these frequency curves, the Monier-Williams data on percentage butter fat, reported by Tocher (1928), clearly demonstrates the varied types of frequency curves found in experimental data.

Pearson used departures of the standardized moments from normal-theory values as a measure of non-normality. For our purposes it is often more convenient to use functions of the moments themselves, cumulants, since for the normal distribution all cumulants above the second are zero. Therefore the value of the standardized cumulant itself measures the departure from normality. In the above "standardized" implies that the r'th moment about the mean or the r'th cumulant is divided by the r'th power of the standard deviation of the distribution.

The second justification for the use of normal-theory procedures holds only for a very limited class of tests. The possible dangers of using the assumptions of normality in the more modern tests developed by R. A. Fisher, and Student, eg. analysis of variance, was early appreciated by such writers as E. S. Pearson (1931) and P. R. Rider (1929, 1931). For whereas the central limit theorem could be used to justify the normality assumptions on tests which depended only upon the distribution of the mean, no such theorem was available for tests involving a ratio of the mean and the sample standard deviation as found in the t test and analysis of variance.

Indeed, when non-normality occurred, it was known that the sample variance did not follow, even approximately, the distribution it would
if the population were normal. This is exemplified by the fact that although the mean value of the sample variance is independent of the parent distribution, its variance depends directly on the standardized fourth moment, \( \beta_2 = \mu_4 / \mu_2^2 \). Furthermore investigations of Le Roux (1931) and Sophister (1928) showed the distribution of the sample variance to be extremely dependent upon the type of parent distribution involved.

However, these facts did not necessarily throw much light on what was to be expected for quantities like those involved in the analysis of variance tests, which are in the form of ratios. E. S. Pearson (1931) calculated the mean and variance of the F statistics of the analysis of variance. He obtained the rather surprising result that the mean and standard deviation of the statistic (for general non-normal parent distributions) were the same to order \( N^{-1} \) as when parent normality is assumed.

This result is particularly surprising as it seemed to come about due to the canceling of one effect by another. The two effects of non-normality were to markedly change the distribution of the sample variance and to destroy the independence of the numerator and denominator. Pearson noted, however, that the effects of these two phenomena largely cancelled. This confirmed the small effect of non-normality found earlier by Rider (1929), who considered the behavior of the two-sided t test when sampling from an approximately rectangular distribution. He sampled from a population in which the integers 0, 1, ..., 9, were all equally likely.
Subsequent investigations were made by R. C. Geary (1936), who emphasized that the apparent insensitivity to normality was not shared by the one-sided t test in which the sample mean was compared with some standard value. He also showed that insensitivity to normality was not to be expected in the comparison of two sample variances.

These results were confirmed by Gayen (1949, 1950) who assumed the parent distribution followed the Edgeworth series. In spite of these many investigations, discussions of the role of the assumption of normality in the validity of statistical tests have failed to point out that those tests which are insensitive to these assumptions are those which compare two or more sample means.

This matter is clarified in a paper, largely overlooked, by J. W. Tukey (1948) in which he points out that there is a need to develop tests in many cases to make them more independent of nuisance parameters. He specifically indicates the need to studentize tests for variances for the population fourth moment.

Sensitivity to normality, already demonstrated by E. S. Pearson, R. C. Geary and A. K. Gayen in the case of comparing two variances, has recently been shown to be extreme when a large number of variances are compared, G. E. P. Box (1953).

1.2 Nature and Use of Permutation Tests

The theory of testing hypotheses, as developed by Neyman and Pearson, (1928), rests on the properties of likelihood. In their methods some parent distribution is assumed and a test is derived on
the principle that, while keeping the probability of rejecting the null hypothesis when it is true at a fixed value, the test should be selected which will be most likely to reject this hypothesis when some specified alternative is true. Tests possessing this property have been developed using the "likelihood ratio" principle.

Tests so derived might or might not be sensitive to the assumptions made; there is nothing in the formulation of the problem to ensure insensitivity to the assumptions. In fact it is known, Box (1953) for example, that the likelihood ratio test for differences in means (analysis of variance) is extremely insensitive to the assumption of normality, whereas the corresponding likelihood ratio test for differences in variances, L₁ or Bartlett's form of this test, is extremely sensitive to this assumption.

Fisher says, in his Design of Experiments (1935),

It has been mentioned that "Student's" t test, in conformity with the classical theory of errors, is appropriate to the null hypothesis that the two groups of measurements are samples drawn from the same normally distributed population. This is the type of null hypothesis which experimenters, rightly in the author's opinion, usually consider it appropriate to test, for reasons not only of practical convenience, but because the unique properties of the normal distribution make it alone suitable for general application. There has, however, in recent years, been a tendency for theoretical statisticians, not closely in touch with the requirements of experimental data, to stress the element of normality, in the hypothesis tested, as though it were a serious limitation to the test applied. It may, nevertheless, be legitimately asked whether we should obtain a materially different result were it possible to test the wider hypothesis which merely asserts that the two
series are drawn from the same population, without specifying that this is normally distributed.

Here, Fisher has already pointed out the necessity for investigating the sensitivity of a normal-theory test to parent non-normality, an issue which many subsequent authors have clearly missed.

Fisher goes on to say,

In these discussions it seems to have escaped recognition that the physical act of randomization, 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.

This remarkable test of R. A. Fisher, which depends upon the values in the sample alone, has come to be known as the randomization or permutation test; we shall use the latter expression. In such a test one evaluates all the differences in means which could have been generated by rearrangement of the sample and considers the proportion of cases which are more extreme than that observed.

Fisher demonstrated that for the particular example he considered, (an experiment due to Darwin concerning cross fertilized and self fertilized plants tested in pairs), that the null probability by the permutation tests 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) says, "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."

E. S. Pearson (1937) questioned the emphasis which Fisher gave to the importance of the permutation principle in the above statement. He said,

I am concerned ... with the question of whether there is something fundamental about the form of the test suggested, so that it can be used as a standard against which to compare other more expeditious tests, such as Student's. It seems to me that Fisher is overstating the claim of an extremely ingenious device ....

Pearson goes on to question Fisher's view that randomization is the central principle of test construction since it is possible, for example, to produce many permutation tests for comparing position parameters by using position statistics other than the mean, such as the midrange (midpoint), etc. He cites an example of a sampling experiment performed with a rectangular parent population in which the permutation test based on the midrange detects alternative hypotheses more often than does this test using the sample mean as the statistic.

Pearson goes on to develop his thesis by saying,

Now of course in practice it is extremely unlikely that we should deal with variables whose probability distribution is rectangular, but I have introduced these examples because it seems to me to suggest that in problems of this kind it is impossible to make a rational choice between alternative tests unless we introduce some information beyond that contained in the sample data, i.e. some information as to the kind of alternatives with which we are likely to be faced.
He concludes by saying,

It is true that when variation departs from the normal, the test will not give quite accurate control of the risk of wrong rejection of $H_0$ (although the error will usually be small), while a test based on randomization will continue to do so. It is in this that the value of the randomization test lies; but as I have pointed out, in so far as this latter test is applied to means, it cannot be regarded as unique, and for wide departures from normality it could probably be improved by use of other central estimates.

We thus have two separate ideas which at first sight seem to provide a paradoxical situation. If we are prepared to assume a particular parent distribution, we could pick out a "best" test in the Neyman-Pearson sense, but it may be more sensitive to this assumption than we would like it to be. Whereas if we do not assume any particular parent distribution, we can use a permutation test which is exact in the sense that in using it the null hypothesis will be rejected in only a stated percentage of the time when it is true, but we have no clear guidance as to what statistic to base the permutation test on.

The principle of the permutation test was developed further by Pitman (1936, 1937) and Welch (1937, 1938). They compared the permutation distribution of the $t$ statistic and the Beta form of the analysis of variance statistic, when the null hypothesis was true, with the normal-theory distribution of these statistics, by calculating the moments of the permutation distribution and moments of the normal-theory distributions of these test criteria.
1.3 Rejection of Preliminary Tests and Introduction of Robustness

Numerous publications emphasizing the assumptions made in statistical tests, e.g. Eisenhart (1947) have sometimes led users of the tests to be in a state of nervous agitation concerning whether or not they are justified in using them. Results of Pearson, Geary, Gayen and Box all indicate that in some cases these worries may be justified.

One method by means of which the user of statistical tests has attempted to assuage his conscience is by performing preliminary tests to determine whether he is justified in making the assumptions required in the main test on the same data. It has been pointed out however, Box (1953), that such procedure is unsatisfactory. Consider, for example, the analysis of variance test which is insensitive to departures of $\beta_2$ from the normal-theory value as opposed to the Bartlett test for comparing variances, which is extremely sensitive to such departures. Now suppose the analysis of variance were to be applied in the situation when $\beta_2$ were large and suppose the experimenter decided that since the analysis of variance assumed equal variances within groups, he would first perform the Bartlett test. Even if the group variances were really equal, the Bartlett test would tend to show a significant result because the population $\beta_2$ was larger than the normal value; consequently the experimenter might be afraid to apply the analysis of variance, because he thought the assumption of equal variances in the main test was not satisfied. In fact the large value of $\beta_2$ which had caused the misleading conclusion with the Bartlett
test would have had little effect in upsetting the analysis of variance test so that the experimenter would have been completely mislead by this series of tests.

Furthermore, it has been shown by Welch (1937), David and Johnson (1951), and Box (1953) that even if differences in variances had occurred, this would have produced little effect on the analysis of variance, providing (as would often be the case) the groups were of equal size. It seems that the principle of using one test to see if another test should be performed is a bad one. If carried to its logical conclusion, it could lead to an endless regression of tests; the test for equality of means could be preceded by a test for equality of variances, this in turn could be preceded by test for normality, the test for normality by a test for the independence of observations, etc.

The final result of such a series of tests would be some complicated function of the power of the various component criteria and their sensitivity to assumptions; and as has been demonstrated above could lead to incorrect conclusions. What seems to be required is that a test should stand on its own feet. The outcome of statistical tests takes the form of probability statements, and the human mind cannot appreciate small differences in probability. Therefore it is not necessary to insist on exactness, but only on avoidance of gross and misleading errors. The test should be such that departures from assumptions of the type and order to be expected in practice would not affect it unduly. Tests which possess this property of being reasonably independent of assumptions may be called "robust".
One method of devising such tests would be to specify a parent population which was sufficiently elastic to take into account all the situations likely to be met in practice so that quantities measuring differences in variances, departures from normality, etc. appeared in the test criterion itself. In practice it is extremely difficult to specify such parent distributions which provide criteria whose sampling distributions are mathematically tractable.

The so called randomization procedure and in particular the simplification which comes about by approximating the randomization distribution using its moments may be used to provide robust test criteria. To demonstrate this we must consider the nature of the permutation test argument.
Chapter II
PERMUTATION TEST FROM THE POINT OF VIEW OF THE NEYMAN - PEARSON
THEORY OF TESTING STATISTICAL HYPOTHESES

2.1 Introduction

According to the Neyman-Pearson theory (1933) we should develop statistical tests from the following considerations. We wish to test some hypothesis, $H_0$, concerning the nature of the probability law governing $N$ observations, $x_1, x_2, \ldots, x_N$ (for example that the sample observations are drawn from a normal universe with mean, $\mu = 0$ and unknown variance, $\sigma^2 > 0$) and we have in mind some alternative hypothesis $H_1$ (for example that the sample observations are drawn from a normal universe with mean, $\mu_1 > 0$, and unknown variance, $\sigma^2 > 0$). To do this we select a region, $w$, called the "critical region" in the sample space, and adopt the rule that if the sample point is contained in $w$, the null hypothesis will be rejected, otherwise it will be accepted.

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, $\alpha$, chosen in advance. This value, $\alpha$, is called the risk of error of the first kind, that is the error of rejecting $H_0$ when it is true.

(ii) When $H_1$ is true the chance of rejecting $H_0$ will be as large as possible. This latter chance is called the power of the test under the alternative, $H_1$; if the power is subtracted from unity, we have the risk of error of the second kind, $\beta$, the
error of failing to reject $H_0$ when it is false.

When a region can be found such that both conditions are satisfied we have a "best" critical region. Suppose we denote by $p_0(x_1, x_2, \ldots, x_N)$ the probability law when $H_0$ is true and by $p_1(x_1, x_2, \ldots, x_N)$ the probability law when $H_1$ is true. Then according to Neyman and Pearson we should choose the region, $w$, such that the two following conditions are satisfied.

\begin{equation}
(1) \int \ldots \int_{(w)} p_0(x_1, x_2, \ldots, x_N) \, dx_1 \, dx_2 \ldots \, dx_N = \alpha
\end{equation}

and

\begin{equation}
(2) \int \ldots \int_{(w)} p_1(x_1, x_2, \ldots, x_N) \, dx_1 \, dx_2 \ldots \, dx_N = 1 - \beta
\end{equation}

is a maximum. For the sake of brevity, we shall adopt the following vector notation for the above $n$-fold integrals.

\begin{equation}
(i) \int_{(w)} p_0(x) \, dX = \alpha
\end{equation}

\begin{equation}
(ii) \int_{(w)} p_1(x) \, dX = 1 - \beta
\end{equation}

where: $X$ = the vector, $(x_1, x_2, \ldots, x_N)$

Let us consider first condition (i). In the Neyman-Pearson development $p_0$ is assumed to have some stated form. For example:

\begin{equation}
p_0 = (2\pi \sigma^2)^{N/2} \exp\left[ - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_0)^2 \right]
\end{equation}

The test is derived so that condition (i) is satisfied, provided the assumption about the parent distribution is correct. However, if the
assumption concerning the form of \( p_0 \) were not true, then (i) might or might not be approximately satisfied, depending on the type of region, \( w \), being considered; see for example Box (1953).

2.2 Permutation Test

Suppose that the sample \((x_1, x_2, \ldots, x_N)\) be designated by the vector, \( X \), the permutations of the sample by \( X_i \) \((i = 1, \ldots, N!)\) and the set of these \( N! \) permutations by \( S(X) \) or \( S \). We shall denote the probability density associated with the sample, \( X \)'s, by \( p_0(X_i) \).

Now we may define the conditional probability,

\[
p_0(X_i / S) = \frac{p_0(X_i)}{\sum_{i=1}^{N!} p_0(X_i)}
\]

If we assume that all permutations of a given sample have the same probability density, \( i.e. \)

\[
p_0(X_i) = p_0(X_i) \quad i = 1, 2, \ldots, N!
\]

then

\[
p_0(X_i / S) = \frac{1}{N!}
\]

To construct a critical region, \( w \), choose, if possible, an integer of such that \( \alpha = q / N! \), which satisfies condition (1). Then arrange that \( q \) out of the \( N! \) permutations of each set, \( S(X) \), are contained in \( w \), and \((N! - q)\) are outside \( w \). The type I error may then be written as
\[
\int \frac{q}{\sum_{i=1}^{\Delta} \int p_0(x_i/s) \cdot p_0(x) \, dx} = \frac{q}{N!} \int p_0(x) \, dx = \alpha.
\]

where: \( \Delta \) = the entire sample space.

In picturing the critical region defined by (12), consider collecting in clusters all samples (vectors) which differ only in the ordering of the values, but having the same set of values in some order. In each of these clusters, \( S(x) \), select \( q/N! \) vectors, \( x_1 \), to be contained in the critical region, \( \omega \), and the remainder to be outside the critical region. Having chosen \( \alpha \) per cent of each cluster to be in the critical region, the probability of any sample in the entire sample space being in the critical region is \( \alpha \).

2.3 Permutation Tests with More than One Population

In the previous section we have assumed that the probability distribution under the null hypothesis is such that the probability density is the same for every rearrangement of all the observations. Such a null hypothesis would be appropriate, for example, when two or more treatments were tested on specimens of reasonably homogeneous basic material; two specific cases would be the t test to compare two treatments and the one-way classification analysis of variance to compare \( k \) treatments.

A somewhat different situation occurs when two or more treatments are compared within \( k \) blocks containing \( n \) observations each, for a total of \( N = nk \) observations. Since the observations within each block
are presumed more homogeneous than observations not in the same block, the probability density under the null hypothesis is not the same for all N observations, because a different block parameter occurs in each of the k groups. In such a case it would be appropriate to assume that only rearrangement within a block would leave the null probability density unchanged.

Provided therefore, that each rearrangement within each block has the same probability density, we could again construct a region, w, of size \( \alpha \) by arranging that q out of the \((n!)^k\) within-block permutations of the N observations are contained in w, with \( q = \alpha(n!)^k \). The type I error may then be denoted by the N-fold integral analogous to (9)

\[
\int_{(A)} \frac{k}{n} p_{0j}(X) \cdot \sum_{j=1}^{q} \frac{k}{n} p_{0j}(X_j/S) \, dX = \alpha
\]

(10)

In this and the previous type of permutation test discussed in section 2.2, if q is not an integer, we can of course obtain a test at approximately the desired significance level by taking the number of arrangements included in the critical region to be the nearest integer to q. Equations (9) and (10) are essentially of the same form and we include both types of test in the discussion which follows.

Now we need make no assumption about the form of the distribution function \( p_0(X) \) except that

\[
p_0(X_1) = p_0(X)
\]

(11)

for same set, \( X_1, X_2, \ldots, X_\lambda \), of rearrangements of the sample, X.
(where $\lambda$ = number of permissible rearrangements of the sample) That is to say that samples containing the same observations will have the same probability density for several orderings of the observations. This would include, for example, all hypotheses of the form

$$P_0(X) = \prod_{i=1}^{N} f(x_i)$$

(12)

i.e. each observation is distributed independently with the same density function, whatever its form. It would also include, for example, the hypothesis that the observations were distributed multivariate normally and were all equally correlated. It would, on the other hand, exclude the hypothesis that the observations were serially correlated.

2.4 Form of Permutation Test to be Used

By using a permutation test, then, we can satisfy exactly Neyman and Pearson's first condition without seriously restrictive assumptions. We now consider the second condition that the critical region, $w$, should be chosen such that the chance of accepting the alternative, $H_1$, when it is true, shall be as large as possible.

As was originally pointed out by E. S. Pearson (1937), by using the permutation test the probability of rejection of the null hypothesis when it is true will be maintained at the desired level, $\alpha$. However, the probability of rejection of the null hypothesis, when the alternative hypothesis is true, and consequently the power will depend directly on the specific form of the parent population. Since the configuration of the critical region, $w$, in the entire sample space, $\Omega$, 


depends upon the statistic used in the permutation test, the power also
depends upon the choice of this statistic. We can only satisfy Neyman
and Pearson's second condition, therefore, if we are prepared to be
specific about the class of probability density functions which we had
in mind in our alternative hypothesis, and base our method of selection
of the critical region on this parametric alternative hypothesis. If
the alternative, $H_1$, is so specified, Lehman and Stein (1949) have
shown how it is possible 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 distribu-
tion against the alternative that they came from two different
distributions, one of which had a larger location parameter than did
the other. The assumption that the form of the alternative distribu-
tion was the normal would lead to a test in which for each sample
the q points in the critical region were those for which the largest
differences in means (in the anticipated direction) occurred. If the
distribution were assumed to be rectangular, a more powerful test would
be based on the comparison of midranges.

In practice the statistician's feeling about the parent distribu-
tion could usually best be expressed in terms of a distribution of
possibilities rather than in any one particular possibility. This
mental, "prior distribution of distributions", might be imagined to
have some central value, but its range would, or perhaps should, make
the statistician reluctant to treat this value as if it were the only
one that could occur, especially if departure from this "central" distribution would lead to serious errors. In particular he would be reluctant to make the assumption of a specific parent null distribution if it were possible to show (as it is for example with tests to compare variances) that such an assumption could lead to serious inaccuracy in estimating the first kind of error, or significance level.

On the other hand, if \( \alpha \) is fixed at the desired level by the use of the permutation test, and being faced with the necessity for choosing some specific alternative distribution (or implying such a choice by "intuitive" selection of a criterion) it would seem natural for the statistician to base his criterion on a statistic appropriate for what he supposed to be the central alternative distribution. Even though the statistician might expect this distribution seldom if ever, to be realized exactly, he 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 midranges 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 this type of experiment was centered 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; hence, this would lead to a permutation test based on some different statistic.
2.5 Permutation Moments

Suppose we have selected some criterion, \( g(x) \), of the \( N \) observations in our experiment, then for any given sample a permutation distribution of \( g(x) \) generated by all permissible rearrangements of the observations is obtained and the \( h \)th permutation moment of this distribution is given by

\[
M_h(x) = \mathbb{E}_{\{\mathcal{S}\}} \left[ \frac{g^h(x)}{S} \right] = \frac{\lambda}{\pi} \sum_{i=1}^{\pi} g^h(x_i) \cdot p(x_i/S)
\]

(13)

where the summation is over all permissible arrangements and would include both the test situation of section 2.2 and that of section 2.3.

2.6 Overall Moments from Permutation Moments

The usual "overall" \( h \)th moment for the statistic or function \( g(x) \), is given by

\[
\mu_h = \mathbb{E}_{(\Delta)} \left[ g^h(x) \right]
\]

(14)

By use of (6), this may be written as

\[
\mu_h = \mathbb{E}_{(\Delta)} \left[ \mathbb{E}_{(\mathcal{S})} \left( g^h/S \right) \right]
\]

(15)

where: \( \Delta \) is entire sample space.

Substituting (13) in (15), we may write

\[
\mu_h = \mathbb{E}_{(\Delta)} \left[ M_h^\prime \right]
\]

(16)

Thus, as was originally indicated by Welch (1937), the overall moments may be evaluated by taking expectations of the permutation moments.
2.7 Example of a Permutation Test

At this point it may be helpful to study a particular example of the permutation test. Suppose that an experiment has been carried out in which k pairs of observations have been made. One observation within each pair has been made with treatment A applied and one with treatment B. Apart from application of the treatment, conditions have been kept as uniform as possible within each pair but differences in average level may occur from pair to pair. These data may be denoted by

<table>
<thead>
<tr>
<th>Block</th>
<th>Treatments</th>
<th>Treatment Differences (A-B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>1</td>
<td>(x_{11})</td>
<td>(x_{12})</td>
</tr>
<tr>
<td>2</td>
<td>(x_{21})</td>
<td>(x_{22})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>(k)</td>
<td>(x_{k1})</td>
<td>(x_{k2})</td>
</tr>
<tr>
<td>means</td>
<td>(x_{\cdot1})</td>
<td>(x_{\cdot2})</td>
</tr>
</tbody>
</table>

This is the familiar situation encountered in the "paired observation" t test. It can equivalently be regarded as an example of a randomized block design having k blocks and n = 2 treatments with a total of (kn = N) observations. The null hypothesis is that discussed
in section 2.3, that within each pair the probability density is unchanged by interchanging the observations. If the alternative hypothesis were that

\[ x_{ij} = \mu + \tau_j + \beta_i + \varepsilon_{ij} \] (17)

where: \( \mu \) = mean of all the observations
\( \tau_j \) = treatment effect, a constant \( (j = 1, 2) \)
\( \beta_i \) = block constants \( (i = 1, 2, \ldots, k) \)
\( \varepsilon_{ij} \) = normal \( (0, \sigma^2) \) independent random variable

then Ishman and Stein (1949) show that the best critical region is that based on the difference between the sample means, \( x_{1} - x_{2} = y \), referred to the distribution generated by all permissible permutations of the \( N \) observations, i.e., all possible rearrangements within blocks. This is equivalent to basing the test on the mean of the \( k \) differences, \( y_{1}, y_{2}, \ldots, y_{k} \), where the observed mean differences is referred to the \( 2^k \) mean differences generated by associating all possible combinations of plus and minus signs with the \( k \) differences, \( y_{i} \). This is the form of the test as originally suggested by R. A. Fisher (1935).

As a specific example suppose that in a particular experiment in which there were \( k = 10 \) pairs of numbers, the absolute treatment differences (the differences without regard to sign) were 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Then there are \( 2^{10} = 1024 \) possible ways in which we can attach a positive or negative sign to these numbers and consequently \( 2^{10} \) values of the mean, \( y \). These values constitute the permutation distribution of \( y \), and this is shown for the present example in Figure 1.
Figure 1
Permutation Distribution of The Sample Mean

Figure 2
Permutation Distribution of The Statistic, W
Suppose \( \delta \) is the true difference in the means of the populations, then to test the null hypothesis

\[ H_0: \; \delta = 0 \]

against the alternative

\[ H_1: \; /\delta/ > 0 \]

at the significance level, \( \alpha = 50/1024 \), we would include in the critical region those samples that give values of \( /y_\cdot/ > 3.7 \), since this inequality is true for the 50 most extreme values lying outside the arrows in the distribution in Figure 1. Alternatively, the region, \( y_\cdot > 3.7 \), would supply the critical region at the level of significance, \( \alpha = 25/1024 \), when testing against the alternative

\[ H_1: \; \delta > 0 \]

It will be noted that the same critical region would be obtained had we calculated the permutation distribution for the \( t \) statistic itself because

\[
t = \frac{y_\cdot}{\sqrt{\frac{1}{k(k-1)} \sum (y - y_\cdot)^2}}^{1/2}
\]

(18)

Equivalently, if we were testing the double-sided alternative hypothesis, we could have used the corresponding analysis of variance criterion,
\[ F = t^2 = \frac{\frac{k y^2}{*}}{(S - ky^2) / (k-l)} \]  

\[ \frac{2}{k} \sum_{j=1}^{2} (x_{*,j} - x_{*}*^2 / 1} \]

\[ = \frac{k}{\sum_{j=1}^{2} (x_{ij} - x_{*,j} - x_{*,}^2 / (k-l)} \]

\[ = \frac{S_t / l}{S_e / (k-l)} \]  

where:

\[ S_t = \text{treatment sum of squares} \]

\[ S_e = \text{error sum of squares} \]

Alternatively we could have used the criterion

\[ W = \frac{S_e}{S_e + S_t} = \frac{\sum y^2 - ky^2}{\sum y^2} \]  

which may be written as

\[ W = \frac{1}{1 + (k-l)t^2} \]  

which is a monotonic decreasing function of \( t^2 \).

For this example, using equation (22),

\[ W = \frac{385 - 10y^2}{385} \]

The permutation distribution of \( W \) can be calculated from that of \( y \), and is shown in Figure 2.
2.8 Approximation to the Permutation Test

Evaluation of the permutation distribution, or of such part of it as is necessary to determine the critical value of the statistic, using the procedure above is laborious. To make the permutation theory of practical value, use is made of the approximation to the permutation distribution based on the evaluation of its moments by the method proposed by Fitman (1936, 1937) and by Welch (1937, 1938).

Let us consider the particular test considered above. Of the several statistics demonstrated above, we shall derive the moments of the criterion, \( W \), because expectations of a ratio are not required in this case due to the constant denominator over the permutation distribution. Using the \( W \) statistic we have for its moments with respect to the permutation distribution

\[
M_1'(W) = \frac{k-1}{k} \quad (24)
\]

\[
M_2(W) = \frac{2(k-1)}{k^2(k+2)} \left[ 1 - \frac{(b_2-3)}{k-1} \right] \quad (25)
\]

The corresponding moments assuming normality are

\[
E(W) = \frac{k-1}{N} \quad (26)
\]

\[
V(W) = \frac{2(k-1)}{Nk^2(k+2)} \quad (27)
\]

It will be noted that the first moment of the permutation distribution is the same as that for the distribution based on the assumption of normality, while the variance of \( W \) in the permutation distribution
differs from that of the distribution based on normal theory in a term of order $k^{-1}$, involving the sample value of the fourth moment ratio,

$$b_2 = \frac{(k+2) \sum y^4}{(\sum y^2)^2}$$ (28)

It will be recalled that for normal theory, $W$ follows a Beta distribution, the cumulative distribution of which may be written

$$p(W < W_0) = I_{W_0} \left( \frac{1}{2} (k-1), \frac{1}{2} \right)$$ (29)

where Karl Pearson's notation for the incomplete Beta ratio is

$$I_x(p, q) = \frac{\int_0^x t^{p-1} (1-t)^{q-1} \, dt}{\int_0^1 t^{p-1} (1-t)^{q-1} \, dt}$$ (30)

The moments of the latter distribution are known to be

$$E(t) = \mu_1' = \frac{p}{p + q}$$ (31)

$$V(t) = \mu_2 = \frac{pq}{(p+q)^2 (p+q+1)}$$ (32)

Solving these in order to express $p$ and $q$ in terms of $\mu_1'$ and $\mu_2$, we obtain

$$p = \frac{\mu_1' (\mu_1' - \mu_1' - \mu_2)}{\mu_2}$$ (33)

$$q = \frac{(1 - \mu_1') \, p}{\mu_1'}$$ (34)
The permutation distribution of $W$ is of course discontinuous. However, its values lie between 0 and 1 and Pitman (1937) has shown that its third and fourth moments agree reasonably closely with those of a Beta distribution. It is therefore reasonable to approximate the permutation distribution by a Beta distribution, equating the first two moments of the two distributions. These are given in (24) and (25) for the permutation distribution and in (26) and (27) for the normal distribution.

This procedure will in general have the effect of changing the parameters, $p$ and $q$, in the approximating distribution from their normal theory values of $(k - 1)/2$ and $1/2$. However, since from (34)

\[
\frac{1 - \mu_1}{\mu_1'} \text{ and from (24) and (26) the mean of } W \text{ is the same for the permutation distribution as for the normal-theory distribution, it follows that both parameters, } p \text{ and } q, \text{ will be changed by the same factor. We will denote this factor by the symbol, } d.
\]

Thus we may write the type I error of the approximate permutation test in the form

\[
P(W < W_0) = \frac{1}{W_0} \left( \frac{1}{2} (k - 1) d, \frac{1}{2} d \right) = \alpha
\]

(35)

where the modifying constant

\[
d = 1 + \frac{(k + 2)(b_2 - 3)}{k(k + 2 - b_2)} = 1 + \frac{b_2 - 3}{k \left[ 1 - \frac{b_2}{k + 2} \right]}
\]

(36)
or approximately for large \( k \)

\[
d = 1 + \frac{b_2 - 3}{k}
\]  

(37)

We can now transform (35) back to the \( F \) form. Thus, finally we have that, as an approximation for the permutation test, we should perform the usual \( F \) test but instead of employing \( 1 \) and \( k - 1 \) degrees of freedom. In the example discussed above we find that with the observed values of the absolute treatment differences, \( 1, 2, \ldots, 10 \), and with the critical region, \( w \), defined by \( |y| > 3.7 \), the error of the first kind for the permutation test is \( 4.88 \) per cent. Using the approximation to the permutation test, we get \( 4.95 \) per cent as opposed to the normal theory value of \( 4.52 \) per cent.

2.9 The Critical Region for the Approximate Permutation Test

It will be noted that we have defined our test procedure for the approximate permutation test and hence the critical region, \( w \), in terms of whether or not the inequality

\[
F > F_\alpha(b_2)
\]  

(38)
is satisfied, where \( F_\alpha(b_2) \) refers to the \( \alpha \) percentile point of the \( F \) table with degrees of freedom modified by the factor, \( d \), which depends upon the sample value of \( b_2 \). This is in contrast with the more conventional test which defines the critical region, \( w \), as those values which satisfy the inequality

\[
F > F_\alpha
\]  

(39)
in which the left hand side of the inequality depends upon the observations and the right hand side is some fixed tabled value independent of the sample.

Tests of the sort defined by (38) have been suggested for other problems. For example Welch (1937, 1949, 1951) has proposed tests for homogeneity of means when the variances may differ. These are of the form

\[ F > F_\alpha(s_1^2, s_2^2, \ldots, s_k^2) \]  \hspace{1cm} (40)

where \( s_1^2, s_2^2, \ldots, s_k^2 \) are the sample variances.

There is of course no particular reason, except perhaps convenience, for preferring tests given inequalities of the form of (39). All we are really concerned with is to define a critical region, \( W \), in the sample space. It is of some interest to consider the critical region for the modified tests of the last section and compare it with the normal-theory critical region.

Unfortunately it is possible to show the critical region geometrically only for \( k \), 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 is to be justified only on the grounds of an approximation to the permutation test, it too will be of little value. However, we shall see later that, to some extent, the modified criterion can be justified independently of the fact that it approximates the permutation test.
The critical region of any test independent of scale must necessarily be conical (since a set of observations, \( y_1, y_2, \ldots, y_k \) has the same significance as a set, \((\lambda y_1, \lambda y_2, \ldots, \lambda y_k)\). Both the normal-theory test and the approximate permutation test are of this type and their critical regions for \( k = 3 \) are shown in Figures 3a and 3b.

Figure 3a is a sketch of the three-dimensional sample space. In Figure 3b a section of the critical region in the plane, \( y_1 + y_2 + y_3 = 10 \) is shown in greater detail. By including the scales of \( y_1, y_2, \) and \( y_3 \), we have a diagram in trilinear coordinates and we can see at a glance which types of samples are included in \( w \) by each of the two tests.

It can be seen that the approximate permutation test differs from the normal-theory test chiefly in selecting as significant samples, those samples which are found non significant by the standard \( t \) test because an outlier increases the variance. To take a hypothetical example with more observations, suppose that we have ten differences, nine of which are equal to +1 and the remaining one to +11. Then we find that \( t = 2.0 \) and entering the usual tables with one and nine degrees of freedom, the corresponding probability is \( 0.0766 \).

Using the modified test, we find that \( b_2 = 10.402 \) and \( d = 6.588 \); thus we should enter the tables with 6.6 and 59 degrees of freedom, which gives the corresponding probability of \( 0.00190 \). The exact randomization test would, of course, give a value of \( 1/512 = 0.00195 \), since this sample and the corresponding sample in which all the differences are negative are the most extreme possibilities out of the 1024.
Figure 3(a)

Normal Theory
and
Approximate-Permutation
Critical Regions
Figure 3(b)

Comparison of Normal-Theory
And Approximate-Permutation

Critical Region, \( W_{10} \).
2.10 Use of the Permutation Test to Estimate the Approximate Effect of Non Normality

We shall use the example of the test on paired observations discussed above to illustrate one further application of the permutation approach.

We have seen in section 2.7 that it is possible to derive the overall h'th moment of a function of the observations, \( g(X) \), by considering first the permutation moments of the function and then taking expectation of these permutation moments over all possible samples.

For example, from (22), we have the criterion

\[
W = \frac{Se}{Se + St} \tag{41}
\]

whose exact moments with respect to a general, non-normal, parent population of \( x \)'s are

\[
E(\hat{w}) = \frac{1}{k} \tag{42}
\]

\[
V(\hat{w}) = \frac{2(k - 1)}{k^2(k + 2)} \left[ 1 - \frac{1}{k - 1} \left[ E(b_2) - 3 \right] \right] \tag{43}
\]

where:

\[
b_2 = (k + 2) \frac{\sum y^4}{(\sum y^2)^2}
\]

In particular when the distribution is normal

\[
E(b_2) = \beta_2 = 3
\]
and we obtain the normal moments given by (26) and (27).

When \( E(b_2) \neq 3 \) we can as an approximation fit a beta distribution to the distribution of \( W \) and we have for the approximate probability when the distribution is not normal

\[
P(W < w) = I_w \left[ \frac{1}{2} (k - 1) \delta, \frac{1}{2} \right]
\]

(48)

where the expression for \( \delta \) is the same as for \( d \) except with \( b_2 \) replaced by \( E(b_2) \). That is

\[
\delta = 1 + \frac{E(b_2) - 3}{k \left[ 1 + \frac{E(b_2)}{k + 2} \right]}
\]

(45)

or approximately if \( k \) is large

\[
\delta = 1 + \frac{E(b_2) - 3}{k}
\]

(46)

As before we may now change the Beta function back to the \( F \) form and the approximate distribution under non normality of the statistic, \( t^2 \), is given by the \( F \) distribution with \( \delta \) and \( \delta(k - 1) \) degrees of freedom.

Since \( b_2 \) is in the form of a ratio, it is necessary to expand the denominator in a power series in a manner similar to that employed by Pearson (1931) and several other writers. Using this technique, we obtain to order \( k^{-1} \)

\[
E(b_2) = \frac{1}{k + 2} \left[ (k + 3)b_2 + 3b_2^2 - 2b_4 \right]
\]

(47)
Now with formulas (47) and (45), the approximate effect of non-normality may be estimated for various values of $k$ and $\beta_2$. Table 2 shows a few such values.

Table 2
Type I Error for Paired t Test
Nominal $\alpha = .05$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\beta_1^2$</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>.0552</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.0552</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>.0540</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>.0512</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.0512</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>.0518</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>.0508</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.0508</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>.0508</td>
</tr>
</tbody>
</table>
It should be noted that no value has been designated for the value of $\beta_4$. For practical purposes, forms of frequency functions are usually designated by the first four moments, though this characterization is not unique. Various classes of distribution functions may be used when fitting by moments, two of which seem to suggest themselves most readily for this situation.

The one assumption that could be made in this case would be to say that deviations from normality will be assumed to occur in the first four moments, while the sixth moment will be given its normal theory value, $\beta_4 = 15$. A second assumption would be that the fitted curve was of the Pearson type, which defines the fifth and higher moments as a function of the first four moments. Throughout this investigation we have used the latter convention.

2.11 Summary

In this chapter we have utilized the fact that the permutation test may be used to perform two useful functions.

1. It may be used to give a modified test procedure which will in effect compensate for departures from assumptions. This is effected by calculating a factor, $d$, a statistic used to modify the normal-theory degrees of freedom.

2. It can be used to provide approximate estimates of the effect of departures from assumptions of standard normal-theory tests. This is effected by calculating a factor, $\delta$, a parameter used to modify the normal-theory degrees of freedom.
We shall now consider these applications in more detail. In Chapter III modified t tests and analysis of variance tests based on the permutation moments of Pitman (1937) and Welch (1937, 1938) are indicated. In Chapter IV the approximate effects of departures from assumptions, using the procedures mentioned in section 2.10, are indicated. In Chapters V and VI these methods are applied to provide a robust test to compare variances.
Chapter III
MODIFIED t TESTS AND ANALYSIS OF VARIANCE

Using the methods described in section 2.8, approximate permutation tests have been derived for several tests comparing means. As indicated, these tests all have a critical region of the form

$$w: F > F_{\alpha}(X) = F_{\alpha}(d_{n1}, d_{n2})$$  \hspace{1cm} (47)

where: $d$ = a sample estimate of a measure of non normality or variance heterogeneity.

Therefore the test is uniquely defined by stating the formula for the modifying factor, $d$, as a function of the observed sample. These are given below.

3.1 Randomized Blocks

$$d = 1 + \frac{(ks - k + 2)V^2 - 2k}{k(s - 1)(k - V^2)}$$  \hspace{1cm} (48)

where:

$$V^2 = \frac{\sum_{i=1}^{k}(k_{2i} - k_{2o})^2}{(k - 1)k_{2o}}$$

$V$ = sample coefficient of variation of block variances

$k_{2i}$ = sample variance in the $i$'th block

$s$ = number of treatments

$k$ = number of blocks
3.2 Paired t Test (special case of randomized blocks)

\[ d = 1 + \frac{b_2 - 3}{k(1 - \frac{b_2}{k + 2})} \] (49)

where:

\[ b_2 = \frac{\sum_{i=1}^{k} y_{i1}^2}{\left( \sum_{i=1}^{k} y_{i1} \right)^2} \]

\[ y_{i1} = x_{i1} - x_{i2} \] = treatment differences in i'th block

3.3 Analysis of Variance (one way classification)

\[ d = 1 + c_2 \frac{\frac{N + 1}{N(N - 1)} + \frac{N + 1}{N - 1} A}{1 - c_2 \left( \frac{1}{N} - A \right)} \] (50)

where:

\[ A = \frac{N + 1}{2(k - 1)(N - k)} \left[ \frac{k^2}{N} - \frac{k}{N} \sum_{t=1}^{k} \frac{1}{n_t} \right] \]

\[ c_2 = \frac{k_4}{k_2^2} \]

\[ k_4, k_2 = Fisher's k statistics \]

\[ k = number of groups \]

\[ n_t = number of observations per group \]

\[ N = \sum_{t=1}^{k} n_t = total number of observations \]
Noting that for equal-sized groups, \( A = 0 \), the formula simplifies in this special case to

\[
d = 1 + \frac{(N + 1)c_2}{(N - 1)(N - c_2)}
\]  
(51)

It is interesting to note that all of these tests are asymptotically equal to the normal-theory test, because the modifying factor, \( d \), always tends to unity with increasing sample size. This then confirms the robust nature of the normal-theory tests for these cases.
Chapter IV

EFFECT OF NON NORMALITY AND UNEQUAL VARIANCES ON T TESTS AND
ANALYSIS OF VARIANCE

Using the methods described in section 2.10, modified tests may
be derived for situations where the parent populations deviate in a
specified manner from the parent populations assumed in the derivation
of standard tests. As indicated, these tests all have a critical
region of the form

\[ F > F_{x}(6n_{1}, 5n_{2}) \]  \hspace{1cm} (52)

where: \( \delta \) = a measure of the extent to which the standard assumptions
of normality and homoscedasticity fail to hold in the
postulated parent population.

Therefore the test is uniquely determined by the formula for the
modifying factor, \( \delta \), as a function of the type of parent population.
A factor, \( \delta \), near unity indicates that the standard test is very
nearly like the approximate permutation test, and therefore quite
robust with respect to the failure of the standard assumptions con-
considered. A value of \( \delta \) not near unity should warn the experimenter
against the use of the standard test if it is expected that the
assumptions will fail in the manner postulated. Then the use of a
more robust test is suggested; for some situations these are given in
Chapter III.
4.1 Formulae for $\delta$.

4.1.1 Randomized Blocks

Effect of Non-Normality:

$$
\delta = 1 + \frac{ks(s - 2) \lambda_2}{ks(s - 1)^2ks - k - 2 + k(s - 1)^2 \lambda_2}
$$

(53)

where:
- $k$ = number of blocks
- $s$ = number of treatments
- $\lambda_2$ = Population standardized 4th cumulant = $\left(\beta_2 - 3\right)$

For very large samples,

$$
\delta = 1 + \frac{2}{ks} \lambda_2
$$

(54)

Effect of Unequal Block Variances

$$
\delta = 1 + \frac{(ks - k + 2) E(A_1) - (s - 1)}{k(s - 1)(1 - E(A_1))}
$$

(55)

where:
- $E(A_1) = \frac{s + 1}{s - 1} C_2 + \frac{6}{s - 1} C_2^2 - \frac{8}{s - 1} C_3$

$$
C_2 = \frac{\sum K_{2i}^2}{\left(\sum K_{2i}\right)^2}
$$

$$
C_3 = \frac{\sum K_{2i}^3}{\left(\sum K_{2i}\right)^3}
$$

$K_{2i}^2 = \sigma_i^2$ = Population variance of $i$th block.
The Simultaneous Effect of Non-Normality and Unequal Block Variances

\[ s = 1 + \frac{(ks - k + a) E(A_1 + A_2) - (s - 1)}{k(s - 1)(1 - E(A_1 + A_2))} \]  \hspace{1cm} (56)

where:

\[ E(A_2) = \frac{\sum K_{21}^2 \sum K_{4i}}{(\sum K_{21})^4} \]

\[ - \frac{1}{s} \frac{\sum K_{4i} K_{21}}{(\sum K_{21})^3} \]

\[ + \frac{1}{s} \frac{\sum K_{4i}}{(\sum K_{21})^2} \]

where:  \( K_{4i} \) = Population 4th cumulant in i\textsuperscript{th} block.

Note:  \( E(A_1) \) represents the effect of inequality of variance, while \( E(A_2) \) represents primarily the effect of non-normality.

4.1.2 Analysis of Variance (One-way classification)

Effect of Non-Normality

\[ s = \frac{1 + 2N E(C_2)}{1 - \left[ \frac{1}{N} - \sum E(C_2) \right]} \]  \hspace{1cm} (57)

where:  \( k \) = number of groups

\( n_t \) = number of observations per group
\[ N = \sum_{t=1}^{k} n_t \quad \text{total number of observations} \]

\[ U = \frac{N}{2(k - 1)(N - k)} \left[ \frac{k^2}{N} - \frac{k}{\sum_{t=1}^{k} \frac{1}{n_t}} \right] \]

\[ E(C_2) = \lambda_2 + \frac{1}{N} \left( -3\lambda_2^2 - 10\lambda_2 - 12\lambda_1^2 - 2\lambda_4 \right) \]

\[ \lambda_r = \frac{k - \frac{2}{r + 2}}{k} \]

For equal sized groups the factor, U, is zero and the formula simplifies to

\[ \delta = \frac{1}{1 - \frac{1}{N} E(C_2)} = \frac{N}{N - E(C_2)} \]

Since it is desired to tabulate the effect of non-normality in terms of the third and fourth cumulants or moments, there is no unique value which can be assigned to the sixth cumulant occurring in the expression for \( E(C_2) \). Two reasonable approaches suggest themselves. The sixth cumulant may be given its normal value, 0, for all cases, or the fitted distribution may be restricted to the set of Pearson curves, in which case the sixth cumulant is a known function of the third and fourth cumulants. We shall adopt the latter convention.

The same problem arises in t test for paired treatments and the F test for comparing two variances. Here again the convention of using the Pearson system to determine the sixth moment will be used. This
The formula is

\[ \beta_4 = \frac{5 \left( \frac{1}{2} \beta_3 + \left(1 + \frac{1}{2} \alpha \right) \beta_2 \right)}{1 - \frac{3}{2} \alpha} \tag{58} \]

where

\[ \alpha = \frac{2\beta_2 - 3\beta_1 - 6}{\beta_2 + 3} \]

4.1.3 t Test for Comparing Two Means

The simple t test for comparing two means would be handled by one of the two preceding formulae, depending upon whether it has equal or unequal numbers of observations in the two treatments.

4.1.4 t Test for Paired Treatments (randomized blocks with 2 treatments)

\[ t = 1 + \frac{E(b_2) - 3}{k + 1 - \frac{E(b_2)}{k + 2}} \tag{59} \]

where

\[ E(b_2) = \frac{1}{k + 2} \left[ (k + 3)\beta_2 + 3\beta_2^2 - 2\beta_4 \right] \]

\( b_2 \) and \( \beta_2 \) = sample and population fourth standardized moments respectively of the differences between the two observations in each block.

4.1.5 F Test for Comparing Two Variances, with Means Known

\[ t = \frac{1}{\left[ 1 + \frac{1}{N} \cdot \frac{N - 1}{E(c_2)} \right]} - \frac{2}{N} \tag{60} \]
where: \( N = \) total number of observations

\[
E(C_2) = \lambda_2 + \frac{1}{N} \left( -3\lambda_2^2 - 13\lambda_2 - 20\lambda_1^2 \right) = 6
\]

4.1.6 F Test for Comparing Two Variances, with Means Unknown

\[
\delta = 1 + \frac{(N + 2)(1 + \frac{1}{2}\lambda_2) - N}{N - 2(1 + \frac{1}{2}\lambda_2)}
\]  \hspace{1cm} (61)

where: \( N = \) total degrees of freedom for two variance estimates.

This completes the formulae for the factor, \( \delta \), for all those modified F tests which appear in Section 3. For several of the other tables a factor, \( \delta \), is also appropriate for modifying degrees of freedom, but these are not included in Section 4 since they may be found elsewhere in the literature as indicated.

4.2 Tabled Values of Type I Error Where Failure of Assumptions Occur in Testing Means

While the deviation of the factor, \( \delta \), from unity is a measure of sensitivity of a test to failure of the assumptions, it is helpful to translate this effect for a few particular cases into the change in type I error. This is done readily with the aid of the formulae of the previous section by noting that the standard tests have a critical region of the form

\[
w: F > F_{\alpha}
\]

such that \( 1 - \int_0^{F_{\alpha}} dF(n_1, n_2) = \alpha \)
Since $F$ is not distributed as $F(n_1, n_2)$, but rather as $F(5n_1, 5n_2)$, the actual type I error when using the standard test is not $\alpha$, but rather

$$a = 1 - \int_0^F \alpha \, dF(5n_1, 5n_2)$$

(63)

With this use of the approximate permutation test, several tables have been prepared, giving the actual type I error when the standard test is applied in situations where specified deviations from the assumptions occur. Wherever possible excerpts have been taken from tables published elsewhere to give a more complete picture of the effects of failure of assumptions in a broad class of problems. In these cases, the approximate permutation theory does not permit evaluating these errors.

Two types of designs for comparing two or more means will be considered in this section, randomized blocks and the one-way classification analysis of variance. The effect of non-normality will be considered for both, and the effect of inequality of variances will be considered for the first of these. Also student's $t$ test for comparing a sample mean with a hypothesized value will be considered.

4.2.1 Randomized Blocks - Inequality of Block Variances

Exact values for the type I error in randomized blocks with unequal block variances are taken from Box (1951) and given below. These are compared with values obtained by the approximate-permutation test theory.
Table 3
Comparison of Exact and Approximate-Permutation Type I Error
for Randomized Blocks with Unequal Block Variances
Nominal $\alpha = .05$

<table>
<thead>
<tr>
<th>No. of Treatments (s)</th>
<th>No. of Blocks (k)</th>
<th>Block Variances ($k_{21}$)</th>
<th>Exact $\alpha$</th>
<th>Approx. $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>1, 2, 3</td>
<td>.0425</td>
<td>.0436</td>
</tr>
<tr>
<td>ii</td>
<td>5</td>
<td>1, 2, 3</td>
<td>.0427</td>
<td>.0426</td>
</tr>
<tr>
<td>iii</td>
<td>11</td>
<td>1, 1, 3</td>
<td>.0376</td>
<td>.0388</td>
</tr>
<tr>
<td>iv</td>
<td>5</td>
<td>1, 1, 3</td>
<td>.0391</td>
<td>.0411</td>
</tr>
<tr>
<td>v</td>
<td>3</td>
<td>1, 1, 1, 1, 1, 3</td>
<td>.0447</td>
<td>.0413</td>
</tr>
<tr>
<td>vi</td>
<td>3</td>
<td>1, 1, 1, ..., 1, 3</td>
<td>.0486</td>
<td>.0478</td>
</tr>
</tbody>
</table>

4.2.2 Randomized Blocks - Inequality of Treatment Variances

Approximate type I errors for randomized blocks with unequal variances in the treatments are taken from Box (1951) and given below. Permutation theory does not give an approximation for this situation.
Table 4
Type I Error for Randomized Blocks with Unequal Treatment Variances - Nominal $\alpha = .05$

<table>
<thead>
<tr>
<th>No. of Blocks \ (k)</th>
<th>No. of Treatments \ (s)</th>
<th>Treatment Variances \ ($k_{21}$)</th>
<th>Type I Error \ ($\alpha$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>11</td>
<td>3, 2, 3</td>
<td>.0549</td>
</tr>
<tr>
<td>ii</td>
<td>5</td>
<td>3, 2, 3</td>
<td>.0559</td>
</tr>
<tr>
<td>iii</td>
<td>11</td>
<td>1, 2, 3</td>
<td>.0593</td>
</tr>
<tr>
<td>iv</td>
<td>5</td>
<td>1, 2, 3</td>
<td>.0612</td>
</tr>
<tr>
<td>v</td>
<td>3</td>
<td>1, 2, 3</td>
<td>.0692</td>
</tr>
<tr>
<td>vi</td>
<td>3</td>
<td>1, 2, 3</td>
<td>.0709</td>
</tr>
</tbody>
</table>

4.2.3 Randomized Blocks - Non Normality

Permutation theory permits obtaining the following approximate values for the type I error.
Table 5

Type I Error for Randomized Blocks with Non-
Normal Parent Population - Nominal \( \alpha = .05 \)

<table>
<thead>
<tr>
<th>No of Blocks (k)</th>
<th>No. of Treatments (s)</th>
<th>Population Fourth Cumulant (( \lambda_4 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-1.5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.0775</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>0.0592</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>0.0535</td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>0.0516</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.0552</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>0.0514</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>0.0518</td>
</tr>
</tbody>
</table>

4.2.4 Paired t Test (Special case of randomized blocks)

Since there is only one degree of freedom for error within each
block, under the null hypothesis, it is not possible to separate the
effect of non normality and inequality of block variances in the case
of the paired t test. Therefore it will be convenient to consider
these two failures of the assumptions by non-normal third and fourth
moments of the variable, \( y \), the differences between the treatment
yields within each block.

Although these figures were given in Table 2 of section 2.10,
they will be repeated here for completeness of this section.
Table 6
Type I Error for Paired t Test
Nominal \( \alpha = .05 \)

<table>
<thead>
<tr>
<th>k</th>
<th>( \beta_1^2 )</th>
<th>( \beta_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>.0552</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.0552</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>.0540</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>.0512</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.0512</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>.0518</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>.0508</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.0508</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>.0508</td>
</tr>
</tbody>
</table>

4.2.5 Analysis of Variance (one-way classification) Effect of Non Normality

The effect of non normality on the analysis of variance for the comparison of \( k \) means was evaluated by Gayen (1950b) by the use of the Gram Charlier series to represent the parent populations. Selected values from his table are compared with results obtained by the simpler method of the approximate permutation test. Both results are given below.
Table 7

Type I Error for Comparing k Means in The Analysis of Variance (with 5 observations for each of 5 treatments)

Non Normal Parent Populations - Nominal $\alpha = .05$

<table>
<thead>
<tr>
<th>$\lambda_2$</th>
<th>0.00</th>
<th>0.50</th>
<th>1.00</th>
<th>1.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0</td>
<td>0.0524</td>
<td>0.0529</td>
<td>0.0534</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0524)</td>
<td>(.0521)</td>
<td>(.0516)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.0500</td>
<td>0.0505</td>
<td>0.0510</td>
<td>0.0515</td>
</tr>
<tr>
<td></td>
<td>(.0500)</td>
<td>(.0500)</td>
<td>(.0491)</td>
<td>(.0482)</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0452</td>
<td>0.0457</td>
<td>0.0462</td>
<td>0.0467</td>
</tr>
<tr>
<td></td>
<td>(.0568)</td>
<td>(.0528)</td>
<td>(.0468)</td>
<td></td>
</tr>
</tbody>
</table>

Note: upper values from Gayen, lower values permutation theory

4.2.6 Analysis of Variance (one-way classification)

Effect of Inequality of Variances

From Box (1951) we have taken a few values to indicate the effect of inequality of variances within groups for situations where the groups are equal and unequal. It should be noted that with equal groups the test shows the robustness which has been found in the previous tests, but that when a large or small variance occurs in a
group which has a very small percentage of the total observations, the effect of heteroscedasticity is quite severe.

Table 8

Type I Error for Comparing Three Means in The Analysis of Variance

<table>
<thead>
<tr>
<th>Group Variances</th>
<th>Number of Observations in Each Group</th>
<th>Total N</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_1^2 ) ( \sigma_2^2 ) ( \sigma_3^2 )</td>
<td>( n_1 ) ( n_2 ) ( n_3 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 1</td>
<td>5 5 5</td>
<td>15</td>
<td>.0587</td>
</tr>
<tr>
<td>1 1 3</td>
<td>7 5 3</td>
<td>15</td>
<td>.1070</td>
</tr>
<tr>
<td>1 1 3</td>
<td>9 5 1</td>
<td>15</td>
<td>.1741</td>
</tr>
<tr>
<td>1 1 3</td>
<td>1 5 9</td>
<td>15</td>
<td>.0131</td>
</tr>
</tbody>
</table>

4.2.7 t Test for Comparing Two Treatments

Effect of Non Normality

By use of permutation theory we may obtain the approximate effect of non normality on the t test.
Table 9
Type I Error for Comparing Two Means Non-
Normal Parent Populations - Nominal $\alpha = .05$

<table>
<thead>
<tr>
<th>No. of Blocks</th>
<th>$\lambda_2$</th>
<th>$\lambda_1^2$</th>
<th>$\lambda_1^1$</th>
<th>$\lambda_1^{1.5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-1</td>
<td>.0562</td>
<td>.0493</td>
<td>.0438</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>.0500</td>
<td>.0425</td>
<td>.0358</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>.0737</td>
<td>.0598</td>
</tr>
<tr>
<td>11</td>
<td>-1</td>
<td>.0534</td>
<td>.0520</td>
<td>.0510</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>.0500</td>
<td>.0485</td>
<td>.0471</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>.0555</td>
<td>.0492</td>
</tr>
<tr>
<td>21</td>
<td>-1</td>
<td>.0518</td>
<td>.0515</td>
<td>.0512</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>.0500</td>
<td>.0496</td>
<td>.0492</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>.0496</td>
<td>.0478</td>
</tr>
</tbody>
</table>

4.2.6 Testing a Mean Against A Hypothesized Value (Student's t Test)

While it is not possible to estimate the effect of non normality on the t test in the case of testing a value of the sample mean against a hypothesized value of the mean by the permutation-theory method, Geary (1936) has estimated the effect of non normality for samples of 10.
Table 10
Type I Error for One Sided t Test Sample

Size = 10  Nominal $\alpha = .025$

<table>
<thead>
<tr>
<th>$\lambda_2$</th>
<th>$\lambda_2$</th>
<th>Type I Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,</td>
<td>1</td>
<td>.024</td>
</tr>
<tr>
<td>1/2</td>
<td>0</td>
<td>.041</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>.072</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>.086</td>
</tr>
</tbody>
</table>

Here again there is insensitivity to the population fourth moment, but in contrast to tests for comparing two or more means, rather small deviations of the third cumulant from zero cause an appreciable change in probability.

4.3 Summary of Tests for Means

From the foregoing tables it can be seen that a large variety of tests for comparing means are highly robust to both non normality and inequality of variances. Therefore the practical experimenter may use these tests of significance with relatively little worry concerning the failure of these assumptions to hold exactly in experimental situations. The rather striking exception to this rule is the sensitivity of the analysis of variance to variance heterogeneity when the groups are of unequal size.
Student's t test for testing a sample mean against a hypothesized value, in contrast to the tests for comparing two or more means, is quite sensitive to skewness in the parent population.
Chapter V

TESTS ON VARIANCES, MEANS ASSUMED KNOWN

Since it has been demonstrated that tests for comparing variances are extremely sensitive to non normality, an approximate permutation test will be derived by the method of fitting a Beta distribution to the first two permutation moments as outlined in section 2.8.

5.1 Test to Compare Two Variances, Means 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. There is no loss of generality in assuming that the two known means are zero. The statistic most commonly used to test this hypothesis is

\[
F = \frac{\sum_{i=1}^{n_1} X_i^2 / n_1}{\sum_{j=1}^{n_2} Y_j^2 / n_2}
\]

(64)

where:
- \(X_i\) = the \(i\)-th observation of the first sample
- \(Y_j\) = the \(j\)-th observation of the second sample

An equivalent statistic for testing this same hypothesis is

\[
W = \frac{n_1 \sum_{i=1}^{n_1} x_i^2 \sum_{j=1}^{n_2} y_j^2}{\sum_{i=1}^{n_1} x_i^2 + \sum_{j=1}^{n_2} y_j^2}
\]

(65)
Since this statistic is an exact equivalent of the more common F, and lends itself to simpler algebraic manipulation, the theoretical development will be concerned with the distribution of W. As W is distributed as a Beta variate for samples drawn from a normal population, its first two moments are

\[ E(W) = \mu_1(W) = \frac{v_1}{v_1 + v_2} \]  \hspace{1cm} (66)\]

\[ V(W) = \mu_2(W) = \frac{2v_1 v_2}{(v_1 + v_2)^2(v_1 + v_2 + 2)} \]  \hspace{1cm} (67)\]

where \( v_1, v_2 \) = degrees of freedom

The moments of \( W \) with respect to the permutation distribution are

\[ M_1 = n_1/N \]  \hspace{1cm} (68)\]

\[ M_2 = \frac{2n_1 n_2}{N^2(N - 1)} \left( 1 + \frac{1}{2} c_2 \right) \]  \hspace{1cm} (69)\]

where: \( c_2 = b_2 = 3 \)

\[ b_2 = \frac{N(\sum x^4 + \sum y^4)}{(\sum x^2 + \sum y^2)^2} \]

\[ N = n_1 + n_2 \]

If these moments of the permutation distribution of \( W \) are equated to the corresponding normal moments, the following relationship is found:
\[ v_1 = n_1 d; v_2 = n_2 d \]  \hspace{1cm} (70)

where:
\[ d = \left[ \frac{1}{1 + \frac{1}{2} \frac{N - 1}{N}} \right] = \frac{2}{N} \]  \hspace{1cm} (71)

Since there is a one to one correspondence between the \( W \) statistics here investigated and the \( F \) statistic used to test the same hypothesis, this result suggests that the appropriate statistic is
\[ F(n_1 d, n_2 d) \]  \hspace{1cm} (72)

rather than the conventional
\[ F(n_1, n_2) \]  \hspace{1cm} (73)

That is, the appropriate statistic is the conventional \( F \) with degrees of freedom modified by the sample fourth moment as indicated.

There are however two questions concerning the \( F \) approximation to the permutation test.

1. How good is the moment approximation to the permutation test?
2. It is well known that when the parent distribution is normal, the standard \( F \) test is uniformly most powerful. Therefore it is of interest to estimate how much power is lost if the parent distribution is normal by using the modified test in order to obtain the property of robustness.
In order to answer these two questions a rather extensive empirical sampling experiment has been performed.

5.2 Sampling Experiment to Study Two Procedures for Comparing Two Variances

In order to study the power function and robustness of the standard F test and the modified F test, the power functions and the null distributions of these statistics were investigated for the rectangular, normal and truncated double-exponential parent distributions.

The empirical sampling procedure involved drawing 2000 samples of size 20 from each of the three populations. These were paired to give 1000 values of the standard F and the modified F for each population. The appropriate probability associated with each of these F's was estimated from a set of graphs prepared from Pearson's "Tables of The Incomplete Beta Function". These probabilities for the two statistics and three distributions are summarized on Figures 4, 5 and 6.

In addition to the calculations above, which evaluate the behavior of these statistics when the null hypothesis is true, the distributions of these statistics were estimated for three alternatives to the null case:

\[ H_1: 2\sigma_1^2 = \sigma_2^2 \]

\[ H_2: 4\sigma_1^2 = \sigma_2^2 \]

\[ H_3: 6\sigma_1^2 = \sigma_2^2 \]
In order to decrease the amount of IBM calculation, a dodge was used in drawing the original 40,000 samples. For each of the three populations, one hundred values of the ordinate of the density functions were calculated at the 0.5, 1.5, 2.5, ..., 99.5 percentile points. The squares, cubes and fourth powers of each of these ordinates were calculated on a desk calculator, and all four of these quantities were punched on one IBM card for each of the one hundred 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 one 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 2,000 samples of 20 were drawn.

Before analyzing the results of this experiment it is instructive to note the imperfection in the three populations due to this truncation caused by selecting only one hundred values to approximate continuous distributions. Comparisons of the theoretical and actual fourth moments, $\beta_2$, are shown below.
Table 11
Parent Populations for Sampling Experiment

Values of $\beta_2$

<table>
<thead>
<tr>
<th></th>
<th>Rectangular</th>
<th>Normal</th>
<th>Double Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical</td>
<td>1.8000</td>
<td>3.0000</td>
<td>6.0000</td>
</tr>
<tr>
<td>Actual</td>
<td>1.7998</td>
<td>2.8340</td>
<td>4.7301</td>
</tr>
</tbody>
</table>

Actually deviations of these populations from their theoretical counterparts is of no serious consequence, as the purpose of the experiment was to obtain populations representing three degrees of kurtosis centered about the normal.

The results indicating the robust property of the modified $F$ test are shown on Figures 4, 5, and 6. These are frequency distributions for 1000 pairs of samples, each showing the frequencies of the probabilities calculated from the standard and modified $F$ statistics when sampling from each of the three parent distributions. For a sufficiently large number of samples these distributions should be rectangular for a completely robust test. This is equivalent to saying that five per cent of the values of the statistic should occur in each five percentile range of their distributions.

The charts establish empirically the failure of the normal-theory $F$ test for variances to be robust with respect to deviations of kurtosis
Figure 4
Rectangular Parent Population

Standard F Test

Robust F Test

Nominal Percentile Groups
Figure 5
Normal Parent Population

Standard F Test

Robust F Test

Frequency

Nominal Percentile Groups
Figure 6

Double Exponential Parent Population

Standard F Test

Robust F Test

Nominal Percentile Groups
from normal. They also establish the relative robustness of the modified F test for these same leptokurtic and platykurtic populations.

For the rectangular parent population, the standard F test shows 0.7 per cent of the values below the 5 per cent point and 0.6 per cent of the values above the 95 per cent point. The modified test corrects almost perfectly for this lack of robustness in the standard test, showing 4.5 per cent of the values below the 5 per cent point and 4.7 per cent of the values above the 95 per cent point.

For a robust test, fifty samples would be expected in each cell of the chart. The use of chi square to compare the actual frequencies with this ideal behavior illustrates the improvement with the modified test. With nineteen degrees of freedom, the value of chi square is reduced from 254.6 to 21.84 by the modification of the F test; this clearly indicates that the standard F test is not behaving properly, while the observed frequencies of the modified test are compatible with the hypothesis of a robust test. The results of sampling from all three parent populations may be summarized by the following table.
Table 12
Sampling Results Summary for Comparison of Standard(S) F
and Modified (M) F Tests

<table>
<thead>
<tr>
<th>Parent Population</th>
<th>Rectangular</th>
<th>Normal</th>
<th>Double Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Modified</td>
<td>S</td>
</tr>
<tr>
<td>Per Cent Below 5%</td>
<td>0.7</td>
<td>4.5</td>
<td>3.0</td>
</tr>
<tr>
<td>Point</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per Cent Above 95%</td>
<td>0.6</td>
<td>4.7</td>
<td>3.0</td>
</tr>
<tr>
<td>Point</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi Square</td>
<td>254.16</td>
<td>21.8</td>
<td>36.12</td>
</tr>
<tr>
<td>(19 d.f.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probability of</td>
<td>less</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exceeding Observed Chi Square*</td>
<td>than</td>
<td>.001</td>
<td>.290</td>
</tr>
</tbody>
</table>

* Assuming H₀: The test is robust with respect to the parent-population sampled.

Having established the relative robustness of the modified test, the question arises as to how much power is lost by the modification of the standard test, when in fact the parent population is normal. Chart 7 shows the power curve for the standard and modified F test for α = .05. The power curve for the standard F test agrees closely with the theoretical power function of the test.

From the graph it appears that there is very little loss of power due to modifying the F test. For example, the probability of detecting
a variance ratio of 3:1 is reduced from about .75 to about .71 by the use of the modified test. Since this difference in power is subject to a sampling error of .02, the ninety-five per cent confidence limits on the loss of power is (.00 = .08).

It is interesting to note that the power curve for the standard F test with 18 and 18 degrees of freedom, $\alpha = .05$, coincides almost perfectly with the modified F test curve for 20 and 20 degrees of freedom. This indicates a loss of power of about ten per cent in the sense of requiring ten per cent more observations with the modified F than with the standard test to obtain the same power curve.

Since the standard F test does not have the proper 5 per cent intercept for the non-normal populations, the comparison of power curves for the two tests in these cases is of relatively less interest. However, these are shown in Figures 8 and 9. What can be noted is that the modified test is most powerful for the rectangular and least powerful for the double exponential. That is to say the power decreases with increasing $\beta_2$.

5.3 Test to Compare k Variances, Means Known

The next simplest case of a test for the equality of variances is that of comparing k groups, where the mean for each group is known. Again the known means may be assumed equal to zero.

The statistic most commonly used to test the equality of k variances is the Bartlett statistic; note that here the formula is slightly modified due to the zero means of the groups.
\[ M = N \ln \left( m_{2t} \right) - \sum_{t=1}^{k} n_t \ln \left( m_{2t} \right) \]  

where:

\[ m_{2t} = \frac{1}{n_t} \sum_{i=1}^{n_t} x_{ti}^2 \text{ = sample variance of group } t \]

\[ m_{2} = \frac{1}{k} \sum_{t=1}^{k} m_t \text{ = average sample variance} \]

\[ n_t \text{ = number observations in group } t \]

\[ k \text{ = number of groups} \]

\[ X_{ti} \text{ = observation } i \text{ in group } t \]

\[ N = \sum_{t=1}^{k} n_t \]

The critical region used for the statistic is

\[ W_{x} : M > (1 + A) x^2 (k - 1) \]  

where \( A \) is a modifying quantity of order \( N^{-1} \) proposed by Bartlett (1934). Since, however, the test depends on \( \beta_2 \) to order \( N^0 \); there is little point in considering this refinement, and we shall consider the critical region defined by

\[ M > x^2 (k - 1) \]  

(76)

Since it has been shown, Box (1953), that this test is extremely sensitive to non-normality, an approximate permutation test has been derived by fitting the first two moments, as outlined in section 2.8.

The first step in the derivation is to evaluate to order \( N^0 \), the permutation moments of \( M \), denoted by \( E (M) \) and \( V (M) \),

\[ E (M) = \left( k - 1 \right) \left( \frac{b_2 - 1}{2} \right) \]  

(77)
\begin{equation}
V_M = 2(k - 1) \left( \frac{b_2 - 1}{2} \right)^2
\end{equation}

where:

\[ b_2 = \frac{\sum_{s=1}^{n} x_s}{\left[ \sum_{s=1}^{n} x_s^2 \right]^{\frac{1}{2}}} \]

Since the mean and variance of chi square with \( k - 1 \) degrees of freedom are

\[ E(\chi^2) = k - 1 \]
\[ V(\chi^2) = 2(k - 1) \]

and since \( b_2 \) is a constant over the permutation distribution, the statistic

\[ M' = \frac{M}{b_2 - 1} = \frac{2M}{b_2 - 1} \]

immediately suggests itself since its mean and variance follow directly from (77) and (78) as

\[ E(M') = k - 1 \]
\[ V(M') = 2(k - 1) \]

This then gives the approximate permutation test, where the critical region, \( w_\alpha \), is defined by

\[ w_\alpha : \frac{2M}{b_2 - 1} > \chi_\alpha^2 (k - 1) \]

No empirical sampling was carried out for this test.
Chapter VI

TESTS ON VARIANCES, MEANS ASSUMED UNKNOWN

6.1 Theoretical Justification

In Chapter 5 it was possible to obtain the first two permutation moments of the F statistic to compare two variances and the Bartlett statistic, \(M\), to compare \(k\) variances. In these cases where the mean was known, this was possible essentially because of the fact that both statistics could be put in a form where the only terms occurring in the denominator of the statistic, eg.

\[
\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} x_{ij}^2
\]

are constant over all permutations of the sample. This eliminates the problem of getting the expectation of a ratio with variable numerator and denominator.

In the case of the means unknown, terms like (82) become

\[
\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} (x_{ij} - \overline{x}_i)^2
\]

which are not constant over all permutations, as the values of \(x_i\) change for each permutation. Therefore the method of the previous chapter, that of obtaining the approximate permutation test, does not here seem feasible.

The approach to the derivation of test to compare variances for the case of means assumed unknown is to select rather arbitrarily the
criterion analogous to the one used in the previous chapter, where the means were assumed known, and justify its use by finding the general, non-normal, moments of the statistic so chosen. The verification that this statistic has the correct mean and variance to order \( N^0 \) will be considered justification for its use.

Considering the modified Bartlett statistic, we may write this in the form

\[ M' = \frac{M}{1 + \frac{1}{2} c_2} \quad (85) \]

where: \( c_2 = k_4/k_2^2 \)

\( k_4, k_2 = \text{Fisher's k statistics} \)

It can then be shown that to order \( N^0 \) the mean and variance are

\[ E (M') = k-1 \quad (86) \]

\[ V (M') = 2(k-1) \]

for any parent population. This justifies the test criterion

\[ w : M' > \chi^2_{\alpha} (k-1) \quad (87) \]

That is, we have shown that the mean and variance of the test criterion, \( M' \), are the same to order \( N^0 \) as the mean and variance of the chi square distribution to which it is referred. This was not true of the Bartlett criterion, \( M \).
6.2 Empirical Sampling Experiment

Since this test has the correct mean and variance only to order \( N^0 \) it is of interest to find out how much more robust this modified criterion is than the standard Bartlett test. In order to study the robustness and to evaluate the power, an empirical sampling experiment has been performed, drawing 2000 samples of 20 observations from each of the three populations described in chapter 5. These were subdivided into 200 sets of 10 samples each and for each set of 10 samples both the standard \( M \) and the modified statistic, \( M' \), were computed to obtain three sampling distributions of each statistic.

In addition to the investigation of the properties of these test criteria in the null case, the same number of samples was drawn from populations with different variances to estimate the power with respect to the alternative hypothesis,

\[
H_1: \quad \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1 \\
\sigma_5^2 = \sigma_6^2 = \sigma_7^2 = 1.7 \\
\sigma_8^2 = \sigma_9^2 = \sigma_{10}^2 = 2.6
\]  

(88)

The results of the sampling from the null parent populations are shown on Figures 10, 11 and 12. These verify the extreme sensitivity of the normal-theory test to non normality. While the frequencies are not exactly as they should be in the distribution of the probabilities for the modified statistic, \( M' \), a great improvement is achieved relative to the original \( M \) statistic. The most serious aberration from "ideal"
Bartlett Statistic, $M_{100}$
($k=10$, $n=20$)

Figure 10
Empirical Sampling Distributions
Approximate Double-Exp. Parent Population

Modified Bartlett Statistic, $M'$
($k=10$, $n=20$)

Frequency

Nominal Percentiles

Frequency

Nominal Percentiles
Figure 11

Empirical Sampling Distributions
from
Normal Parent Population

Bartlett Statistic, $M$
($k=10$, $n=20$)

Frequency

Modified Bartlett Statistic, $M'$
($k=10$, $n=20$)

Frequency

Nominal Percentiles

Nominal Percentiles
Figure 12
Empirical Sampling Distributions from Rectangular Parent Population
behavior is the occurrence of 37 cases rather than the expected 20 in the upper decile when sampling from a rectangular parent population. However, some solace may be taken in the fact that this population represents a more severe deviation from normality than is commonly encountered in experimental situations.

Since an "ideal" distribution of probabilities would give 20 samples in each decile grouping, we may make a chi square test of the hypothesis that these empirical distributions come from a rectangular distribution. The results of these computations show:

Table 13

Empirical Comparison of The Robustness of The Two Criteria, M and $M'$

<table>
<thead>
<tr>
<th>Parent Distribution</th>
<th>Test Criterion</th>
<th>$\chi^2(9)$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectangular</td>
<td>M</td>
<td>550.8</td>
<td>.000 0001</td>
</tr>
<tr>
<td></td>
<td>$M'$</td>
<td>36.1</td>
<td>.000 04</td>
</tr>
<tr>
<td>Normal</td>
<td>M</td>
<td>21.2</td>
<td>.011 70</td>
</tr>
<tr>
<td></td>
<td>$M'$</td>
<td>22.9</td>
<td>.007 38</td>
</tr>
<tr>
<td>Double Exponential</td>
<td>M</td>
<td>475.8</td>
<td>.000 0001</td>
</tr>
<tr>
<td></td>
<td>$M'$</td>
<td>6.7</td>
<td>.669</td>
</tr>
</tbody>
</table>

While inspection of Figure 12 for the rectangular parent population indicates that the modified statistic is much less sensitive to
non normality than Bartlett's criterion, the significance level of .000 04 for the chi square test indicates that we have enough evidence here to detect the fact that $M'$ is not a perfectly robust criterion. Figure 11 for the normal population gives a significance level of approximately .01 for both $M$ and $M'$.

This fact plus an inspection of the individual pairs of values of $M$ and $M'$ indicate that this is not so much evidence of lack of robustness, but rather evidence of an "unlikely" set of 200 statistics. That is, the similar failure of both graphs in Figure 11 to be rectangular suggests that a repetition of this portion of the sampling experiment would give two similar, but more nearly rectangular distributions. Figure 10 shows that $M'$ has modified a very sensitive test so that the new test is extremely robust to a rather leptokurtic parent population.

While it is desirable to have a test which is more robust with respect to non normality, it is also important that the gain in robustness is not accompanied by a great loss in power when the parent population is normal. Therefore an equal number of samples have been drawn from the non-null parent population (88). For the error of the first kind fixed at $\alpha = .05$, the probabilities of detecting heterogeneity of variance of the magnitude of $H_1$ for both the standard and modified statistic when sampling from the normal parent population are
Table 14

Power of The Test Criteria, M and M'

<table>
<thead>
<tr>
<th>Test</th>
<th>Power w.r.t. $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartlett</td>
<td>.815</td>
</tr>
<tr>
<td>Modified Bartlett</td>
<td>.810</td>
</tr>
</tbody>
</table>

where the values in the table are subject to a sampling standard error of about .028.

6.3 Summary of Robust Tests for Variances

While the modification of the Bartlett test for comparing $k$ variances has the correct moments only to order $N^0$, it appears that it gives a considerably more robust test with very little loss of power.
Chapter VII

SUMMARY AND CONCLUSIONS

The primary result obtained in this dissertation is the derivation of the modified Bartlett test for comparing k variances. This test criterion has the critical region

\[ W: M' = M \left( 1 + \frac{1}{c_2} \right)^2 \chi^2_{k-1} \]

where:

\[ M = N \ln(k_2) - \sum n_t \ln(k_{2t}) \]

\[ c_2 = k_4 / k_2^2 \]

\[ k_2, k_4 = \text{Fisher's k statistics} \]

The robustness and power of this approximate permutation test has been compared with the regular Bartlett criterion, \( M \), by an empirical sampling experiment. This investigation has shown that, with no appreciable loss of power for normal parent populations, the sensitivity of the type I error to non normality has been greatly reduced, as the following table indicates.

<table>
<thead>
<tr>
<th>Parent Population</th>
<th>( M )</th>
<th>( M' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>.090</td>
<td>.130</td>
</tr>
<tr>
<td>Rectangular</td>
<td>.005</td>
<td>.185</td>
</tr>
<tr>
<td>Approximate Double Exponential</td>
<td>.555</td>
<td>.095</td>
</tr>
</tbody>
</table>

* based on 200 samples each
Having developed the idea of using approximate permutation tests as robust testing criteria, the approximate permutation tests have been obtained for several situations where it is desired to compare means. The essential feature to be noted is that these criteria are the same as the standard tests to order $N^0$; this explains the robustness of the standard tests, and also serves as a justification for the standard tests which is independent of the assumptions made in their standard derivations. This feature is in contrast with tests to compare variances for which the normal-theory and the approximate permutation criteria do not agree to order $N^0$.

The fact that the general (non-normal) moments may be obtained readily from the permutation moments of these statistics enables us to derive the effect of assumption failures in the type I error directly from the form of the robust tests. Therefore, as a by-product of having developed approximate permutation tests, we have a rather simple method of evaluating the effect of non-normality and variance heterogeneity on several tests for comparing means. These results show close agreement with previously published tables which have been obtained by more complex methods.

In addition to these newly prepared tables, and their previously published counterparts, excerpts have been taken from several sources to give a more complete picture of the effects of assumption failures on a large number of standard tests.
Further Research Required

The robustness of all existing parametric tests should be investigated. Those showing only slight sensitivity should be designated as robust, useful procedures. Those showing great sensitivity to the failure of the underlying assumptions should be so classified in order that research may be initiated to find more robust procedures for these situations. The development of new tests in this latter situation, then, represents the second area for future research, which, of course, must await completion of the first phase.

In particular the large number of multivariate tests should be studied from this point of view as there are certain to be several tests which fall into each classification. Also the robustness of tests in variance component analysis, regression analysis, and serial correlation might be investigated.

In view of the failure of the modified Bartlett statistic, $M'$, to show as much robustness as desired, in this rather small empirical investigation, a more extensive sampling investigation on an electronic computer seems highly desirable. If this confirmed the failure of $M'$ to behave as well as might be desired, it might be possible to develop a better statistic than $M'$. 
REFERENCES


Pitman, E. J. G. 1937. Significance tests which may be applied to samples from any population. Biometrika 29: 322-335.

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


Appendix A

NOTES ON GENERAL TECHNIQUE OF FITTING A BETA DISTRIBUTION

BY THE FIRST TWO MOMENTS

Some relationships among the parameters and moments of the Beta distribution should be noted. The mean and variance of a Beta variate, with parameters, \((r, s)\), are known to be

\[
\mu = \frac{r}{r + s} \tag{1}
\]

\[
\nu = \frac{rs}{(r + s)^2 (r + s + 1)} \tag{2}
\]

From these it follows that

\[
r = \mu \nu (\mu - \mu^2 - \nu) \tag{3}
\]

Let \(r, s\) = normal-theory Beta parameters

\(r^*, s^*\) = parameters of Beta distribution fitted to general, non-normal population

We have shown earlier that these may be written as

\[
r^* = \Delta r \tag{4}
\]

\[
s^* = \Delta s
\]

This permits us to write

\[
\Delta = \frac{r^*}{r} = \frac{\mu^*}{\nu} \cdot \frac{\nu}{\mu} \cdot \frac{(\mu^* - \mu^{*2} - \nu^*)}{(\mu - \mu^2 - \nu)} \tag{5}
\]

Let \(\varepsilon = \frac{\nu^*}{\nu} \tag{6}\)
and assume

$$\mu = \mu^*$$

Then (5) may be written in the simpler form

$$\Delta = \frac{1}{\varepsilon} \left( \frac{\mu - \mu^2 - V\varepsilon}{\mu - \mu^2 - V} \right)$$

(7)

which may also be written in the form

$$\Delta = 1 + \frac{\mu}{\varepsilon} \cdot \frac{(1 - \mu)(1 - \varepsilon)}{\mu - \mu^2 - V}$$

(8)

This relationship may be used to obtain the factor, $d$, for robust tests, where the Beta distribution is fitted to the permutation moments. Then $V^*$ and $\mu^*$ are the mean and variance with respect to the permutation distribution. The same relationships may be used to obtain the factor, $\delta$, for evaluating the effect of assumption failures. The $V^*$ and $\mu^*$ are the mean and variance with respect to the general (e.g., non-normal) parent population. For all the cases considered in both types of problems, it was found that $\mu = \mu^*$; therefore, this assumption has been made above to simplify the expression of $\Delta$. 
Appendix B

CALCULATION OF THE FACTOR, d, FOR ROBUST F TESTS

This type of calculation will be illustrated by the one-way classification of analysis of variance for comparing k means.

For reasons noted earlier, we shall consider the statistic,

\[ W = \frac{\text{treatment sum of squares}}{\text{treatment + error sum of squares}} \] (1)

From Welch (1938) we have the moments of \( W \).

**Permutation Moments:**

\[ \mu^* = M_1 = \frac{k - 1}{N - 1} \] (2)

\[ V^* = M_2 = \frac{2(k - 1)(N - k)}{(N + 1)(N - 1)^2} \left(1 - \frac{k^2}{Nk_2}\right) \]
\[ - \frac{k_4}{(N - 1)^2 k_2^2} \left(\frac{k^2}{N} - \sum \frac{1}{n_t}\right) \] (3)

**Normal-Theory Moments:**

\[ \mu = \mu(N) = \frac{k - 1}{N - 1} \] (4)

\[ V = \mu_2(N) = \frac{2(k - 1)(N - k)}{(N + 1)(N - 1)^2} \] (5)

where: \( n_t = \text{number of observations per group} \)

\[ N = \sum_{t=1}^{k} n_t = \text{total number of observations} \]
With the moments given in (3) and (5), we write equation A(6) for this case

\[ \varepsilon = \frac{V^*}{V} = (1 - \frac{c_2}{N}) - \frac{c_2 \left( \frac{k^2}{N} - \sum \frac{1}{n_t} \right)}{2(k - 1)(N - k)} \frac{1}{N + 1} \]  

(6)

where: \[ c_2 = \frac{k_4}{k_2^2} \]

By the use of equation A(8) we may then write

\[ d = 1 + c_2 \frac{\frac{N + 1}{N(N - 1)} + \frac{N + 1}{N - 1} U}{1 - c_2 \left( \frac{1}{N} - U \right)} \]

(7)

where:

\[ U = \frac{N + 1}{2(k - 1)(N - k)} \left( \frac{k^2}{N} - \sum \frac{1}{n_t} \right) \]

For large \( N \), this may be approximated by

\[ d = 1 + \frac{1 + NU}{N - c_2 + NUc_2} \]  

(8)

For equal sized groups, where \( U = 0 \), we have

\[ d = 1 + \frac{(N + 1)c_2}{(N - 1)(N - c_2)} \]  

(9)
Appendix C

CALCULATION OF FACTOR, $c$, FOR ESTIMATING THE
EFFECT OF ASSUMPTION FAILURES

To estimate the effect of assumption failures on the type I error,
the Beta distribution is fitted by means of the first two moments to
the general (non-normal) moments of $W$. Making use of the fact (esta-
blished in section 2.6) that the general moments about the origin can
be obtained by taking the expectation of the permutation moments, we
may proceed with the following argument:

\[ \mu_1' = E(M_1') = M_1' \] \hspace{1cm} (1)

because in all cases considered $M_1'$ is a function only of the number of
observations in each classification, and not of the values of those
observations.

\[ \mu_2' = E(M_2') \] \hspace{1cm} (2)

where:
- $M_1'$ = $r^\text{th}$ permutation moment about the origin
- $\mu_r'$ = $r^\text{th}$ general moment about the origin

Now from these relationships we may write

\[ \mu_2 = \mu_2' - (\mu_1')^2 = E(M_2') - \left( E(M_1') \right)^2 \]
\[ = E(M_2' - M_1'^2) \]
\[ = E(M_2) \] \hspace{1cm} (3)
That is, we have shown that the general variance, for the cases considered in this paper which satisfy (1), equals the expectation of the permutation variance. With the aid of this result, we may write equation A(6) in its special form for fitting the Beta distribution to the general moments as

$$\delta = 1 + \frac{\mu}{\varepsilon} \cdot \frac{(1 - \mu)(1 - \varepsilon)}{\mu - \mu^2 - \mu_2}$$

(4)

where: $\delta = \text{factor for modifying degrees of freedom}$

$$\varepsilon = \frac{\mu_2}{M_2}$$

Then having chosen the critical value of $F$ for the normal-theory test such that:

$$P(F > F_\alpha) = \alpha$$

(5)

we wish to calculate the probability that $F'$, the $F$ statistic obtained from sampling non-normal parent populations where the null hypothesis is true, exceeds the critical value,

$$P(F' > F_\alpha) = 1 - \int_0^{F_\alpha} dF(\delta_1, \delta_2) = \alpha'$$

(6)

where: $\alpha' = \text{the true probability of rejecting } H_0, \text{ when it is true in non-standard parent population}$

The numerical calculations are made in the standard fashion with the aid of Tables of The Incomplete Beta Function.
Having established the general method, consider a particular case, the one-way classification analysis of variance, with an equal number of observations in each group. From equation (3), we may write

\[ v^* = \mu_2(G) = E(M_2) = \frac{2(k-1)(N-k)}{(N+1)(N-1)^2} \left( 1 - \frac{1}{N} E(c_2) \right)^2 \]  

(7)

Therefore the only change in the modifying factor, \( d \), for the robust test, and the factor, \( \delta \), for evaluating non-normality is replacement of \( c_2 \) by \( E(c_2) \). Therefore

\[ \delta = 1 + \frac{(N+1) E(c_2)}{(N-1)(N - E(c_2))} \]  

(8)

To find \( E(c_2) \) in terms of the population parameters, write \( c_2 \) in the form,

\[ c_2 = \frac{k_1}{k_2^2} = \frac{k_1 + (k_4 - k_4)}{k_2^2 \left( 1 + \frac{k_2 - k_3}{k_2} \right)^2} \]  

(9)

Letting \( \Delta_r = k_r - k_r \)

and using the series expansion

\[ \frac{1}{(1 + z)^2} = 1 - 2z + 3z^2 - 4z^3 + \ldots \]

we may rewrite (9) in the form

\[ c_2 = k_2^{-2} \left( k_4 + \Delta_4 - \frac{\Delta_4}{k_2} \right) \left( 1 - 2 \frac{\Delta_4}{k_2} + (\frac{\Delta_4}{k_2})^2 \right) \]  

(10)
Using the method of evaluating moments and mixed moments of k-statistics by F. N. David (1949), the expectation of $c_2$ to order $N^{-1}$ is

$$\mathbb{E}(c_2) = \lambda_2 + \frac{1}{N} (3\lambda_2^2 - 10\lambda_2 - 12\lambda_1^2 - 2\lambda_4)$$

(11)

where:

$$\lambda_r = \frac{K_r + 2}{r + 2}$$

$$K_r = r^{th} \text{ cumulant}$$

Substituting (11) into (8) and simplifying, we have

$$\delta = 1 + \frac{(N + 1) \mathbb{E}(c_2)}{(N - 1) \left( N - \mathbb{E}(c_2) \right)}$$

(12)

$$\delta = \frac{2\mathbb{E}(c_2) + N(N - 1)}{(N - 1) \left( N - \mathbb{E}(c_2) \right)}$$

(13)

Since $\mathbb{E}(c_2)$ has been evaluated only to order $N^{-1}$ and since $N^2$ is large relative to $\mathbb{E}(c_2)$, this may, for practical purposes, be approximated by the simpler expression,

$$\delta = \frac{1}{\mathbb{E}(c_2)} \quad \frac{1}{N}$$

(14)
Appendix D

METHOD OF CONSTRUCTION OF CRITICAL REGION FOR THE
APPROXIMATE PERMUTATION TEST, SECTION 2.10

The procedure for obtaining the cross section of the critical region, \( W \), on the plane,

\[
X_1 + X_2 + X_3 = 10
\]

of the approximate permutation test is outlined. Since the critical region, \( W \), is defined by

\[
W : F > F_\alpha(d, 2d)
\]

those points on the boundary of \( W \) will be defined by

Boundary of \( W \): \( F = F_\alpha(d, 2d) \) \( (1) \)

Since

\[
d = \frac{(n + 2)(b_2 - 3)}{n(n + 2 - b_2)} = \frac{5(b_2 - 3)}{3(5 - b_2)} \quad (3)
\]

and

\[
b_2 = (n + 2) \frac{\sum x^4}{(\sum x^2)^2} = 5 \frac{\sum x^4}{(\sum x^2)^2} \quad (4)
\]

it is possible to draw contour lines of constant \( d \).

Also, \( F \) is defined as a function of the sample by

\[
F = \frac{\frac{n}{2} \sum \frac{x^2}{2}}{\sum (x - \bar{x})^2} \quad (5)
\]
Figure 13
Graphical Construction of Approximate Permutation Test Critical Region

Concentric Circles = Contours of Equal $F_{10}(b_2)$
Other Contours of Equal $b_2$, i.e. equal $d$. 
permitting us to draw contours of constant $F$. If in particular we
draw the contours of constant $F$. If in particular we draw the contour
line of $F = F_{10}$ $(d, 2d)$, and note its intersection with the contour
line of constant $d$, we obtain points on the boundary of $W$.

In Figure 13 four values of $d$ have been selected and contour
lines of fixed $d_1$ and fixed $F_{10}$ $(d_1, 2d_1)$ have been drawn. However,
because of near-tangential intersection of the two sets of contour
lines, a graphical solution is not satisfactory to define the boundary
line. Therefore the following procedure is used.

A. Sketch contours of equal $b_2$

B. Sketch contours of equal $F(b_2)$

Note approximate intersection points graphically

B. 1. From graph, estimate a point of intersection, $X$, eg. on $b_2 = 3$

2. Calculate $F_{10}(d)$

3. Read one coordinate of $X$, say $x_1$, as closely as possible

4. Solve for $x_2, x_3$ on the plane which give $F_{10}(d)$ by solving

$$x_2, x_3 = \frac{P \pm (2Q - F^2)^{1/2}}{2}$$

where $P = x_2 + x_3 = 10 - x_1$

$$Q = x_2^2 + x_3^2 = \frac{100}{3} + \frac{200}{3F_{10}(d)} - x_1^2$$

Proof:

$$F = \frac{\frac{1}{3} (\sum x)^2}{\frac{1}{2} \left( \sum x^2 - \frac{1}{3} (\sum x)^2 \right)}$$
\[ \sum x^2 = \frac{100}{3} + \frac{200}{3F} \]
\[ x_2^2 + x_3^2 = \frac{100}{3} + \frac{200}{3F} - x_1^2 = Q \]
\[(ii) \quad x_2^2 + \left[ F (x_2 + x_3) - x_2^2 - F x_2 + x_3^2 \right] = 0 \]
\[\text{i.e.} \quad x_2^2 + (F - x_2)^2 - Q = 0 \]
\[ x_2^2 - F x_2 + \frac{1}{2} (F^2 - Q) = 0 \]
\[ x_2 = F \pm \frac{1}{2} (2Q - F^2)^{1/2} \]

5. Find \( x_3 = 10 - x_1 - x_2 \)

6. Solve for
\[ b_2 = (n + 2) \frac{\sum x^4}{(\sum x^2)^2} = 5 \frac{\sum x^4}{(\sum x^2)^2} \]

7. Compare this with assumed value of \( b_2 \). Repeat this procedure to obtain a trial and error solution - the process must be continued until the assumed value of \( b_2 \) has been bracketed, so that graphical interpolation may be used.

A sample calculation of the trial and error procedure indicated above is given.

1. Assume \( d = 6 \)
   \[\text{i.e.} \quad b_2 = 4.5 \]
2. Look up in tables \( F_{10}(6, 12) = 2.331 \)
3. Choose \( x_1 = 7.7 \)
4. Calculate

\[ P = 10 - 7.7 = 2.3 \]

\[ \sum x^2 = \frac{100}{3} + \frac{200}{3(2.331)} = 61.933 \]

\[ \sum x^2 - x_1^2 = Q = 61.933 - (7.7)^2 = 2.643 \]

\[ x_2 = P \pm \frac{1}{2}(2Q - P^2)^{1/2} = \]

\[ = 2.3 \pm \frac{1}{2}(2 \cdot 2.643 - (2.3)^2)^{1/2} \]

\[ = 1.15 \]

5. Solve for

\[ x_3 = 10 - 7.7 - 1.15 = 1.15 \]

6. From these we calculate the sample value,

\[ b_2 = 4.587 \]

7. Compare this to the assumed value \( b_2 = 4.500 \) and choose a new estimate \( x_1 \) to decrease the calculated value 4.587 nearer the assumed value, 4.500.

The results of these calculations for four contours of equal d values are:

<table>
<thead>
<tr>
<th>( b_2 )</th>
<th>( d )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( F\cdot10(b_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.143</td>
<td>0.50</td>
<td>3.00</td>
<td>4.55</td>
<td>2.45</td>
<td>28.123</td>
</tr>
<tr>
<td>3.000</td>
<td>1.00</td>
<td>5.56</td>
<td>2.67</td>
<td>1.77</td>
<td>8.526</td>
</tr>
<tr>
<td>4.500</td>
<td>6.00</td>
<td>1.67</td>
<td>0.66</td>
<td>7.67</td>
<td>2.331</td>
</tr>
<tr>
<td>4.945</td>
<td>60.00</td>
<td>0.23</td>
<td>0.64</td>
<td>9.13</td>
<td>1.3203</td>
</tr>
</tbody>
</table>
Appendix E
DERIVATION OF ROBUST F TEST FOR COMPARING
TWO VARIANCES, MEANS KNOWN

Consider the statistic,

\[ W = \frac{\sum_{j=1}^{n_1} X_{1j}^2}{\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} X_{ij}^2} \]  

(1)

with the assumption that

\[ X_{ij} \sim \text{IN}(0, \sigma_i^2) \]

and under the null hypothesis

\[ H_0: \sigma_1^2 = \sigma_2^2 \]

it is well known then that \( W \) has a Beta distribution with parameters, \((n_1/2, n_2/2)\), enabling us to write the moments,

\[ E(W) = \mu(N) = \frac{n_1}{N} \]  

(2)

\[ V(W) = \mu_2(N) = \frac{2n_1n_2}{(n_1 + n_2)^2(n_1 + n_2 + 2)} \]  

(3)

where: \( N = n_1 + n_2 \)

Noting that the denominator of \( W \) is constant over all permutations of

the observations, we find the permutation moments,

\[ \bar{W}_1 = \frac{\frac{n_1}{N} S_2}{S_2} = \frac{n_1}{N} \]  

(4)
where: \[ S_r = \sum_{i=1}^{n_1} \sum_{j=1}^{x_{i,j}} \]

The second permutation moment with respect to the origin is

\[ E(v^2) = \frac{1}{2} \frac{S_2}{S_2^2} E \left( \sum_{j=1}^{n_1} x_{i,j}^4 + \sum_{i \neq j} x_{i}^2 x_{j}^2 \right) \]  \( \text{(5)} \)

Noting that

\[ E(x_1^4) = \frac{1}{N} S_4 \]  \( \text{(6)} \)

\[ E(x_1^2 x_j^2) = E(x_1^2) \cdot E(x_j^2 / x_1) \]

\[ = \frac{E(x_1^2) \cdot (S_2 - x_1^2)}{N - 1} \]

\[ = \frac{S_2^2 - S_4}{N(N - 1)} \]  \( \text{(7)} \)

Now using (6) and (7) to evaluate (5), we have

\[ E(v^2) = \frac{1}{S_2^2} \left( \frac{n_1}{N} S_4 + \frac{n_1(n_1 - 1)}{N(N - 1)} (S_2^2 - S_4) \right) \]

\[ = \frac{n_1 n_2}{N^2(N - 1)} b_2 + \frac{n_1(n_1 - 1)}{N(N - 1)} \]  \( \text{(8)} \)

where: \[ b_2 = \frac{N S_4}{S_2^2} \]

From this it follows that

\[ \bar{v}_2 = \frac{E(v^2)}{p} - \bar{v}_1^2 = \frac{2 n_1 n_2}{N^2(N - 1)} \zeta^{-1} + \frac{1}{2} \sigma_2^2 \]  \( \text{(9)} \)
where: $c_2 = b_2 = 3$

Making use of equation A(6) and A(7) to obtain the modifying factor, $d$, we obtain

$$e = \frac{M_2}{\mu_2(N)} = \frac{N + 2}{N - 1} \left(1 + \frac{1}{2} c_2\right)$$  \hspace{1cm} (10)

and

$$d = \frac{1}{6} \left(\frac{\mu - \mu^2 \mu_2(N)}{\mu - \mu_2(N)}\right)$$

$$= \frac{N - 1}{N + 2} \left(1 + \frac{1}{2} c_2\right) \left[\frac{n_1}{N} - \frac{n_1^2}{N^2} - \frac{2n_1 n_2}{N^2 (N - 1)} \left(1 + \frac{1}{2} c_2\right)\right]$$  \hspace{1cm} (11)

$$= \frac{N - 1}{N} \left(1 + \frac{1}{2} c_2\right)^{-1} - \frac{2}{N}$$  \hspace{1cm} (12)
Appendix F

DERIVATION OF MODIFIED BARTLETT STATISTIC, M',
MEANS UNKNOWN

For the case where the means of the k groups are assumed known, there will be no loss of generality to consider these means all zero. Adopt the notation

\[
M = n \ln(m) - \sum_{t=1}^{k} n_t \ln(m_t)
= - \sum_{t=1}^{k} n_t \ln\left(\frac{m_t}{m}\right)
\]  

(1)

where:
\[m_t = \frac{1}{n_t} \sum_{i=1}^{n_t} y_{it}^2\]
\[m = \frac{1}{N} \sum_{t=1}^{k} n_t m_t = \frac{1}{N} s_2\]
\[n_t = \text{number of observations per sample}\]
\[k = \text{number of groups}\]
\[N = \sum_{t=1}^{k} n_t = \text{total number of observations}\]
\[y_{it} = i-th observation in t-th group\]
\[s_2 = \sum_{t=1}^{k} \sum_{i=1}^{n_t} y_{it}^2 = \sum_{a=1}^{N} y_a^2\]

Let \[x_t = \frac{m_t - m}{m}\]

Then we may rewrite (1) in the form
\[ M = - \sum_{t=1}^{k} n_t \ln(1 + x_t) \]
\[ M = - \sum_{t=1}^{k} n_t (x_t - \frac{1}{2}x_t^2 + \frac{1}{3}x_t^3 - \cdots) \]  
(2)

Now we shall proceed to find the expectation of \( M \) with respect to the permutation distribution term by term. Adopt the following notation.

\[ x_t = \frac{m_t - m}{m} = \frac{1}{n_t} \sum_{i=1}^{n_t} \frac{y_{1i}^2}{N} - \frac{1}{N} S_2 \]

\[ = \frac{1}{S_2} \cdot \frac{N}{n_t} \cdot \sum_{i=1}^{n_t} \left( y_{1i}^2 - \frac{1}{N} S_2 \right) \]

\[ = \frac{1}{S_2} \cdot \frac{N}{n_t} \cdot \sum_{i=1}^{n_t} \left( y_{1i}^2 - \overline{y}^2 \right) \]

\[ = \frac{1}{S_2} \cdot \frac{N}{n_t} \cdot \sum_{i=1}^{n_t} w_i \]  
(3)

where:

\[ w_i = y_{1i}^2 - \overline{y}^2 \]

\[ \overline{y}^2 = \frac{1}{N} \sum_{t=1}^{k} \sum_{i=1}^{n_t} y_{1it}^2 \]
Now let us evaluate the first term in the series expansion of $M$. This by inspection is zero. Then writing the second term we have

$$II = \frac{1}{2} \sum_{t=1}^{k} n_t x_t^2$$

Using equation (3), we write

$$x_t^2 = \frac{N^2}{n_t^2} \cdot \frac{1}{S_2} \left( \sum_{i=1}^{n_t} w_i^2 \right)^2$$

$$= \frac{N^2}{n_t^2} \cdot \frac{1}{S_2} \left( \sum_{i=1}^{n_t} w_i^2 \right) + \sum_{i \neq j} w_i w_j \cdot f$$

Now consider some sub calculations which must be made before proceeding.

**Type A Calculation** - Expectations of powers and products of $w$'s.

All expectation operators will be with respect to the permutation distribution.

$$E(w^2) = \frac{1}{N} S_2'$$ \hspace{1cm} (5)

$$E(w_i w_j) = E(w_i) \cdot E(w_j / w_i) = E(w_i) \cdot \frac{S_1' - w_i}{N - 1}$$

$$= \frac{-E(w_i^2)}{N - 1} = -\frac{1}{N(N - 1)} S_2'$$ \hspace{1cm} (6)

where:

$$S_k' = \sum_{a=1}^{N} w_a^k$$ \hspace{1cm} (7)
noting that

\[ S_{1}' = 0 \]

---

Type B Calculation - Express \( S_k' \) in terms of sums such as \( S_r' \), and standardized moments, \( b_r' \).

\[
S_{2}' = \sum_{a=1}^{N} w_a^2 = \sum_{a=1}^{N} (y_a^2 - \frac{1}{N} s_2^2)^2
\]

\[
= \sum_{a=1}^{N} \left( y_a^2 - \frac{2}{N} s_2 y_a + \frac{1}{N^2} s_2^2 \right)
\]

\[
= s_4 - \frac{1}{N} s_2^2
\]

(9)

Dividing through by \( s_2^2 \), equation (9) may be written in the form

\[
\frac{S_{2}'}{s_2^2} = \frac{1}{N} (b_2 - 1)
\]

(10)

where:

\[ b_2 = \frac{N s_4}{s_2^2} \]

---

Returning to the main argument of evaluating the expectation of the second term in the series expansion, we have

\[ E(II) = \frac{1}{2} \sum_{t=1}^{k} n_t \cdot E(x_t^2) \]
\[
E(II) = \frac{1}{2} \sum_{t=1}^{k} n_t \cdot \frac{N^2}{n_t} \cdot \frac{1}{S^2} \cdot E \left[ \sum_{i=1}^{n_t} w_i^2 + \sum_{i \neq j} w_i w_j \right]
\]

\[
E(II) = \frac{1}{2} \sum_{t=1}^{k} n_t \cdot \frac{N^2}{n_t} \cdot \frac{1}{S^2} \cdot \left[ \frac{n_t}{N} S_2' - \frac{n_t (n_t - 1)}{N(N - 1)} S_2' \right]
\]

---

**Type C Calculation - Series expansion of coefficients**

To simplify expressions of the type above, it will be helpful to adopt a uniform way of combining coefficients, for which the following notation will be of assistance.

Let

\[
A_0 = \frac{n_t}{N}
\]

\[
A_1 = \frac{n_t (n_t - 1)}{N(N - 1)}
\]

\[
A_r = \frac{n_t (n_t - 1) \cdots (n_t - r)}{N(N - 1) \cdots (N - r)}
\]  \(\text{(11)}\)

---

 Returning to the main argument, we have

\[
E(II) = \frac{1}{2} \cdot \sum_{t=1}^{k} n_t \cdot \frac{N^2}{n_t} \cdot (A_0 - A_1) \frac{S_2'}{S_2^2}
\]

\[
= \frac{1}{2} \sum_{t=1}^{k} \frac{N^2}{n_t} \left( A_0 - A_1 \right) \cdot \frac{1}{N(b_2 - 1)}
\]

Taking terms of \(A_1\) to order \(N^{-1}\), we get
\[ E(II) = \frac{1}{2}(b_2-1) \cdot N \cdot \sum_{t=1}^{k} \frac{1}{n_t} \cdot \left\lceil \frac{n_t(n_t-1)}{N(N-1)} \right\rceil \]

\[ = \frac{1}{2}(b_2 - 1)(k - 1) \cdot \frac{N}{N - 1} \quad (12) \]

q.e.d. \( E(II) \)

Following this procedure gives to order \( N^{-1} \)

\[ E^\sim = \frac{1}{3} \sum_{t=1}^{k} n_t x_t^3 \cdot \]

\[ = \frac{N^2}{3(N-1)(N-2)} \cdot (k-1) \cdot \left( \sum \frac{1}{n_t} - \frac{1}{N} \right) = \frac{3}{N} (3b_2 - b_4 - 2) \quad (13) \]

and

\[ E^\sim = \frac{1}{4} \sum_{t=1}^{k} n_t x_t^4 \cdot \]

\[ = \frac{N^2}{3(N-1)(N-2)} \cdot (k-1) \cdot \left( \sum \frac{1}{n_t} - \frac{1}{N} \right) = \frac{2}{N} \left( \frac{2}{4} + \frac{2b_2}{2} + \frac{4b_2^2}{4} \right) \quad (14) \]

Collecting terms gives to order \( N^{-1} \)

\[ E(M) = \frac{N}{N-1} \cdot (k-1) \cdot \left[ \frac{b_2-1}{2} + \frac{N}{3(N-2)} \cdot \left( \sum \frac{1}{n_t} - \frac{1}{N} \right) \right. \]

\[ \left. \cdot \left( (3b_2 - b_4 - 2) + \left( \frac{1}{n_t} - \frac{1}{N} \right) = \frac{2}{N} \left( \frac{2}{4} + \frac{2b_2}{2} + \frac{4b_2^2}{4} \right) \right. \right] \quad (15) \]
To find the variance of $M$, first evaluate the expectation of
$M^2$, which as a consequence of (2), may be written directly as

\[ M^2 = \frac{1}{4} \sum_{t=1}^{k} n_t^2 x_t^4 - \frac{1}{3} \sum n_t^2 x_t^5 + \frac{13}{36} \sum n_t^2 x_t^5 \]

(I) \hspace{1cm} (II) \hspace{1cm} (III)

\[ + \frac{1}{4} \sum_t \sum_s n_t n_s x_t x_s^2 - \frac{1}{3} \sum_t \sum_s n_t n_s x_t x_s^3 \]

(IV) \hspace{1cm} (V)

\[ + \frac{1}{4} \sum_t \sum_{t \neq s} n_t n_s x_t x_s^4 + \frac{1}{5} \sum_t \sum_{t \neq s} n_t n_s x_t x_s^5 + \ldots \]

(VI) \hspace{1cm} (VII)

(16)

Following the pattern of finding the mean of $M$, we shall evaluate the
first term in some detail and give the results for the rest. The
first term may be written

\[ I = \frac{1}{4} \sum_{t=1}^{k} n_t^2 x_t^4 \]

(17)

With the aid of (3) we write

\[ I = \frac{1}{4} \sum r_t^2 \cdot \frac{1}{s_2} \cdot \frac{N^h}{n_t} \cdot \left( \sum v_1 \right)^4 \]

(18)

---

Type D Calculation

By the use of tables of symmetric functions by F. N. David and
M. G. Kendall (1949) expressions of the following type may be evaluated
directly. From the table we get

\[
\sum_{i \neq j} w_j^4 = \sum w_j^4 + 6 \sum \sum w_j^2 w_i^2 w_k^2 + \sum \sum \sum w_i^2 w_j^2 w_k^2 w_l^2
\]

Type A Calculation

A series of calculations somewhat more complex than the type A calculations in the previous section will be shown in detail here.

\[
E(w_1^4) = \frac{1}{N} S_{i^1}
\]

\[
E(w_1^2 w_j) = E(w_1^2) \cdot E(w_1/i) = E(w_1^2) \cdot \frac{1}{N - 1} (S_{i^1} - w_1)
\]

\[
= -\frac{1}{N(N - 1)} S_{i^1}
\]

\[
E(w_1^2 w_j^2) = E(w_1^2) \cdot E(w_1^2) = E(w_1^2) \cdot \frac{1}{N - 1} (S_{i^1} - w_1)
\]

\[
= \frac{1}{N(N - 1)} \sum_{i \neq j} w_j^4 - S_{i^1}
\]

\[
E(w_1^2 w_j w_k) = E(w_1^2) \cdot E(w_1/i) \cdot E(w_1/i, j)
\]

\[
= E(w_1^2) \cdot E(w_1/i) \cdot \frac{1}{N - 2} (S_{i^1} - w_1 - w_j)
\]

\[
= E(w_1^2) \cdot \frac{1}{(N - 1)(N - 2)} \cdot \sum_{i \neq j} w_1(S_{i^1} - w_1) - (S_{i^1} - w_1^2)
\]

\[
= \cdot \sum_{i \neq j} (S_{i^1} + S_{j^2})
\]
\[
\frac{1}{N(N-1)(N-2)} \left( \sum_{i=1}^{N} s_i \right)^2 \left( \sum_{i=1}^{N} s_i \right)^{-1}
\]

\[E(w_{1,j}w_{k,j}) = E(w_{1,j}) \cdot E(w_{j}/i) \cdot E(w_{k}/i, j) \cdot E(w_{i,j,k}) \]

\[(P) \quad (Q) \quad (R) \quad (T)\]

\[= \frac{1}{N-3} \left( \sum_{i=1}^{N} w_i \right)^2 \left( \sum_{i=1}^{N} w_i \right)^{-2}
\]

\[= \frac{1}{N-2} \left( \sum_{i=1}^{N-1} w_i (s_i - w_i) \right)^2 \left( \sum_{i=1}^{N-1} (s_i - w_i) \right)^{-1}
\]

\[= \frac{1}{N-1} \left( \sum_{i=1}^{N-2} w_i (s_i - w_i) \right)^2 + 2 \left( \sum_{i=1}^{N-1} w_i (s_i - w_i) \right) - \left( \sum_{i=1}^{N-1} s_i \right)^{-2}
\]

\[= \frac{1}{N-1} \left( \sum_{i=1}^{N-2} w_i (s_i - w_i) \right)^2 + 3 \left( \sum_{i=1}^{N-2} s_i \right) - \left( \sum_{i=1}^{N-2} s_i \right)^{-2}
\]

\[= \frac{1}{N(N-1)(N-2)(N-3)} \left( \sum_{i=1}^{N} s_i \right)^2 \left( \sum_{i=1}^{N} s_i \right)^{-1}
\]

---

**Type C Calculation**

Because of the greater complexity of the evaluation of the variance, and because we are seeking an answer correct only to order \(N^{-1}\), we shall expand slightly on the type C calculation illustrated in the previous section. There we adopted the notation of (11).
\[ A_r = \frac{n_t(n_t - 1) \ldots (n_t - r)}{N(N - 1) \ldots (N - r)} \]

In evaluating the variance we shall rewrite these coefficients as

\[ A_0 = \frac{n_t}{N} \]

\[ A_1 = \frac{n_t}{N} \left( \frac{n_t}{N} - \frac{1}{N} + \frac{n_t}{N^2} - \frac{1}{N^2} + \frac{n_t}{N^3} - \frac{1}{N^3} + \frac{n_t}{N^4} - \frac{1}{N^4} \right) \]

\[ A_2 = \frac{n_t}{N} \left( \frac{n_t^2}{N^2} + \frac{3n_t^2}{N^3} - \frac{3n_t^2}{N^4} + \frac{7n_t^2}{N^5} - \frac{9n_t^2}{N^6} + \frac{2n_t^2}{N^7} \right) \]

\[ A_3 = \frac{n_t}{N} \left( \frac{n_t^3}{N^3} + \frac{6n_t^3}{N^4} - \frac{6n_t^3}{N^5} + \frac{25n_t^3}{N^6} - \frac{36n_t^3}{N^7} + \frac{11n_t^3}{N^8} \right) \]

Returning to the main argument as we left it in (18), and making use of the auxiliary calculations, D, A, and C, we may write

\[ E(I) = \frac{1}{4} \cdot \frac{N^4}{s_i^4} \cdot \sum_{t=1}^{k} \frac{1}{n_t^2} \left[ A_0 S_{i_t}^4 - 4A_1 S_{i_t}^4 + 3A_1 (S_{i_t}^2 - S_{i_t}) \right] \]

\[ + 6A_2 (2S_{i_t}^2 - S_{i_t}) + A_3 (3S_{i_t}^2 - 6S_{i_t}) \]

\[ E(I) = \frac{1}{4} \cdot \frac{N^2}{s_i^4} \cdot \sum_{t=1}^{k} \frac{1}{n_t^2} \left[ (A_0 - 7A_1 + 12A_2 - 6A_3)S_{i_t}^4 \right] \]

\[ + (3A_1 - 6A_2 + 3A_3)S_{i_t}^2 \]

We now wish to express the \( S_{i_t}^r \) in terms of the \( S_i \); therefore we perform another

\[ \sum_{r=1}^{k} \frac{1}{n_t^2} \]

\[ \sum_{r=1}^{k} \frac{1}{n_t^2} \]
Type B Calculation

\[ S'_4 = \sum_{a=1}^{N} N_a^4 v_a^4 = \sum_{a=1}^{N} \left( y_a^2 - \frac{1}{N^2} S_2^2 \right)^4 \]

\[ = \sum_{a=1}^{N} y_a^8 - \frac{4}{N} S_2^4 y_a^6 + \frac{6}{N^2} S_2^2 y_a^4 - \frac{4}{N^3} S_2^2 y_a^2 + \frac{1}{N^4} S_2^4 \]

\[ \therefore \frac{S'_4}{S_2^4} = \frac{1}{N^3} \left( b_6 - 4b_4 + 6b_2 - 3 \right) \]

where:

\[ b_6 = \frac{N^3 S_8^8}{S_2^4}, \quad b_4 = \frac{N^2 S_6^6}{S_2^4}, \quad b_2 = \frac{N S_4^4}{S_2^4} \]

In a similar manner or by using (10), we have

\[ \frac{S'_2}{S_2^2} = \frac{1}{N^2} (b_2 - 1)^2 \]

Again returning to the main argument as we left it in (19), we have

\[ E(I) = \frac{1}{4} \sum_{t=1}^{k} \frac{1}{n_t} \left[ N(b_6 - 4b_4 + 6b_2 - 3)(A_0 - 7A_1 + 12A_2 - 6A_3) \right. \]

\[ \left. + N^2(b_2 - 1)^2 (3A_1 - 6A_2 + 3A_3) \right] \]

\[ = \frac{1}{4}(b_2 - 1)^2 \left( 3k - 6 + \frac{3}{N^2} \sum n_t^2 \right) \]

\[ + \frac{1}{4N} \left[ (b_6 - 4b_4 + 6b_2 - 3)(N \sum \frac{1}{n_t} - 7k + 12 - \frac{6}{N^2} \sum n_t^2) \right] \]
\[+ (b_2 - 1)^2 (\frac{1}{N} \sum \frac{1}{n_t} + 21k - 36 + \frac{18}{N^2} \sum n_t^2) \]

q. e. d. E(I)

In a similar manner we may calculate the remainder of the terms in (2). The results of these calculations are

\[E(IV) = \frac{1}{4} (b_2 - 1)^2 \left( k(k - 1) - 2(k - 1) + 3 \left( 1 - \frac{\sum n_t^2}{N} \right) \right) \]

\[
\begin{array}{cccc}
-1 & +4 & -6 & b_6 \\
+4 & -16 & +24 & b_4 \\
+3 & -12 & +18 & b_2 \\
-12 & +48 & -72 & b_2 \\
+6 & -24 & +36 & (1)
\end{array}
\]

\[E(II) = \frac{1}{3} \left( b_4 b_2 - 3b_2^2 + 5b_2 - b_4 - 2 \right) \]

\[\cdot \left( \frac{1}{N} \sum \frac{1}{n_t} - \frac{k}{N} + 50 \frac{1}{N} - 20 \frac{\sum n_t^2}{N^3} \right) \]

\[E(V) = -\frac{1}{3N} \left( b_4 b_2 - 3b_2^2 + 5b_2 - b_4 - 2 \right) \cdot \left( \frac{1}{n_t} \sum \frac{1}{n_t} \right) \]

\[- (N \sum \frac{1}{n_t} - k) = 6k(k - 1) + 20(k - 1) - 20(1 - \frac{1}{N^2} \sum n_t^2) \]

\[E(III) = \frac{1}{N} \left( b_2^3 - 3b_2^2 + 3b_2 - 1 \right) \cdot \left( \frac{1}{n_t} \sum \frac{1}{n_t} - 45k + 45 - \frac{15}{N^2} \sum n_t^2 \right) \]
\[ E(VI) = \frac{1}{9N} \sum \left( -b_2^3 - 3b_2^2 + 3b_2 - 1 \right) \sqrt{3Nk} \sum \frac{1}{n_t} - 6N \sum \frac{1}{n_t} \]
\[ + 30k - 6k^2 - 36 + 15 \frac{1}{N} \sum n_{t-1}^2 \]

\[ E(VII) = \frac{1}{4N} \left( -b_2 - 1 \right)^3 \sqrt{-9k^2 + 27k - 33} + 15 \frac{1}{N} \sum n_{t-1}^2 \]

Further terms in the series expansion give terms only to order \( N^{-2} \).

Noting the relationship

\[ M_2 = M_2' - \frac{M_2}{1} \]

we proceed to collect terms and obtain

\[ V(M) = 2(k - 1) \left( \frac{b_2^3 - 1}{2} \right)^2 + \text{terms of order } N^{-1} \]

which may be tabulated as follows.

Terms of order \( N^{-1} \) in \( V(M) = \)

\[
\begin{array}{cccccc}
\frac{1}{36N} \cdot & N \sum \frac{1}{n_t} & k^2 & k & \text{(1)} \\
\hline
b_6 & + 9 & - 9 & - .18 & + 18 \\
b_4 & + 72 & - 114 & + 72 \\
b_4b_2 & - 108 & + 36 & + 216 & - 114 \\
b_3^2 & + 168 & - 36 & - 288 & + 156 \\
b_2^2 & - 207 & + 9 & + 306 & - 108 \\
b_2 & + 72 & - 72 & \\
(1) & - 6 & + 6 & \\
\end{array}
\]
Appendix G

DERIVATION OF MEAN AND VARIANCE OF MODIFIED BARTLETT STATISTIC, MEANS UNKNOWN

Bartlett's statistic may be written in the form

\[ M = N \ln(k_2) - \sum_{t=1}^{k} n_t \ln(k_{2t}) \]  \hspace{1cm} (1)

where: \( k_{2t} \) = Fisher's k statistics, sample variance for t-th group

\( k_2 \) = Weighted average sample variance for k groups

Write the sample variances in the form

\[ k_2 = K_2 (1 + \frac{k_2 - K_2}{K_2}) = K_2 (1 + \frac{\Delta_2}{K_2}) \]  \hspace{1cm} (2)

where: \( K_2 \) = population 2-nd cumulant, \( \sigma^2 \)

\( \Delta_2 = k_2 - K_2 \)

Recall the series expansion

\[ \ln (1 + x) = x - \frac{1}{2} x^2 + \frac{1}{3} x^3 - \cdots \]  \hspace{1cm} (3)

which permits us to write

\[ M = \frac{1}{2} K_2^{-2} \left( \sum n_t \Delta_{2t}^2 - N \Delta_2^2 \right) - \frac{1}{3} K^{-3} \left( \sum n_t \Delta_{2t}^3 - N\Delta_2^3 \right) + \cdots \]  \hspace{1cm} (4)

Now consider the modifying factor

\[ \frac{1}{1 + \frac{1}{2} c_2} \]
Since \(|c_2| < 1\) with high probability, we may write the series expansion

\[(1 + \frac{1}{2}c_2)^{-1} = 1 - \frac{1}{2}c_2 + (\frac{1}{2}c_2)^2 - (\frac{1}{2}c_2)^3 + \ldots\]  

(5)

Since \(c_2\) itself is a ratio of two random variables, it is necessary to obtain a series expansion of its denominator as follows:

\[
c_2 = \frac{k_4}{k_2} = \frac{k_4 + (k_4 - K_4)}{\sqrt{k_2(1 + \frac{k_2 - K_2}{k_2})} \lambda^2}
\]

(6)

where: \(k_4 = \) Fisher's fourth \(k\) statistic
\(K_4 = \) fourth population cumulant

Without loss of generality we may assume \(K_2 = 1\) and then write

\[
c_2 = \frac{k_4 + \Lambda_4}{(1 + \Delta_2)^2} = (k_4 + \Lambda_4)(1 - 2\Delta_2 + 3\Delta_2^2 - 4\Delta_2^3 + \ldots)
\]

(7)

where: \(\Delta_r = k_r - K_r\) \((r = 2, 4)\)

Combining the three series expansions, (4), (5) and (7), we may write the modified Bartlett statistic, \(M'\), as

\[
M' = \frac{M}{1 + \frac{1}{2}c_2} = \zeta_\frac{1}{2}(\sum n_t \Delta_{2t}^2 - N\Delta_2^2) - \frac{1}{3}(\sum n_t \Delta_{2t}^3 - N\Delta_2^3)
+ \frac{1}{4}(\ldots \ldots \ldots) + \frac{1}{2}(k_4 + \Lambda_4)(1 - 2\Delta_2 + 3\Delta_2^2 - \ldots)
+ \frac{1}{4}(k_4 + \Lambda_4)^2(1 - 2\Delta_2 + 3\Delta_2^2 - \ldots)^2 - \frac{1}{6} \ldots
\]

(8)
By use of the technique of F. N. David (1969) to evaluate the expectations of powers and cross products of the Δ functions, the expected value of $M'$ to order $N^0$ may be found. From these tables it is immediately apparent that for powers and cross products of the $Δ$'s of order three and greater, the expectations are of order $N^{-1}$ or greater, so that we need consider only terms of order two in the $Δ$'s. Therefore we may shorten (9) to

$$M' = \frac{1}{2} \left[ \sum n_t Δ_{2t}^2 - NΔ_2^2 \right] \times \left[ 1 - \frac{1}{2} K_4 + \left( \frac{1}{2} K_4 \right)^2 - \cdots \right]$$

(10)

$$= \frac{1}{2} \left[ \sum n_t Δ_{2t}^2 - NΔ_2^2 \right] \times \left[ 1 + \frac{1}{2} K_4 \right]^{-1}$$

(11)

Now taking expectations with the aid of David's tables, we obtain

$$E(M') = \frac{1}{2} (k - 1)(K_4 + 2)(1 + \frac{1}{2} K_4)^{-1}$$

(12)

$$= k - 1 \quad \text{q.e.d. expectation}$$

By use of the same three series expansions we find the variance of $M'$ to order $N^0$ to be

$$E(M'^2) = k^2 - 1$$

(13)
\[ V(M') = E(N'^2) - \left( E(M') \right)^2 = k^2 - 1 - (k - 1)^2 = 2(k - 1) \] (14)

This justifies the assertion that \( M' \) is distributed approximately as chi square with \( k - 1 \) degrees of freedom.

It is of some interest to obtain the variance to order \( N^{-1} \), though the result is so complex to be of little value.