A COMPARISON OF METHODS IN USE FOR EVALUATING CONTRASTS AMONG MEANS

A Thesis
Presented to
THE DEPARTMENT OF ACTUARIAL MATHEMATICS
AND STATISTICS
THE UNIVERSITY OF MANITOBA

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
BRIAN DOUGLAS MACPHERSON
April 1963

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

In the analysis of experimental data, the experimenter is faced with the problem of attempting to isolate the particular effect due to an individual treatment or combination of treatments. Methods have been proposed to accomplish this purpose but the various methods are known to give widely differing results in many instances. This study considers these methods and their properties in order to recommend which method is the best to use (1) in a general situation and (2) when particular situations are encountered either in the design of the experiment or the philosophy of the experimenter.

The study includes an outline of the following procedures,

(a) Least Significant Difference (LSD)
(b) Fisher's Modified technique
(c) Tukey's Allowance procedure
(d) Scheffé's test
(e) Dunnett's test
(f) Student-Newman-Keuls procedure
(g) Tukey's 1953 procedure
(h) Duncan's New Multiple Range Test
(i) Scheffé's Modified technique

with particular emphasis given to the method of application and critical values used in each. As an aid to the understanding of the differences between the critical values of the different procedures, a description of the distribution of the range and the tables of the range, studentized range, and special tables for Duncan's New Multiple Range Test and Dunnett's test is given.

The procedures are compared on the basis of several experimental situations and statistical techniques. These situations and techniques are such that should they be

encountered in an experiment, the multiple comparison technique to be used in analysing the experiment is indicated. The effect on the choice of the test procedure by the inclusion of a control treatment in the experiment and by the a priori specification of the treatments to be tested is considered. The statistical techniques of confidence limits, all possible contrast and an F-test in the analysis of variance are discussed. As a guide in the application of the methods and an illustration of the topics considered, a series of four experiments is analysed and presented.

It is found that in the analysis of an experiment with a control treatment, Dunnett's procedure should be used. If the tests are specified a priori, the LSD procedure is valid but this is the only situation in which it may be applied. Scheffé's method is very insensitive but should be applied to the testing of contrasts involving more than two means. Tukey's Allowance procedure is recommended if confidence limits are to be placed about the true treatment mean difference. In the general situation, a strong recommendation is given to the Student-Newman-Keuls procedure.

CHAPTER I

THE PROBLEM AND DEFINITIONS OF TERMS USED

In the statistical analysis of experimental material, one of the most common questions to which an answer is desired is whether observed results are caused by particular conditions imposed upon the material by the experimenter or by the operation of chance factors only. The decision as to whether or not the treatment or variety effect in general is due to chance, offers very little difficulty to the statistician. When an attempt is made to isolate the particular causal effect or effects, increased complications are encountered.

I. THE PROBLEM

Statement of the Problem

In recent years, statistical literature has contained many articles dealing with this particular problem of determining the individual effects. As a result, the statistician is faced with a wide variety of differing procedures all purporting to do the same work.

In spite of the claims of the various authors, it has been noted with great frequency, that the application of each of the methods to the same data will result in a host of different conclusions. Since the methods do not give identical results, it should be possible to find one which is superior or certain situations in which a particular method is the best to use.

It is the purpose of this study to examine these proposed methods to determine (1) whether for a given set of experimental conditions there exists a best method, and (2) whether the characteristics of a method are particularly suitable or unsuitable for various types of experimental conditions.

Importance of the Study

The ability to isolate a particular treatment or variety as being superior to other treatments or varieties is becoming increasingly important. Larger and more complex experimental designs are being used in many scientific fields such as agriculture, chemistry, and medicine for the purpose of putting a new material or group of materials to a discriminating statistical analysis. Many of these experiments are easily analyzed to the stage where the next step is the actual isolation of the causal effects that are statistically important. If, at this stage, a procedure is used which is not valid, is too sensitive, or is not sensitive enough, not only will misleading results be obtained, but much valuable scientific work will have been lost. It is very important that some guidance be given to those people who may be faced with this problem so that the information available in the data will be fully utilized.

II. DEFINITIONS OF TERMS AND PROCEDURES USED

Throughout this study several terms and procedures are used freely. The following is a listing of several of these terms together with their definitions and symbolic representation.

Population

A population consists of all possible values of a variable. It is usually desired that the parameters of the population be obtained so that the population may be fully described. By virtue of the fact that the population may be infinite or finite, it may or may not be possible to obtain the actual parameter values.

Sample

A sample is a part of a population. The usual situation encountered is that it is impossible to observe the entire population at any one time. For the purpose of studying the

characteristics of the population a representative part of it is obtained and inferences about the population are made on the basis of the sample results.

Statistic

A value obtained from a sample with a view to characterizing a parameter of the population from which the sample is obtained.

Arithmetic Mean

The sum of the values of a number of variables divided by the total number of variables. Standard notation refers to the mean as $(\bar{x})$ if obtained from a sample, or (μ) if obtained from a population. The arithmetic mean is calculated by the use of the formula:

$$\bar{x} = \frac{\sum x_i}{n}$$

where $\sum$ denotes the summation over the i sample values. The arithmetic mean is a measure of location or central tendency.

Variance, or Mean Square

The population variance is the average value of the squared deviations of the individual variables from the population mean. Symbolically it is represented by σ^2 and written:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

The sample variance is an unbiased estimate of the population variance and is represented symbolically by s^2 where

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The variance is a measure of the dispersion of the variables about their central value.

Standard Deviation

This is defined as the square root of the variance and is denoted by (σ) for a population and (S) for a sample.

Standard Error

Since it is possible to obtain a large number of samples from a population, it is to be expected that statistics calculated from these samples will themselves be subject to random variation. A measure of this variation is the standard deviation or standard error of the sample statistic. The standard error referred to most frequently in this study is the standard error of the mean, denoted by ($S_{\bar{x}}$) for the sample values. The relation between the standard deviation of the variables and the standard error of a mean is given by:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

Range

The range of a sample is defined as the difference between the maximum and minimum values in the sample. It may be written symbolically as:

$$w = x_{max.} - x_{min.}$$

Standard Range

The ratio of the range of a sample to the standard deviation of the population from which the sample is drawn.

Studentized Range

The ratio of the range of a sample to the sample value of the standard deviation. It may be written as:

$$q = \frac{w}{S}$$

The values of the range and the standard deviation are obtained from independent samples.

Degrees of Freedom

A term which is used to denote the number of independent comparisons that can be made among the members of a sample. Throughout this study, the degrees of freedom will be denoted by "f".

Analysis of Variance

A procedure by which the total variation contained in a set of observations may be separated into components readily associated with defined sources of variation used to classify the observations. The variation is measured as the sum of squares of the deviations from the mean. To illustrate the procedures and terms closely associated with the analysis of variance, an example is presented.

The observations in Table I are percentages of barley kernels dehulled during an abrasive test. The object of the experiment, arranged as a randomized complete block with six complete replications, is to test for differences among hull characteristics of four barley varieties.

TABLE I

PERCENTAGES OF BARLEY KERNELS DEHULLED DURING AN ABRASIVE TEST

Replications	Varieties				Total
	Parkland	MC247	ND_{B16}	ND_{B117}	
I	32.0	11.2	11.2	27.2	81.6
II	29.6	8.4	19.2	36.8	94.0
III	28.8	5.2	18.4	42.8	95.2
IV	34.0	10.0	16.8	50.0	110.8
V	29.6	5.2	11.2	34.8	80.8
VI	24.8	3.6	7.2	35.6	71.2
Total	178.8	43.6	84.0	227.2	533.6

It is possible to obtain a sum of squares identifiable with the variation among the total observations and to subdivide this into effects due to the varieties, the replications, and unexplained experimental error variation.

The total sum of squares is obtained by summing the squares of the individual observations. Since the desired sum of squares is to be in terms of deviations from the mean, correction of this total sum of squares for the contribution due to the mean is necessary. The correction factor is calculated as the square of the grand total divided by the number of observations. The total sum of squares is thus found to be:

$$(32.0)^2 + (29.6)^2 + (28.8)^2 + \ldots\ldots + (34.8)^2 + (35.6)^2 - \frac{(533.6)^2}{24} = 4072.61$$

The sum of squares due to replications is:

$$\frac{(81.6)^2 + (94.0)^2 + \ldots\ldots + (71.2)^2}{4} - \frac{(533.6)^2}{24} = 244.37$$

where the observed replicate totals are obtained from Table I and the divisor four is the number of observations per replicate.

The sum of squares due to varieties is obtained in a similar manner:

$$\frac{(178.8)^2 + (43.6)^2 + (84.0)^2 + (227.2)^2}{6} - \frac{(533.6)^2}{24} = 3560.66$$

where the variety totals are obtained from Table I and the divisor six is the number of replications.

The sum of squares for experimental error can now be obtained by subtracting both the replicate sum of squares and the variety sum of squares from the total sum of squares.

These sum of squares results may be summarized in the standard analysis of variance table. Table II gives the results as follows:

TABLE II

ANALYSIS OF VARIANCE OF THE DATA IN TABLE I

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares
Varieties	3	3560.66	1186.89
Replications	5	244.37	48.87
Error	15	267.58	17.84
Total	23	4072.61	

The column, degrees of freedom, in the analysis of variance table is obtained by the application of the definition given for this quantity. Among the twenty-four observations, twenty-three independent comparisons can be made and as a result, the total number of degrees of freedom for this experiment will be twenty-three. In accordance with the three subdivisions of the total sum of squares, the twenty-three degrees of freedom are divided into three parts representing the number of independent comparisons among varieties, among replicates and among the experimental error. The number of degrees of freedom among varieties and among replicates is one less than the number of items which may be compared within each classification. After removing the five degrees of freedom for replicates and the three degrees of freedom for varieties, the remaining fifteen degrees of freedom are attributed to experimental error.

The column headed "mean square" is found simply by dividing the sum of squares for each classification by the degrees of freedom appropriate to it. The value of the mean square for error provides an unbiased estimate of the error variance σ^2. This estimate, S^2, enables the experimenter to obtain the standard error of a variety mean. It is found by calculating:

$$S_{\bar{x}} = \sqrt{\frac{S^2}{r}}$$

where "r" represents the number of replications. In this example the value obtained is:

$$S_{\bar{x}} = \sqrt{\frac{17.84}{6}}$$
$$= 1.72$$

Having completed the analysis of variance table, the next step in the analysis of the experiment is to test the hypothesis that each of the four varieties has shown equal results. The test is achieved by the use of the ratio of the mean square for varieties to the mean square for error. This ratio follows the F-distribution and its value may thus be compared with a critical value obtained from the tables of the distribution of F in order to judge its significance. The observed value of F in this case is found to be:

$$\frac{1186.89}{17.84} = 66.53$$

Since the mean square for varieties carries three degrees of freedom and the mean square error has fifteen degrees of freedom, this calculated F-statistic must be compared with the tabular F-value with three and fifteen degrees of freedom in the numerator and denominator respectively. The critical value applicable in this case is obtained as:

$$F_{3,15,(0.05)} = 3.29$$

at the 0.05 probability level and,

$$F_{3,15,(0.01)} = 5.42$$

at the 0.01 probability level. Since the calculated value exceeds the tabulated value at probability level 0.01, it can be stated that the probability that the calculated value could have reached the value 66.53 by chance is less than 0.01. The calculated F-statistic is thus declared to be highly significant. Had the F-ratio exceeded the 0.05 critical value but not the 0.01 value, the ratio would be declared significant. It is thus

necessary to reject the claim that the varieties are all alike and conclude that there is evidence to support a claim that at least one of the varieties is different from the rest.

Contrast

A contrast among variety totals is defined as any linear function of the variety totals of the form:

$$C_i = c_{i1}V_1 + c_{i2}V_2 + \ldots\ldots + c_{ik}V_k$$

such that,

$$c_{i1} + c_{i2} + \ldots\ldots\ldots + c_{ik} = 0$$

The contrast carries with it a component of the variety sum of squares with one degree of freedom. The value of the component of the sum of squares is calculated as:

$$\frac{C_i^2}{r\,D_i}$$

where

$$D_i = c_{i1}^2 + c_{i2}^2 + \ldots\ldots + c_{ik}^2$$

and "r" is the number of replications.

Orthogonal Contrasts

Two contrasts of the form:

$$C_1 = c_{11}V_1 + c_{12}V_2 + \ldots\ldots + c_{1k}V_k$$
$$C_2 = c_{21}V_1 + c_{22}V_2 + \ldots\ldots + c_{2k}V_k$$

are said to be orthogonal if the sum of the products of corresponding coefficients of the V_i totals, is equal to zero. This may be represented as:

$$c_{11}c_{21} + c_{12}c_{22} + c_{13}c_{23} + \ldots\ldots + c_{1k}c_{2k} = 0$$

If in a set of contrasts, every pair of contrasts is orthogonal, the set is said to be an orthogonal set of contrasts. If any pair of contrasts is not orthogonal, the set of contrasts is said to be non-orthogonal.

An important property of an orthogonal set of contrasts is that the sum of the component sums of squares will be equal to the sum of squares for the over-all variety effect. Since each contrast carries one degree of freedom, it is possible to have as many contrasts as there are degrees of freedom in the variety sum of squares.

To illustrate the contrast concepts, use is made of the variety totals obtained from the data in Table I.

PARKLAND	MC247	ND_{B16}	ND_{B117}
178.8	43.6	84.0	227.2

Since the sum of squares for varieties contains three degrees of freedom, it is possible to obtain three contrasts each carrying one degree of freedom. It might be of interest to divide the varieties into two groups, one containing varieties Parkland and MC 247 and the other containing ND_{B16} and ND_{B117}. Component sums of squares may thus be obtained representing the difference between the groups and also the differences between the varieties within each of the groups. These effects may be represented in terms of combinations of the coefficients of the variety totals as follows:

1.	+1	+1	-1	-1
2.	+1	-1	0	0
3.	0	0	+1	-1

It is noted that each of these three effects is a contrast since the sum of the coefficients for each effect is zero. They also form an orthogonal set since the sum of the cross products total zero for every pair of contrasts.

The numerical value for these contrasts will be obtained as follows:

1. $\frac{(178.8 + 43.6 - 84.0 - 227.2)^2}{(6)(4)} = 328.56$

2. $\frac{(178.8 - 43.6)^2}{(6)(2)} = 1523.25$

3. $\frac{(84.0 - 227.2)^2}{(6)(2)} = 1708.85$

The sum of these component sums of squares is equal to 3560.66 which is identical with the value given in Table II for the sum of squares due to varieties.

Since each of these components is a sum of squares with one degree of freedom, the particular effect may be tested by obtaining the ratio of the value of the component to the mean square for error. This ratio is an F-ratio and may be compared with tabular value from the tables of the F-distribution with the appropriate degrees of freedom.

An example of a non-orthogonal set of contrasts may be represented in terms of the coefficients of the variety totals as:

1.	+3	-1	-1	-1
2.	+1	+1	-1	-1
3.	0	0	+1	-1

Consideration of these three effects individually reveals that in each case the sum of the coefficients is zero. As dictated by the definition, each effect is thus a contrast. The sum of products of the corresponding variety total coefficients in the three paired combinations of effects is not zero in every case and as a result the set is non-orthogonal.

The calculated component sum of squares for these contrasts are

1. 458.04
2. 328.56
3. 1708.85

The sum of these component effects is equal to 2495.45 which is not equal to the sum of squares for varieties.

CHAPTER II
DESCRIPTION OF METHODS

The multiple comparison techniques that are discussed here differ among themselves in three ways:

a. their purpose

b. their application

c. their results

In spite of these differences, similarities among the tests make it possible to group the methods according to the critical values used. In this chapter, each test is presented, giving where applicable, a brief historical and theoretical account of its development, its method of application and the determination of the critical values used.

The following is a list of the methods presented in this chapter, grouped according to the form of the critical value or values used:

I. Fixed Critical Value
   - a. least significant difference
   - b. Fisher's modified technique
   - c. Tukey's allowance procedure
   - d. Scheffé's test
   - e. Dunnett's test

II. Multiple Critical Values
   - a. Student-Newman-Keuls procedure
   - b. Tukey's 1953 procedure
   - c. Duncan's New Multiple Range test
   - d. Scheffé's modified technique

I. FIXED CRITICAL VALUE

Least Significant Difference (LSD)

The LSD method of making comparisons among means, is perhaps the first method that was used in making additional

tests of significance following a significant analysis of variance F-test. Its widespread use is due to its relative ease of application and its correspondence to statistical techniques known to most experimenters.

The test is applied to differences between all possible pairs of means in the experiment. Each difference is compared with the critical value or LSD and those differences which exceed the critical value are said to be significant.

The critical value may be expressed as:

$$\sqrt{2}\, t(\alpha, f)\, S_{\bar{x}}$$

where $S_{\bar{x}}$ is the standard error of a treatment mean and $t(\alpha, f)$ is the tabular value of student's-t distribution with "f" degrees of freedom at the α level of significance. The critical value may also be expressed in terms of the studentized range as:

$$q(\alpha, 2, f)\, S_{\bar{x}}$$

where $q(\alpha, 2, f)$ is the tabular value of the distribution of the studentized range for two means, at the α level of significance and with f degrees of freedom.

Some writers have indicated that the LSD procedure should not be applied to all possible comparisons among pairs of treatment means, but should be carried out in such a manner that each mean will appear only once in a pair. However valid this may be, most experimenters use the procedure as presented above.

Fisher's Modified Technique

Fisher (1935) has suggested a procedure which should be followed when a non-significant value of F is obtained in the ratio of mean square for treatments to mean square for error. Ordinarily in this situation, the experimenter would not wish to proceed further in attempting to determine whether individual pairs of treatments differ. As a guide to further experimentation, however, a closer look at the treatment differences might be informative.

To ensure that only the large treatment differences be declared significant, the level of significance or the probability of declaring a significant difference when, in fact, the difference is not significant, is made smaller. The net effect of causing the level of significance, α, to be lowered is that the critical value is made larger, thus requiring a larger treatment mean difference to exceed it.

For the experiment testing k treatments, the new value of the level of significance is obtained by dividing α by the number of ways a pair of treatment means may be selected from the k treatment means. The critical value may be written as

$$\sqrt{2}\, t(\alpha/C_2^k,\ f)\ S_{\bar{x}}$$

where $t(\alpha/C_2^k,\ f)$ is the tabular value of student's-t distribution with f degrees of freedom at the probability level α/C_2^k, and $S_{\bar{x}}$ is the standard error of a treatment mean. The critical value may also be expressed in terms of the distribution of the studentized range:

$$q(\alpha/C_2^k,\ 2,\ f)\ S_{\bar{x}}$$

Differences between all pairs of treatment means are obtained and each compared with the critical value. Differences in excess of the critical value are said to be significant at the α level of significance.

Tukey's Allowance Procedure

Tukey (1952) introduced a procedure by which it is possible to construct confidence intervals about the true population mean differences. The usual approach to this test however, is to consider it as giving tests of significance very similar in nature to the LSD procedure previously mentioned. The test has gained much popularity by virtue of its ease of application and because it takes into consideration the number of treatments under test.

All possible differences between pairs of observed treatment means are obtained and each difference is compared with the critical value. If the observed difference exceeds the critical value, it is declared significant and the two treatment means involved are declared to be from different populations.

The critical value used in the test procedure is based on the distribution of the studentized range and may be expressed as:

$$q(\alpha, k, f)\, S_{\bar{x}}$$

where $q(\alpha, k, f)$ is the tabular value of the studentized range at the α level of significance for the k treatment means in the experiment, and with f degrees of freedom.

Scheffé's Test.

Scheffé (1953) devised a method for selecting significant differences among all possible contrasts among means following a significant analysis of variance F-test. The method is such that it is possible to construct confidence limits about the true value of the contrast. The more common approach.however, is to use the method for making tests of significance.

The critical value for the test of significance, or the "allowance" for setting confidence limits, is based on the F-distribution rather than the studentized range distribution. It may be expressed as:

$$\sqrt{2(k-1)\, F_{\alpha, (k-1, f)}}\; S_{\bar{x}}$$

where k is the number of treatments under test and $F_{\alpha, (k-1, f)}$ is the upper α.point of the F-distribution with $k-1$ and f degrees of freedom.

The differences between all possible pairs of observed treatment means are obtained and each difference is compared with this critical value. If the observed difference exceeds the critical value, the difference is declared significant.

This method can be extended to cover contrasts of any form, involving any number of means. The critical value will no longer incorporate the standard error of a treatment mean, but will use instead, the standard error appropriate to the form of the contrast under test. The critical value will remain otherwise unchanged.

Dunnett's Test

A fifth method proposed by Dunnett (1955) is one recommended for making all comparisons with a control treatment. In its application, confidence limits may be constructed about the true difference between treatment and control, or tests of significance may be applied to the observed differences.

As in all previous methods, a single critical value is needed for deciding the significance of an observed difference between treatment and control. All the differences between treatments and the control are tabulated and each difference is compared with the critical value. If the observed difference exceeds the critical value, the treatment involved is said to be significantly different from the control treatment. The critical value may be written as:

$$\sqrt{2}\, D\,(\alpha, k, f)\, S_{\bar{x}}$$

where $D\,(\alpha, k, f)$ is the multivariate analogue of Student's-t distribution at the α level of significance for k treatments, and with f degrees of freedom. Tables of $D\,(\alpha, k, f)$ are given by Dunnett (1955) for one-tailed and two-tailed test procedures.

II MULTIPLE CRITICAL VALUES

Student-Newman-Keuls Procedure

Student (1927) considers the problem of errors which appear in repetitive routine analyses.such as might occur in a commercial analytical chemical firm. Over an extended period of time, experience with the measuring of characteristics of a given material will provide the experimenter.with information about the experimental values to be expected. Student provides a procedure for the rejection of observations which appear not to be in line with the expected results. His procedure involves the use of values, R_i, from the tables of the distribution of the range, the value depending on the number of observations that have been made.

Let R_n be the value of the range expected from a sample of size n, such that the chance of.observing a range greater than R_n is less than some value of P (e.g. $P = 0.05$). If in the analysis an observed range r_n is found to be greater than R_n, an additional observation should be made. If the new range r_{n+1} is less than R_{n+1}, the most discordant observed value is rejected and the new r_n compared with R_n. If r_n is now less than R_n, the mean of the n observations is accepted but if r_n is larger than R_n, a further observation is obtained and the procedure continued with the n+2 observations, until finally a sample of at least n is obtained lying within the required limits.

Keuls (1952) used the results of this procedure and also incorporated the work done by Newman (1939) on the studentized range in presenting a method for judging all pairs of treatment mean differences. The procedure set forward by Keuls (1952) incorporates varying critical values of the studentized range depending upon the number of treatment means lying between the two means being considered.

If an experiment involving k means is being considered, the difference involving the largest and the smallest means would be compared with the critical value for k means. The difference involving the largest and the second smallest would be compared with the critical value for k-1 means and similarly for all possible treatment differences. The critical value may be written as:

$$q(\alpha, p, f)\, S_{\bar{x}} \qquad p = 2, 3, \ldots\ldots, k$$

where $q(\alpha, p, f)$ is the tabular value of the studentized range at the α level of significance for p means with f degrees of freedom.

The difference between the i^{th} and j^{th} treatment means is declared significant if it exceeds the appropriate critical value corresponding to the subset of p means with the i^{th} and j^{th} means as the largest and smallest values in the subset. Symbolically the i^{th} and j^{th} treatments are declared significantly different if,

$$\left| \bar{x}_i - \bar{x}_j \right| > q(\alpha, p, f)\, S_{\bar{x}}$$

where $i \neq j = 1, 2, \ldots\ldots k$ and $p = 2, 3, \ldots\ldots k$.

The only exception to the above rule of rejection is that a treatment difference which falls within a subset with a non-significant range shall be declared non-significant.

Tukey's 1953 Procedure

Tukey (1953) proposed a method which offers a compromise between the Tukey allowance procedure and the Student-Newman-Keuls procedure. The only change from the Student-Newman-Keuls procedure is in the form of the critical value which now assumes a value midway between the critical value for Tukey's allowance and the Student-Newman-Keuls procedures.

If an experiment involving k means is being considered and the two means under test are the largest and smallest in a subset of p means, the critical value may be expressed as:

$$\frac{1}{2}\left[q(\alpha, k, f) + q(\alpha, p, f) \right] S_{\bar{x}}$$

where α is the level of significance, p is the number of means in the subset, k is the total number of means in the experiment, and f is the degrees of freedom. The values of $q(\alpha, k, f)$ and $q(\alpha, p, f)$ are obtained from the tables of the studentized range.

The i^{th} and j^{th} treatment means are declared significantly different if,

$$\left| \bar{x}_i - \bar{x}_j \right| > \frac{1}{2}\left[q(\alpha, k, f) + q(\alpha, p, f) \right] S_{\bar{x}}$$

No treatment difference contained in a subset which has a non-significant range shall be declared significant.

Duncan's New Multiple Range Test

Duncan (1955) has developed a test, very similar to the previously mentioned Student-Newman-Keuls procedure, for making all possible comparisons among treatment means. It differs from the latter test however in that it uses tables of the studentized range specially constructed for this test.

Duncan (1955) has defined a new term, "protection level" (γ) of the test of the hypothesis $\mu_1 = \mu_2$ as the probability of deciding that there is no appreciable difference between the means μ_1 and μ_2 when, in fact, they are equal. A calculation of the protection level thus measures the protection against wrongly finding a significant difference between two equal means. This may be denoted by:

$$\gamma = \Pr\left[\text{decision } (\underline{1, 2}) / \mu_1 = \mu_2 \right]$$

where $(\underline{1, 2})$ denotes that the means 1 and 2 do not differ.

Extending this to the case where three means are involved, four protection levels can be found corresponding to the level defined for the two means case. These may be denoted by:

$$\gamma(1,2) = P_r\left[\text{decision } (1,2)/\mu_1 = \mu_2\right]$$
$$\gamma(1,3) = P_r\left[\text{decision } (1,3)/\mu_1 = \mu_3\right]$$
$$\gamma(2,3) = P_r\left[\text{decision } (2,3)/\mu_2 = \mu_3\right]$$
$$\gamma(1,2,3) = P_r\left[\text{decision } (1,2,3)/\mu_1 = \mu_2 = \mu_3\right]$$

It is easily seen that since α is the probability of deciding that two means differ when in fact they do not differ, γ can be expressed as $1 - \alpha$ for the two mean case. Thus if α is set at 0.05, γ will then be 0.95. The question that arises in any test of n means, given that γ_2 is an appropriate value for the two mean protection level, is what values $\gamma_3, \gamma_4, \ldots\ldots, \gamma_n$ should be regarded as satisfactory for the three mean, four mean, ... n mean protection levels?

It is noted that a test on a subset of p means is actually a test of whether (p-1) orthogonal contrasts between the true population means differ from zero. If these (p-1) contrasts were tested individually, each at the α level of significance, the probability of correctly deciding that there is no difference between the population means would be given by $(1 - \alpha)^{p-1}$. Even though this level is less than 0.95, Duncan finds it unobjectionable because of the choice of 0.95 for each individual test. The test was constructed with protection levels changing with the size of the subset of means being considered. The protection level for the p mean subset thus becomes,

$$\gamma_p = \gamma_2^{p-1}$$

The critical value for the test may be written as,

$$N(\gamma_p, \alpha, p, f)\, S_{\bar{x}}$$

where $N(\gamma_{p,\alpha}, p, f)$ is the tabular value of the distribution of the studentized range for p means at the appropriate protection level based on the α level of significance with f degrees of freedom.

The difference between the i^{th} and j^{th} treatment means is declared significant if it exceeds the critical value appropriate to the subset of means with the i^{th} and j^{th} means as the largest and smallest members of the subset, and provided also that the i^{th} and j^{th} means are not contained in a subset which has previously been declared not significant.

Scheffé's Modified Technique

It is felt that a modification to the method proposed by Scheffé (1953) might be of value in view of the appeal of the sequential procedures of Student-Newman-Keuls and Duncan.

The critical value is now obtained taking into consideration the number of treatment means in the subset. The value may be expressed as:

$$\sqrt{2\,(p-1)\,F_{\alpha,\,(p-1,\,f)}}\;\;S_{\bar{x}}$$

where p is the number of means in the subset and $F_{\alpha,\,(p-1,f)}$ is the upper α point of the F-distribution with (p-1) and f degrees of freedom.

The difference between two means is declared significant provided it exceeds the critical value appropriate to the size of the subset of which the two means are the largest and the smallest values.

III TABLE OF CRITICAL FACTORS

The following table, Table III presents a comparison of the form and the numerical value of the critical values used by each of the tests in an experiment with ten treatment means. The numerical values given are critical value factors only, and must be multiplied by the standard error of a treatment means before the treatment mean differences can be tested. All values are given at the 0.05 level of significance with an infinite number of degrees of freedom and for various subset sizes.

TABLE III

COMPARISON OF CRITICAL VALUE FORMS AND FACTORS FOR VARIOUS TEST PROCEDURES

TEST	Form of the Critical Value	Critical Value Factors Various Subset sizes - p						
		2	3	4	5	6	8	10
L S D	$\sqrt{2}\, t\,(\alpha, f)\, S_{\bar{x}} = q\,(\alpha, 2, f)\, S_{\bar{x}}$	2.77	2.77	2.77	2.77	2.77	2.77	2.77
FISHER'S	$\sqrt{2}\, t\,(\alpha/C_2^k, f)\, S_{\bar{x}} = q\,(\alpha/C_2^k, 2, f)\, S_{\bar{x}}$	4.65	4.65	4.65	4.65	4.65	4.65	4.65
TUKEY'S ALLOWANCE	$q\,(\alpha, k, f)\, S_{\bar{x}}$	4.47	4.47	4.47	4.47	4.47	4.47	4.47
SCHEFFÉ'S TEST	$\sqrt{2\,(k-1)\,F_{\alpha,\,(k-1,\,f)}}\, S_{\bar{x}}$	5.82	5.82	5.82	5.82	5.82	5.82	5.82
DUNNETT'S TEST	$\sqrt{2}\, D\,(\alpha, k, f)\, S_{\bar{x}}$	2.75	2.75	2.75	2.75	2.75	2.75	2.75
STUDENT-NEWMAN-KEULS	$q\,(\alpha, p, f)\, S_{\bar{x}}$	2.77	3.82	3.63	3.86	4.03	4.29	4.47
TUKEY'S 1953	$\frac{1}{2}\left[q\,(\alpha, k, f) + q(\alpha, p, f)\right] S_{\bar{x}}$	3.62	3.89	4.05	4.16	4.25	4.38	4.47
DUNCAN'S NEW MULTIPLE RANGE	$N\,(\gamma_{p,\alpha}, p, f)\, S_{\bar{x}}$	2.77	2.92	3.02	3.09	3.15	3.23	3.29
SCHEFFÉ'S MODIFIED	$\sqrt{2\,(p-1)\,F_{\alpha,\,(p-1,\,f)}}\, S_{\bar{x}}$	2.77	3.46	3.95	4.47	4.70	5.30	5.82

CHAPTER III
CONSTRUCTION OF TABLES

An understanding of the difference between Duncan's test and others necessitates an investigation into the construction of tables of the studentized range. Duncan's test makes use of a table of percentage points of the studentized range which is different from tables used by Keuls (1952), Tukey (1953), and others. An understanding of most testing methods requires, therefore, a closer look at the calculation of critical values or percentage points used in each case.

The present chapter deals with the historical and theoretical background of the distribution of the range and the studentized range. Differences between the table constructed for Duncan's test and the tables used for the tests proposed by others, are illustrated.

I. DISTRIBUTION OF THE RANGE.

Hoel (1954) presents an excellent text book description of the derivation and uses of the distribution of the range. The general form of the distribution of the range for a sample of size n, which has u and v as its largest and smallest members is given by the following definition: If the continuous variable x has the frequency function, f(x), and if x assumes values in the interval (a,b) only, then the frequency function of the range, g (w), for a random sample of size n, is expressed as:

$$g(w) = n\ (n-1) \int_a^{b-w} f(u)\ f\ (u+w) \left[\int_u^{u+w} f\ (x)\ dx \right]^{n-2} du$$

The exact form of the distribution of the range for a sample of size two from a normally distributed population can easily be obtained. The frequency function thus becomes,

$$g(w) = 2\int_a^{b-w} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}u^2\right) \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}(u+w)^2\right\} du$$

$$= \frac{1}{\pi}\int_a^{b-w} \exp\left\{-\tfrac{1}{2}\left[u^2 + (u+w)^2\right]\right\} du$$

$$= \frac{1}{\pi}\int_a^{b-w} \exp\left(-\frac{w^2}{4}\right) \exp\left\{-\frac{(u+w)^2}{2}\right\} du$$

and since $-\infty \leq a < b \leq \infty$

$$g(w) = \frac{1}{\pi} \exp\left(-\frac{w^2}{4}\right) \int_{-\infty}^{\infty} \exp\left\{-\frac{(u+w)^2}{2}\right\} du$$

so that the result

$$g(w) = \frac{1}{\sqrt{\pi}} \exp\left(-\frac{w^2}{4}\right)$$

is obtained.

Unfortunately, the form of the distribution is very cumbersome for sample sizes larger than two, and if the form of the frequency function $f(x)$ is complex.

In spite of the difficulties involved, Tippett (1925) has constructed tables of the probability integral of the standardized range, basing his work on the moments of the distribution of the range. The numerical work used in the preparation of these tables has proved to be very inaccurate as a result of which Tippett's work now holds more historical than practical interest. Nevertheless, this work has provided the impetus and the basis for much of the more recent work on this topic.

McKay and Pearson (1933) also give the distribution function of the range from samples of size n. They give, in particular, the exact form of the distribution when the parent population is rectangular or straight line. In addition to this, they derive the distribution of the range for a sample of size three drawn from a normal population and provide its r^{th} moment.

II TABLES OF THE STUDENTIZED RANGE

In most practical applications of the theory involving the range, the only knowledge available about the population standard deviation, σ, is the sample estimate, S. As a result of this, Newman (1939) considers the sampling distribution of $q = \frac{w}{S}$, the studentized range. In particular, he calculates the 5% and 1% points of the probability integral of the studentized range.

It is known that the probability distribution of S, the sample standard deviation, may be written in the form,

$$p(S) = \frac{f^{\frac{1}{2}f} S^{f-1}}{2^{\frac{1}{2}(f-2)} \Gamma(\frac{1}{2}f) \sigma^f} \exp\left(-\frac{f S^2}{2\sigma^2}\right)$$

Write p(w) for the probability distribution of the range whose precise value is known for samples of size two and three from a normal population. The expected value of the studentized range is expressed as,

$$E(q) = \int_0^\infty q\, p(q)\, dq$$

$$= \int_0^\infty \int_0^\infty w\, S^{-1}\, p(w)\, p(S)\, d w\, d S$$

since w and S are independent. Thus,

$$E(q) = \int_0^\infty w\, p(w)\, d w \int_0^\infty S^{-1}\, p(S)\, d S$$

$$= E(w) \int_0^\infty S^{-1}\, p(S)\, d S$$

Tippett (1925) has tabulated for varying sizes of n, the values of E $(\frac{w}{\sigma})$, thus if

$$E\ (w/\sigma)\ \int_0^\infty \sigma\ S^{-1}\ \ p\ (S)\ d\ S$$

is considered, it is only necessary in the evaluation of E (q) to determine,

$$\int_0^\infty \sigma\ S^{-1}\ p\ (S)\ d\ S$$

Substituting the expression for p (S) given previously, this integral becomes,

$$\frac{f^{\frac{1}{2}f}}{2^{\frac{1}{2}}\ (f-2)\ \Gamma(\frac{1}{2}f)\ \sigma^{f-1}} \int_0^\infty S^{f-2}\ \exp\ (\frac{-f\ S^2}{2\sigma^2})\ dS$$

by making the substitution

$$\frac{f\ S^2}{2\ \sigma^2} = x$$

the integral is found to be equal to

$$\frac{\sqrt{\frac{1}{2}f}\ \ \Gamma\{\frac{1}{2}\ (f-1)\}}{\Gamma(\frac{1}{2}f)}$$

The E (q) has in this way been expressed as:

$$E\ (q) = E\ (\frac{w}{\sigma}) \left[\frac{\sqrt{\frac{1}{2}}\ f\ \ \Gamma\{\frac{1}{2}\ (f-1)\}}{\Gamma(\frac{1}{2}\ f)} \right]$$

Newman (1939) has presented a table of the factors to which E (w/σ) is multiplied to give the E (q).

In order to obtain significance levels for $q = w/S$, it is necessary to evaluate the integral,

$$\alpha = \int_{q_\alpha}^{\infty} p\,(q)\, d\, q$$

$$= \left\{ p\,(w) \left[\int_{o}^{w/q_\alpha} p\,(S)\, d\, S \right] d\, w \right\}$$

Newman (1939) has accomplished the evaluation of these integrals using various quadrature techniques. It should be noted that for the sample of size two, the integral corresponds to the positive half of Student's-t distribution and the relation

$$q_\alpha = \sqrt{2}\, t_\alpha$$

is thus obtained.

This work however is based upon the inaccurate work of Tippett (1925) and the tables constructed thus contain many errors. Recognizing the need for more accurate tables, Pearson and Hartley (1942) recalculated and tabulated a new set of tables for the range. This table expresses the probability that the range in a sample of n observations is less than a given multiple of the population standard deviation.

In the following year, this team of Pearson and Hartley (1943) used this new table of the range to derive tables of the Probability Integral of the Studentized Range. Denote the probability integral by ${}_fP_n(Q)$ where n is the size of the first sample from which the range is calculated and f denotes the degrees of freedom in the standard deviation estimated from a separate sample. The quantity ${}_fP_n(Q)$ thus represents the chance that the ratio $q = w/S$ does not exceed the limit Q.

As f approaches infinity and S^2 approaches σ^2, ${}_fP_n(Q)$ tends toward $P_n(R)$ of the ratio $R = w/\sigma$ taken at the point $R = Q$. It is possible then, by a Taylor's expansion, to represent ${}_fP_n(Q)$ as a quadratic in f^{-1}.

$${}_fP_n(Q) = P_n(Q) + \frac{1}{f}\, a_n(Q) + \frac{1}{f^2}\, b_n(Q)$$

Where

$$a_n(Q) = \frac{1}{4}\left\{Q^2 \frac{d^2P_n}{dR^2} - Q \frac{dP_n}{dR}\right\}$$

and

$$b_n(Q) = \frac{1}{16}\left\{\frac{Q^4}{2} \frac{d^4P_n}{dR^4} - \frac{Q^3}{3} \frac{d^3P_n}{dR^3} - \frac{Q^2}{2} \frac{d^2P_n}{dR^2} + Q \frac{dP_n}{dR}\right\}$$

With the aid of tables of $P_n(Q)$, $a_n(Q)$ and $b_n(Q)$ a table of the 5% and 1% points of the distribution of the studentized range were constructed by inverse interpolation.

May (1952) notes that this quadratic expansion breaks down for very small values of f and large values of q. As a result, corrections and extensions to the Pearson and Hartley tables were obtained by an evaluation of the probability integral at the upper tail.

$$\Pr\ (w/S \geq q) = C_f \int_0^{\infty} \left[4 \frac{w}{q} \not{Z}\left(\frac{w}{q}\right)\right]^f \frac{1}{W} p\ (w/n)\ d\ w$$

where

$$\not{Z}\left(\frac{W}{q}\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}\frac{w^2}{q^2}\right\}$$

and

$$C_f = \frac{2\left\{\frac{1}{4}\sqrt{\pi f}\right\}^f}{\Gamma\left(\frac{1}{2}f\right)}$$

and where p (w/n) is the upper tail probability integral of the range in a sample of size n. This integral was evaluated by numerical quadrature.

Pachares (1959) extended this method to include the upper 10% points and offered corrections to many entries in the 5% and 1% point tables presented by May.

The most recent and most accurate tables to date have been calculated by Harter (1960). Tables are given for the probability integrals of both the range and the studentized range.

III TABLES FOR DUNCAN'S NEW MULTIPLE RANGE TEST

As was pointed out previously, Duncan's New Multiple Range test requires the use of varying probability levels depending upon the number of means contained in a subset. For example, if a subset containing two means is to be tested at the probability level 0.95, a subset of three means is tested at the level $(0.95)^2$ or 0.9025, four means at level $(0.95)^3$ or 0.857375, twelve means at level $(0.95)^{11}$ or 0.568800 and similarly for n means where the level is $(0.95)^{n-1}$.

The ordinary tests such as Tukey's allowance procedure or the Student-Newman-Keuls procedure require that regardless of the subset size, all tests are carried out at the probability level 0.95. As a result, the tables previously discussed give values of the studentized range corresponding to the level 0.95 for every entry in the table. Duncan's test thus requires a special table with entries based on the appropriate probability level for a given subset size.

Beyer (1953) illustrates how the values in the tables for the New Multiple Range test are obtained using tables by Pearson and Hartley. As an example he uses the case in which the subset contains three means, as a result of which, a test on this subset is carried out at the probability level 0.9025. The example also uses eleven degrees of freedom in the independent estimate of the standard deviation. For given values of Q_1, Q_2, and Q_3 of the studentized range, probabilities P_1, P_2, and P_3 are calculated using the tables provided by Pearson and Hartley (1943) and the Taylor's expansion mentioned previously,

$$fP_n(Q) = P_n(Q) + \frac{1}{f} a_n(Q) + \frac{1}{f^2} b_n(Q)$$

The values P_1, P_2 and P_3 are found in such a way that P_1, and P_3 are above and below the probability level 0.9025 and P_2 is closer to 0.9025 than either P_1 or P_3.

Q		P_3 (Q)
3.50	$P_1 = 0.9644 + \frac{1}{11}(-0.39) + \frac{1}{121}(-0.4) = 0.9256$	11
3.25	$P_2 = 0.9439 + \frac{1}{11}(-0.46) + \frac{1}{121}(0.2) = 0.9004$	
3.00	$P_3 = 0.9145 + \frac{1}{11}(-0.51) + \frac{1}{121}(0.1) = 0.8689$	

Interpolation is then carried out to find the value of Q corresponding to the probability 0.9025. Using a second degree polynomial, the interpolated value is found to be 3.27.

A second case with $p = 2$ and $f = \infty$ is given illustrating the use of the Pearson and Hartley tables (1942). A similar procedure as above is carried out with the exception that the probability level is now 0.95. The values used are given as follows:

R	P_2 (R)
2.70	0.9438
2.75	0.9482
2.80	0.9523

As a result of interpolation using a second degree polynomial as above, the value of R corresponding to $P_n(R)$, is equal to 2.77.

Now consider a case, to be used later, in which the subset contains twelve means and the value of the studentized range is based on ninety-six degrees of freedom.

In applying the Student-Newman-Keuls procedure, the test on this subset is based on the probability level 0.95. As a result, interpolation for the value of the studentized range for this subset procedes as follows: Values of P_1, P_2 and P_3 are obtained such that P_2 is nearer 0.95 than either P_1 or P_3. Corresponding to these probabilities, values of the studentized range, Q_1, Q_2 and Q_3 are found and interpolation for the appropriate Q is carried out using a second degree polynomial.

Q	$_{96}P_{12}(Q)$
4.50	$P_1 = 0.9352 + \frac{1}{96}(-1.35) + \frac{1}{9216}(-0.5) = 0.921084$
4.75	$P_2 = 0.9624 + \frac{1}{96}(-1.05) + \frac{1}{9216}(-3.1) = 0.951127$
5.00	$P_3 = 0.9791 + \frac{1}{96}(-0.75) + \frac{1}{9216}(-5.2)$ 2 0.970724

Interpolation for $P = 0.95$ gives the value of Q as 4.74.

For this same subset, with the same degree of freedom, Duncan's New Multiple Range Test requires that the test be carried out at the probability level $(0.95)^{11}$ or 0.5688. The tables given by Pearson and Hartley (1943) are again used and the same procedure as above is employed for finding the value of Q corresponding to $P = 0.5688$.

Q	$_{96}P_{12}(Q)$
3.00	$P_1 = 0.3927 + \frac{1}{96}(-0.18) + \frac{1}{9216}(0.5) = 0.390879$
3.25	$P_2 = 0.5222 + \frac{1}{96}(-0.74) + \frac{1}{9216}(3.7) = 0.514893$
3.50	$P_3 = 0.6442 + \frac{1}{96}(-1.26) + \frac{1}{9216}(6.7) = 0.631801$

The value of Q found by interpolation is 3.36.

Beyer (1953) has calculated complete tables for Duncan's test, basing his work on the procedures set forward by Pearson and Hartley (1943). In order to gain accuracy, he extended the Taylor's expansion of $_fP_n(Q)$ to include the term in f^{-3}. Checking his results against the tables prepared by May (1952), a high degree of accuracy was noted.

IV TABLES FOR DUNNETT'S PROCEDURE

Consider the case in which one treatment is compared with the control. Confidence limits may be placed about the true treatment mean difference basing the procedure on the well known Student's-t distribution. The limits would be found to be:

$$\bar{x}_1 - \bar{x}_o \pm d'' \; S\sqrt{\frac{1}{N_o} + \frac{1}{N_1}}$$

where d'' is chosen such that,

$$\text{Prob} \quad (\,|t| < d'') = P$$

The constant d'' or d' in the case of a one sided confidence limit, can be found from the tables of the distribution of Student's-t.

In generalizing this procedure, where each of p treatments is compared with the control, it is desired that separate confidence limits for each of the p differences be found such that the joint confidence coefficient is equal to a preassigned value P, $(0 < P < 1)$. For the two sided confidence limits, the interval will be found to be of the form,

$$\bar{x}_i - \bar{x}_o \pm d_i'' \; S\sqrt{\frac{1}{N_i} + \frac{1}{N_o}} \qquad i = 1, 2, \ldots\ldots, p$$

In order for these limits to have the desired preassigned joint confidence coefficient P, the p constants d_i'' are found such that,

$$\text{Prob} \quad (\,|t_1| < d_1'', \; |t_2| < d_2'', \ldots\ldots, \; |t_p| < d_p'') = P$$

This requires that the joint distribution of the t_i's be known. Dunnett and Sobel (1954) have found that this distribution is the multivariate analogue of Student's-t distribution. Tables are given for both the one-sided and the two-sided confidence interval approaches.

CHAPTER IV

COMPARISON OF METHODS

In any attempt to evaluate the tests available to an experimenter, for locating significant differences following a significant analyses of variance F-test, several factors must be kept in mind. The design of the experiment, the object of the experiment, and the philosophy of the experimenter will each influence the decision as to which test procedure is to be used.

The present chapter considers several items which are of importance in the selection of a test to be used. These items may be listed as follows:

a. Treatments versus control
b. Analysis of all possible contrasts
c. Confidence limits and tests of significance
d. A priori and a posteriori comparisons
e. the effect of a prior F-test

A series of examples are given which serve to illustrate many of the points raised throughout this study.

I TREATMENTS VERSUS CONTROL

In the design of an experiment, one of the treatments included may be a control treatment. Its purpose is to permit comparisons to be made between a given treatment and this control treatment such that a significance statement may be made regarding the effect of the treatment in relation to the control. In this situation, it may not be that all possible comparisons among treatment means is desired, but only comparisons between the control and each of the other treatments.

In order to achieve the desired results, use may be made of certain of the procedures outlined previously. Tukey's allowance procedure or Scheffe's procedures will give

significance tests or confidence interval statements about these specific treatment mean differences. Dunnett (1953), found that these two procedures, since they were designed for a much more general application, give confidence intervals which are too wide. By considering the control versus treatment situation only, Dunnett's test procedure is able to construct confidence intervals such that the probability that all the confidence intervals of this type in the experiment are simultaneously correct is equal to some preassigned specific probability level. The confidence limits so constructed are placed about the true mean difference between a treatment and the control.

Dunnett's procedure is designed for making a specific type of comparison. Its use in other more general situations, is never justified.

II ANALYSIS OF ALL POSSIBLE CONTRASTS

In many experimental situations, the most meaningful procedure may not be the comparisons among all possible pairs of treatment means. By the very nature of the treatments themselves, the logical procedure may be to divide the treatments into groups and compare the treatment means on this grouped basis. The decision as to whether to consider the problem as one involving all possible contrast comparisons will rest with the experimenter. Once the decision has been made however, the testing of the treatment means becomes the responsibility of the statistician.

The methods outlined by Duncan (1955), Tukey (1953) and Keuls (1952) are applicable only for the case of all possible paired comparisons and should be used only in that situation. Scheffé's procedure has been designed principally with the contrast problem in mind.

The use of Scheffé's method has been shown to be more sensitive than Tukey's allowance test for the contrast situation. It is noted particularly that the critical value in Scheffé's test incorporates the standard error of the particular contrast involved. By virtue of the fact that this standard error will in general be smaller than the standard error of the difference between two treatment means, confidence limits constructed by Scheffé's method will tend to result in shorter intervals than those found by using Tukey's allowance procedure.

The situation is reversed however, when considering the all possible paired comparison approach. Scheffé indicates this fact and cautions that in considering treatment means two at a time, Tukey's allowance procedure is preferred.

When considering the topic of contrasts, it is important to note the difference in procedure involved in orthogonal and non-orthogonal sets of contrasts. It was mentioned, that it is possible to obtain an orthogonal set of contrasts containing as many contrasts as there are degrees of freedom in the sum of squares for treatments in the analysis of variance. While it is possible to obtain a great many orthogonal sets in any one experiment, the set of contrasts to be used is generally prescribed when the experiment is at the designing stage. This being the case, each separate contrast may be tested against the experimental error by the use of the F-statistic.

On the other hand, if the contrasts are not orthogonal as would be the case in considering all possible contrasts, the testing would have to be carried out by the use of Scheffé's technique. It is difficult to see where a situation in which all possible contrasts were to be considered might arise. It seems much more probable that the experimenter, from past experience or because of a particular aim in mind, would be able to instruct the statistician into designing an experiment capable of testing his material without having to resort to a procedure such as that proposed by Scheffé.

III CONFIDENCE LIMITS AND TESTS OF SIGNIFICANCE

While the emphasis here has been towards the significance test, where the observed treatment difference is tested to determine whether the difference is significantly different from zero, much is to be said for a confidence interval approach. Procedures such as Tukey's allowance, Scheffé's test and Dunnett's, can easily be used to set 100 $(1 - \alpha)\%$ confidence limits on the difference between the population values of the treatment means.

A confidence interval statement is, in reality, a test of an infinite number of hypothesis. The limits stated give the upper and lower bounds beyond which a value is said to differ significantly, at a given probability level, from the true value. If, in setting confidence limits about an observed difference between treatment means, it is found that the limits do not enclose zero, the treatment means involved are said to differ significantly. This, of course, is exactly the same information that could be obtained from a test of significance of the difference between these means. The advantage in using the confidence limits is that a reasonable assurance is given that the actual value of the true treatment mean difference lies between these stated limits.

For the two mean case, significance test statements and confidence limit statements are very similar. It is when the multiple confidence limits are applied to the all possible comparisons procedure that difficulty is encountered in the interpretation of the confidence limit results. Ryan (1959) has noted the possibility of inconsistencies or contradictions in results obtained from applying both significance tests and confidence limits to the same experimental situation. In particular he points out that it may happen that limits for a difference between treatment means contain zero whereas, using some significance test procedure, this difference may be declared significantly different from zero. He explains that the procedures applicable to the confidence limit approach employ a single "allowance" for all treatment differences in

an experiment regardless of the size of the subset in which the two means occur, however, the more powerful significance procedures use varying criteria depending upon how many means there are the subset.

To criticize the confidence limit approach on this basis would seem to be very dangerous. Involved in this situation is a confidence limit, obtained perhaps using Tukey's allowance procedure, and a significance test, using the Student-Newman-Keuls test. The criticism should not be directed at the confidence limit approach in general, but at the two different procedures which yielded the contradictory results. Had the confidence limit been replaced by the significance test using the same Tukey allowance procedure, the contradiction would remain. Since different methods have been used, contradictions of this type are to be expected.

To the experimenter, interested more in a yes or no reply to specific questions, the test of significance procedure is undoubtedly the more appealing. The statistician, however, being interested in obtaining as much information from the data as is possible, may well wish to adopt the confidence limit approach.

IV A PRIORI AND A POSTERIORI COMPARISONS

In the designing of experiments, it is common practice to specify exactly what tests are to be carried out following the application of the design to an experimental problem. It is widely held, that this specification of test a priori justifies the use of classical methods. At the present time, the situation with regard to the effect of a priori specification of tests versus the tests suggested by the data a posteriori can not be dogmatically exclaimed. The decision between the use of a classical method such as the LSD procedure and a newer technique such as Duncan's New Multiple Range test depends, very much, on the nature of the specified a priori tests.

It can be stated emphatically, that the LSD procedure should not be used for making all possible paired comparisons between treatment means. The effect of stating, before the experiment has taken place, that all possible comparisons will be made does not alter the fact that these comparisons are not independent and the LSD is not valid.

If on the other hand, a pair of treatments are selected before the data are collected, a different situation is encountered. This case is conceptually identical to the selection of a pair of treatments at random from the set of all possible paired treatments. This randomness validates the use of an ordinary t-test or the LSD procedure for making the test of significance. Similarly, if several pairs of treatments are selected a priori such that the pairs are independent of one another, the LSD may be applied. This latter condition also applies to contrasts of several means and its occurence indicates the use of standard techniques.

After having obtained the data, should it be decided to test the largest mean against the smallest mean, a far different situation arises. The test of this particular pair involves a much more complex result. Should this difference be declared significant, more is being said about the experimental results than in a test of a random pair. In fact, the statement of significance is a claim that there exists one or more significant differences among all pairs of means in the subset containing the largest and the smallest means. As a result of this, the probability that the difference between the largest and smallest mean is significant is greater using the classical procedure than is desired.

The main problem in the question of a priori and a posteriori specified tests is undoubtedly the independence factor. Independence appears to be of less importance to the newer procedures designed for multiple comparisons. The LSD, however, requires that the mean differences being tested be

independent and as a result, if, as is very frequently the case, the tests, outlined in the design of the experiment, are independent, they may be analysed by this classical procedure. Other non-independent situations, whether specified a priori or a posteriori, must employ some other procedure.

V THE EFFECT OF A PRIOR F-TEST

After obtaining the data from an experiment, the usual procedure is to carry out the analysis of variance. A test involving the ratio of the mean square for treatments, and the mean square for error is carried out usually at the 0.95 probability level. The effect of this F-test is to tell the experimenter whether or not the observed treatment means may be considered to have come from a common population with respect to the population mean. Should this F-ratio be declared significant, the experimenter can state, with probability 0.95, that at least one of the means differs from the rest. This is a completely standard and well accepted criterion upon which to base statistical conclusions.

Of the multiple comparison procedures considered, all, with the exception of Duncan's New Multiple Range test, either incorporate this F-statistic as the first step or have included a step consistent with it. The F-ratio can be considered as a type of test of the largest mean against the smallest mean since, as was stated previously, this pair of means involves, inferencewise, all the pairs of means lying between them. Since the F-test statistic may be applied at any desired probability level, P, a range test of the largest mean with the smallest mean which is also carried out at the probability level P, will carry with it, much the same information.

Duncan (1955) has constructed a test criterion which tests the largest mean against the smallest mean at a probability level much lower than that of the F-statistic. The New Multiple Range Test may be applied regardless of whether an analysis of variance F-test is used or not. If one

accepts the view that the F-ratio is insensitive to detecting differences among a set of treatment means, Duncan's test procedure will provide an appealing alternative to the analysis of variance F-test. The more common view, however, is that the F-statistic gives a well founded and a sensitive test of the actual situation. Duncan's procedure, by changing the probability level of the F-statistic appears to unjustly run the risk of finding significant differences when in fact none exist.

The other multiple comparison techniques may alter the probability level P, of the F-statistic but by raising the level rather than lowering it, they cause the entire procedure to remain consistent with standard procedure. The effect of the change is to create a test which is more conservative than may be desired. Only a close look at the consequences of the increase in sensitivity, found in Duncan's procedure, and the lowered risk of declaring too many significant differences with one of the other techniques, will influence the choice of which test to use.

VI APPLICATION OF METHODS

In order to illustrate the various procedures and to show many of the points mentioned previously, a series of examples is presented. The experiment concerns a test of twenty-five varieties of wheat arranged in a balanced lattice design with six replications. The series of four experiments are related to one another in that the same varieties and the same type of design are used in each case. The experiments are members of a series in the Co-operative Wheat Test program carried out at various experimental stations in Canada in the year 1960 by the Canada Department of Agriculture. Differences between the examples are primarily attributed to location and other geographic factors.

The analysis proceeds in much the same way as the example cited in Chapter I. Sums of squares are calculated for the replications, the blocks and the varieties. In this particular design, the replicates, blocks, and varieties carry five, twenty-four and twenty-four degrees of freedom respectively. The mean square values are obtained simply by dividing each sum of squares by its appropriate degrees of freedom.

Since an F-test will be desired to test whether or not all the variety means come from a common population, it is necessary to examine the variety sum of squares very carefully. The complication here is that the varieties occur in different blocks and as a result the effect of the differences between blocks has not been fully eliminated. It is thus necessary to adjust the variety sum of squares for the block effect. The mean square for varieties corrected, is divided by the mean square for error and the resultant F-statistic is compared with the tabular value obtained by entering the tables of the distribution of F with twenty-four degrees of freedom in the numerator and ninety-six degrees of freedom in the denominator for the desired probability level.

The standard error of a corrected variety mean is the value obtained by incorporating the correction for block effects. Since the analysis is carried out using units of measurements in grams, a conversion factor = 0.0647 is necessary to convert means and standard errors into bushels per acre.

Tables IX, X, XI, and XII summarize the results for the four experimental trials considered. Each table gives the location where the experiment was carried out and lists:

i. the analysis of variance and F-test
ii. the standard error of a corrected variety mean
iii. the ranked variety means and multiple comparison results.

In the analysis of variance for each location, the F-statistic is of particular interest. The tabular values of the F-ratio with twenty-four degrees of freedom in the numerator and ninety-six degrees of freedom in the denominator at the 5%, 1% and 0.5% levels of significance are respectively 1.64, 2.00, and 2.44. Comparing the F-ratio in each case with the critical values gives an indication of the numbers of variety mean differences which may be expected to be declared significant by the various multiple comparison procedures.

Table IV lists the critical range factors for the 5% level tests of twenty-five means with ninety-six degrees of freedom in the standard error. These values given for eight different techniques are the factors which, when multiplied by the appropriate standard error, give the critical values necessary for carrying out the significance test. The values in Tables V, VI, VII and VIII are obtained as the product of the standard error of the corrected variety mean for a particular locality and the values in Table IV.

The results of applying the techniques to the ranked variety means for a given locality are presented in a graphical form in Tables IX, X, XI, and XII. It should be noted that the ranked variety means are lettered from A to Y with the highest ranked mean being denoted by A and the lowest mean in rank by Y. As a result of this type of lettering, it may be that a given variety will carry different letters in each example depending upon its rank in relation to the remaining means.

The lines joining the various letters may be interpreted very simply. Any two ranked means underlined with a common line are said not to differ significantly. Any two means which are not underscored by the same line are said to be significantly different.

It is interesting to note the F-statistic from the analysis of variance in relation to the results of the eight techniques applied to the variety means. The data from Foremost, Alberta (experiment 1.) have given rise to a non-significant F value which is interpreted as meaning that there are no differences among the variety means of this experiment. The application of the multiple comparison procedures should be expected to discover no significant differences in this case. In fact, the LSD and Duncan's New Multiple Range test find a sizeable number of differences in direct contradiction to the conclusion reached by considering the analysis of variance F-test. The remainder of the test procedures do not find any differences.

The Evansburg, Alberta experiment (experiment 2.) has yielded an F value which is highly significant but less than the critical value at the 0.5% probability level. To be consistent with this F-statistic result, which indicates that there are significant differences among the variety means, it is expected that the test procedures will detect a few differences. Once again, Duncan's test and the LSD as the only procedures which show any significant results at all.

Advancing a further step in significance, the Regina, Saskatchewan experiment (experiment 3.) has given an F-statistic which is highly significant and exceeds the 0.5% probability level critical value. The interpretation of this significance is that there exists a greater number of significant differences than in the previous experiment among the ranked variety means. As is expected, all test procedures, with the exception of Scheffe's test and the modified Scheffe's procedure, detect a sizeable number of significant differences.

The last example in this series, Morden, Manitoba, (experiment 4.) illustrates the application of the multiple comparison technique to data which have displayed an F-statistic with an extremely large value. The size of the F value is such that a large number of significant differences is expected. With the exception of Scheffe's method, each test procedure has detected many significant differences.

TABLE IV

Comparison of Critical Range Factors for 5% Level Tests of 25 Means with f = 96

TEST	Subset Sizes = p																							
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
L. S. D.	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81	2.81
Fisher's	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44	5.44
Tukey's Allowance	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32
Tukey's 1953	4.07	4.35	4.51	4.63	4.72	4.79	4.85	4.91	4.95	4.99	5.03	5.06	5.09	5.12	5.15	5.17	5.20	5.22	5.24	5.25	5.27	5.29	5.31	5.32
Students-Newman-Keuls	2.81	3.37	3.70	3.94	4.12	4.26	4.38	4.50	4.58	4.66	4.74	4.80	4.86	4.92	4.97	5.02	5.07	5.11	5.15	5.18	5.22	5.25	5.29	5.32
Duncan's New Multiple Range	2.81	2.95	3.05	3.12	3.18	3.22	3.26	3.29	3.32	3.34	3.36	3.38	3.39	3.41	3.42	3.44	3.45	3.46	3.47	3.48	3.49	3.50	3.50	3.51
Scheffé's	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85	8.85
Modified Scheffé's	2.81	3.51	4.02	4.44	4.80	5.13	5.42	5.70	5.95	6.20	6.43	6.66	6.88	7.08	7.29	7.48	7.67	7.84	8.01	8.20	8.37	8.54	8.71	8.85

TABLE V

Co-operative Wheat Test (Foremost, Alberta-1960) Experiment 1.

Critical Range Values for 5% Level Test of 25 Means with f = 96 and $S_{\bar{x}}$ = 1.23 Bushels

TEST	Subset Sizes = p																							
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
L. S. D.	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46
Fisher's	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Tukey's Allowance	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32	5.32
Tukey's 1953	5.01	5.35	5.55	5.69	5.81	5.89	5.97	6.04	6.09	6.14	6.19	6.22	6.26	6.30	6.33	6.36	6.40	6.42	6.45	6.46	6.48	6.51	6.53	6.54
Student-Newman-Keuls	3.46	4.15	4.55	4.84	5.07	5.24	5.39	5.54	5.63	5.73	5.83	5.90	5.98	6.05	6.11	6.17	6.24	6.29	6.33	6.37	6.42	6.46	6.51	6.54
Duncan's New Multiple Range	3.46	3.63	3.75	3.84	3.91	3.96	4.01	4.05	4.08	4.11	4.13	4.16	4.17	4.19	4.21	4.23	4.24	4.26	4.27	4.28	4.29	4.30	4.31	4.32
Scheffé's	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89	10.89
Modified Scheffé's	3.46	4.32	4.94	5.46	5.90	6.31	6.67	7.01	7.32	7.63	7.91	8.19	8.46	8.71	8.97	9.20	9.43	9.64	9.85	10.09	10.30	10.50	10.71	10.89

TABLE VI

Co-operative Wheat Test (Evansburg, Alberta-1960) Experiment 2.

Critical Range Values for 5% Level Tests of 25 Means with f = 96 and $S_{\bar{x}}$ = 1.62 Bushels

TEST	Subset Sizes = p																							
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
L. S. D.	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55
Fisher's	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81	8.81
Tukey's Allowance	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62	8.62
Tukey's 1953	6.59	7.05	7.31	7.50	7.65	7.76	7.86	7.95	8.02	8.08	8.15	8.20	8.25	8.29	8.34	8.38	8.42	8.46	8.49	8.51	8.54	8.57	8.60	8.62
Student-Newman-Keuls	4.55	5.46	5.99	6.38	6.67	6.90	7.10	7.29	7.42	7.55	7.68	7.78	7.87	7.97	8.05	8.13	8.21	8.28	8.34	8.39	8.46	8.51	8.57	8.62
Duncan's New Multiple Range	4.55	4.78	4.94	5.05	5.15	5.22	5.28	5.33	5.38	5.41	5.44	5.48	5.49	5.52	5.54	5.57	5.59	5.61	5.62	5.64	5.65	5.67	5.68	5.69
Scheffé's	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34	14.34
Modified Scheffé's	4.55	5.69	6.51	7.19	7.78	8.31	8.78	9.23	9.64	10.04	10.42	10.79	11.15	11.47	11.81	12.12	12.43	12.70	12.98	13.28	13.56	13.83	14.11	14.34

TABLE VII

Co-operative Wheat Test (Regina, Saskatchewan-1960) Experiment 3.

Critical Range Values for 5% Level Test of 25 Means with f = 96 and $S_{\bar{x}}$ = 1.10 Bushels

TEST	Subset Sizes = p																							
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
L. S. D.	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09
Fisher's	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98
Tukey's Allowance	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85	5.85
Tukey's 1953	4.48	4.79	4.96	5.09	5.19	5.27	5.34	5.40	5.45	5.49	5.53	5.57	5.60	5.63	5.67	5.69	5.72	5.74	5.76	5.78	5.80	5.82	5.84	5.85
Student-Newman-Keuls	3.09	3.71	4.07	4.33	4.53	4.69	4.82	4.95	5.04	5.13	5.21	5.28	5.35	5.41	5.47	5.52	5.58	5.62	5.67	5.70	5.74	5.78	5.82	5.85
Duncan's New Multiple Range	3.09	3.25	3.36	3.43	3.50	3.54	3.59	3.62	3.65	3.67	3.70	3.72	3.73	3.75	3.76	3.78	3.80	3.81	3.82	3.83	3.84	3.85	3.85	3.86
Scheffé's	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74	9.74
Modified Scheffé's	3.09	3.86	4.42	4.88	5.28	5.64	5.96	6.27	6.55	6.82	7.07	7.33	7.57	7.79	8.02	8.23	8.44	8.62	8.81	9.02	9.21	9.39	9.58	9.74

TABLE VIII

Co-operative Wheat Test (Morden, Manitoba-1960) Experiment 4.

Critical Range Values for 5% Level Tests of 25 Means with f = 96 and $S_{\bar{x}}$ = 1.16 Bushels

TEST	Subset Sizes = p																							
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
L. S. D.	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26	3.26
Fisher's	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31	6.31
Tukey's Allowance	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17	6.17
Tukey's 1953	4.72	5.05	5.23	5.37	5.48	5.56	5.63	5.70	5.74	5.79	5.83	5.87	5.90	5.94	5.97	6.00	6.03	6.06	6.08	6.09	6.11	6.14	6.16	6.17
Student-Newman-Keuls	3.26	3.91	4.29	4.57	4.78	4.94	5.08	5.22	5.31	5.41	5.50	5.57	5.64	5.71	5.77	5.82	5.88	5.93	5.97	6.01	6.06	6.09	6.14	6.17
Duncan's New Multiple Range	3.26	3.42	3.54	3.62	3.69	3.74	3.78	3.82	3.85	3.87	3.90	3.92	3.93	3.96	3.97	3.99	4.00	4.01	4.03	4.04	4.05	4.06	4.06	4.07
Scheffé's	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27	10.27
Modified Scheffé's	3.26	4.07	4.66	5.15	5.57	5.95	6.29	6.61	6.90	7.19	7.46	7.73	7.98	8.21	8.46	8.68	8.90	9.09	9.29	9.51	9.71	9.91	10.10	10.27

TABLE IX

Co-operative Wheat Test (Foremost, Alberta-1960) Experiment 1. Results

i. Analysis of Variance and F-test

Source	d.f.	Sum of Squares	Mean Square	F
Replicates	5	330281	66056	
Blocks	24	253334	10556	
Varieties	24	139038	5793	
(corrected)	24	75952	3165	1.60 (not significant)
Error	96	190019	1979	

ii. Standard Error of a Corrected Variety Mean $S_{\bar{x}} = 1.23$ Bushels

iii. Variety Means Ranked in Order and Test Results

	18.4	18.4	18.2	17.9	17.8	17.6	17.4	17.3	17.0	16.7	16.7	16.6	16.3	15.8	15.7	15.5	15.5	15.4	14.9	14.6	14.5	14.4	14.4	14.1	12.0
TESTS	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
L. S. D.																									
Fisher's																									
Tukey's Allowance																									
Tukey's 1953																									
Student-Newman-Keuls																									
Duncan's New Multiple Range																									
Scheffé's																									
Modified Scheffé's																									

Table X

Co-operative Wheat Test (Evansburg, Alberta-1960) Experiment 2. Results

i. Analysis of Variance and F-test

Source	d.f.	Sum of Squares	Mean Square	F
Replicates	5	88,584	17,917	
Blocks	24	189,828	7,910	
Varieties	24	196,720	8,197	
(corrected)	24	165,570	6,898.7	2.12
Error	96	312,733	3,258	

ii. Standard Error of a Corrected Variety Mean $S_{\bar{x}} = 1.62$ Bushels

iii. Variety Means Ranked in Order and Test Results

	50.7	49.6	49.4	48.7	48.7	48.2	48.2	47.9	47.8	47.7	47.7	47.1	47.0	46.8	46.7	46.5	46.2	46.0	44.6	44.6	44.5	43.0	42.9	42.3	42.3
TEST	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
L. S. D.																									
Fisher's																									
Tukey's Allowance																									
Tukey's 1953																									
Student-Newman-Keuls																									
Duncan's New Multiple Range																									
Scheffé's																									
Modified Scheffé's																									

TABLE XI

Co-operative Wheat Test (Regina, Saskatchewan-1960) Experiment 3. Results

i. Analysis of Variance and F-test

Source	d.f.	Sum of Squares	Mean Square	F
Replicates	5	73,904	14,781	
Blocks	24	73,089	3,045	
Varieties	24	176,775	7,366	
(corrected)	24	146,423	6,101	3.65
Error	96	160,335	1,670	

ii. Standard Error of a Corrected Variety Mean $S_{\bar{x}} = 1.10$ Bushels

iii. Variety Means Ranked in Order and Test Results.

	46.0	43.6	43.1	42.8	42.5	42.2	42.0	42.0	41.5	41.5	41.3	41.2	40.9	40.3	40.2	40.2	40.0	39.9	39.5	39.3	38.5	38.2	37.6	37.4	36.3
TEST	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
L. S. D.																									
Fisher's																									
Tukey's Allowance																									
Tukey's 1953																									
Student-Newman-Keuls																									
Duncan's New Multiple Range																									
Scheffé's																									
Modified Scheffé's																									

TABLE XII

Co-operative Wheat Test (Morden, Manitoba-1960) Experiment 4. Results

i. Analysis of Variance and F-test

Source	d.f.	Sum of Squares	Mean Square	F
Replicates	5	17,065	3,413	
Blocks	24	55,460	2,311	
Varieties	24	617,610	25,734	
(corrected)	24	590,378	24,599	14.29
Error	96	177,726	1,851	

ii. Standard Error of a Corrected Variety Mean $S_{\bar{x}} = 1.16$ Bushels

iii. Variety Means Ranked in Order and Test Results

	55.5	52.2	51.9	50.3	50.0	49.6	49.4	48.6	48.2	48.0	47.5	47.4	46.8	45.5	45.3	45.0	45.0	44.7	44.2	44.1	43.8	43.0	41.8	39.2	35.7
TEST	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
L. S. D.																									
Fisher's																									
Tukey's Allowance																									
Tukey's 1953																									
Student-Newman-Keuls																									
Duncan's New Multiple Range																									
Scheffé's																									
Modified Scheffé's																									

Table XIII gives, for each test procedure and for the four experiments, the number of significant differences detected. In an experiment with twenty-five means, the maximum number of significant differences it is possible to detect is 300.

TABLE XIII

THE NUMBER OF SIGNIFICANT DIFFERENCES PER TEST DETECTED IN THE EXPERIMENTAL SERIES

TEST	EXPERIMENTS			
	1	2	3	4
LSD	40	58	126	176
Fisher's	0	0	15	82
Tukey's Allowance	0	0	16	86
Tukey's 1953	0	0	19	95
Student-Newman-Keuls	0	0	21	107
Duncan's New Multiple Range	13	29	71	156
Scheffé's	0	0	0	29
Modified Scheffé's	0	0	0	59

From these results several items of importance can be realized. As was explained earlier, Duncan's New Multiple Range Test, by the lowering of the probability level of the F-test, runs the risk of finding too many significant differences. The examples illustrate this point clearly since Table XIII shows that Duncan's test ranks second only to the LSD in total numbers of significant differences found in each experiment. The striking feature here is that the LSD and Duncan's test were able to find some differences even in the case in which the F-statistic declared there were none.

At the other extreme, Scheffé's test and the Modified Scheffé's test are shown to be very insensitive. It is not until the F-statistic reaches an unusually large value that any significant differences are detected. Even in this case, however, only twenty-nine were located with Scheffé's procedure and fifty-nine with the Modified Scheffé technique. These values are very low in comparison with other procedures. The Modified Scheffé technique, which seemed a promising one, is still hampered with a very high critical value for the largest subset of twenty-five means. The sensitivity has improved but its showing in these examples does not recommend its widespread use.

The failure of the Student-Newman-Keuls procedure to detect any differences in experiment 2 is somewhat surprising. The size of the F-statistic indicates that differences should be discovered. Its failure in this situation, however, serves to illustrate that this procedure as well as Fisher's Modified technique and Tukey's allowance and 1953 procedures tend to be conservative.

In an examination of Table XIII, it is noted that no two tests give exactly the same results in any of the four experiments.

CHAPTER V

CONCLUSIONS

It is now possible to offer to the practising statistician some guidance as to which procedure should be recommended for a given experimental problem. It is not possible to state emphatically that in any one situation there is one and only one test procedure which must be used. Some of the techniques studied have been designed to analyse a particular type of problem. These special techniques may not be the best for analysing the situation for which they were intended however their use in any other, more general situation, is not justified.

If it is desired to test each treatment against the control treatment, Dunnett's procedure should be the method applied and Dunnett's procedure should be used only in this situation.

If orthogonal contrasts are designed into the experiment prior to the collection of the data, the F-ratio may be used for the desired tests of significance. The LSD may be applied to tests involving independent pairs of means specified a priori. If, however, the contrasts are supplied after the data has been observed, or if the contrasts are of the all possible contrast type, the LSD method is not valid and Scheffé's method is preferable.

Duncan's New Multiple Range test and the LSD procedure are undoubtedly the most powerful procedures. The LSD, however, should never be used except when independent comparisons are designed into the experiment. Duncan's test appears to be too sensitive and the risk of declaring too many significant differences is great.

If it is desired to use a confidence limits procedure, the Tukey allowance method is recommended.

If significance test of all possible comparisons are desired, Tukey's 1953 procedure and the Student-Newman-Keuls procedure give very similar results. However, since the critical values are easier to obtain in the latter, the Student-Newman-Keuls procedure is recommended.

At the present time, there is no way, short of a long and detailed sampling study, of obtaining an empirical comparison of all the methods presented here. A large volume of work has been done on the concept of error rates as applied to these methods. Unfortunately any attempt at comparing procedures by the use of this concept, runs into the problem that statisticians can not agree as to which error rate is most important. Until some procedure is outlined which is capable of applying criteria to these tests which are acceptable to the large body of statisticians, statistical philosophy and personal experience will remain the principal guide lines in the choice of which technique is applied.

BIBLIOGRAPHY

Beyer, William H., (1953) "Certain Percentage Points of the Distribution of the Studentized Range of Large Samples." Unpublished Master's thesis, Virginia Polytechnic Institute, 55 pp.

Duncan, D. B., (1955) "Multiple Range and Multiple F Tests," Biometrics, 11: 1-42, March.

Dunnett, C. W., (1955) "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," Journal of the American Statistical Association, 50: 1096-1121, December.

Dunnett, C. W., and Sobel, M., (1955) "Approximations to the Probability Integral and Certain Percentage Points of a Multivariate Analogue of Student's t-distribution," Biometrika, 42: 258-60.

Federer, Walter T., (1955) Experimental Design. New York: The Macmillan, 544 pp.

Fisher, R. A. (1960) The Design of Experiments. New York: Hafner Publishing Company, Inc., 7th Edition, 245 pp.

Harter, H. L., (1957) "Error Rates and Sample Sizes for Range Tests in Multiple Comparisons," Biometrics, 13: 511-536, December.

_____, (1960) "Tables of Range and Studentized Range," Annals of Mathematical Statistics, 31: 1122-1147.

Hartley, H. O., (1955) "Some Recent Developments in Analysis of Variance," Communications on Pure and Applied Mathematics, 8: 47-72, February.

Hoel, Paul G., (1954) Introduction to Mathematical Statistics. New York: John Wiley and Sons, Inc., 2nd Edition, 331 pp.

Keuls, M., (1952) "The use of the 'Studentized Range' in Connection with an Analysis of Variance," Euphytica 1: 112-121.

May, J. M., (1952) "Extended and Corrected Tables of the Upper Percentage Points of the Studentized Range," Biometrika, 39: 192-193.

McKay, A. T. (1933) and Pearson, E. S., "A Note on the Distribution of Range in Samples of n," Biometrika, 25: 415-420.

Newman, D., (1939) "The Distribution of the Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation," Biometrika, 31: 20-30.

Pachares, J., (1959) "Table of the Upper 10% points of the Studentized Range," Biometrika, 46: 461-463.

Pearson, E. S., and Hartley, H. O., (1942) "The Probability Integral of the Range in Samples of n Observations from a Normal Population," Biometrika, 32: 301-308.

_____ (1943) "Tables of the Probability Integral of the Studentized Range," Biometrika, 33: 89-99.

Ryan, T. A., (1959) "Multiple Comparisons in Psychological Research," Psychological Bulletin, 56: 26-47

Scheffé, H., (1953) "A Method for Judging all Contrasts in an Analysis of Variance," Biometrika, 40: 87-104.

Steel, R. G. D., and Torrie, J. H., (1960) Principles and Procedures of Statistics. New York: McGraw-Hill Book Company, Inc., 481 pp.

Student, (1927) "Errors of Routine Analysis," Brometrika, 19: 151-164.

Tippett, L. H. C., (1925) "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," Biometrika 17: 364-387.

Tukey, J. W., (1952) "Allowances for Various Types of Error Rates," Unpublished invited address presented at Blacksburg meeting of Institute of Mathematical Statistics.

_____, (1953) "The Problem of Multiple Comparisons," Ditto, Princeton University, Princeton, New Jersey.

Wine, R. L., (1955) "A Power Study of Multiple Range and Multiple F-Tests," Virginia Polytechnic Institute Technical Report No. 12.