PROPERTIES OF MULTIPLE COMPARISON TEST PROCEDURES

by

DONG HYUN KIM

B. S., Washburn University, Topeka, Kansas, 1960

A MASTER'S REPORT

submitted in partial fulfillment of the

requirement for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

TABLE OF CONTENTS

INTRODUCTION

In an experiment which is designed to compare specific treatments, varieties, or methods, one may wish to decide which population means are actually equal, and which are unequal. In the various fields of agriculture, industry, and physical or biological science, one often is confronted with such problems; therefore it becomes important to have test procedures which have certain desirable properties. It often is also desirable to know how convenient it is to use the suggested test procedures.

In the early 1920's the British statistician R. A. Fisher introduced a statistical technique for analyzing experiments called the analysis of variance. In general, the analysis of variance compares possible sources of variation in the experimental units with an appropriate measure of random sampling error, and leads to decisions to accept or to reject appropriate statistical hypotheses regarding population parameters. The F-test commonly is used to make these decisions. In other words, this is a statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to decide which kinds of effects are important, and to estimate these effects.

Before an analysis of variance can be conducted, it is essential to know the properties of the observed values, i.e., the nature and distributions of the observed values. If r observations $x_1, x_2, \ldots, x_r$, assumed to be on r random variables, are linear combinations of m unknown effects $\alpha_1, \alpha_2, \ldots, \alpha_m$ plus errors, $\epsilon_1, \epsilon_2, \ldots, \epsilon_r$, one usually can express x in the form of a mathematical model such as:

$$x_i = \alpha_1 + \alpha_2 + \ldots + \alpha_m + \epsilon_i \quad (i = 1, \ldots, r)$$

The α's are more or less idealized formulations of some aspects or the observations which underly the phenomena of interest to the investigators. If the $\{\alpha_j\}$ measure the unknown effects in some describable way they can be defined as parameters. A model is called a fixed-effects model if all these α_j are unknown constants. A model in which all α_j are random variables, except one which is used to represent the "general mean" the model is called a random-effects model. One other model called a mixed-effects model is a combination of the fixed and random models in which at least one effect is random and at least one is a fixed effect. 'Fixed effects' means that the treatments involved are the only treatments in the experiment in which investigators are interested in the amount of effects. 'Random effects' means the treatments are chosen randomly or systematically from a large group of treatments and the investigator is interested not only in those particular treatments but also in the whole group of treatments.

A multiple comparison test is applied to fixed-effects because it only concerns testing differences among various treatment means involved in the experiment; and this implies fixed effects with their corresponding means.

A one way classification for a fixed-effects model refers to the effects of inequalities among the true means of several (univariate) treatments, $\mu_1, \mu_2, \ldots, \mu_k$. It is assumed that the k treatments have a common variance, σ^2, and that independent random samples of equal size r were taken from the k populations. This can be expressed in the mathematical model:

$$x_{ij} = \mu_i + \epsilon_{ij} \quad (i = 1, \ldots, k;\ j = 1, \ldots r)$$

$$= \mu + (\mu_i - \mu) + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$$

ϵ_{ij} are NID(0, σ^2),

where

μ = population mean

$\tau_i = (\mu_i - \mu)$ = fixed effect for the ith treatment.

The null hypothesis which the F-test rejects or accepts at a stated level of significance ($= \alpha$) is

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_k$$

or equivalently

$$H_o : \tau_i = 0 \quad \text{for all} \quad i = 1, \ldots, k$$

If the F-test rejects the null hypothesis there remains the problem of deciding which treatment means are different from each other.

One problem in statistical testing is the determination of the probability of rejecting the null hypothesis when it is true. This probability is called the probability of a Type I error. The other sort of error possible is the acceptance of the null hypothesis when it is false. One must also determine the probability of this error which is called the probability of a Type II error. These probabilities can be determined easily when specific alternative hypotheses are involved. These probabilities become difficult to compute, however, when one is interested in equality or inequality of several treatments as considered for example in tests of all two comparisons.

This report deals with various procedures designed to make realistic comparisons among treatment means. These procedures have been based on somewhat different approaches because several logical points of view are possible regarding the relative importance of the two kinds of errors. What balance should one strike between the probabilities of Type I and Type II errors? Some multiple comparison procedures described in the literature are very cautious about Type I errors. Others try to sacrifice some of this caution in favor of lower probabilities of Type II errors. Other procedures, however, are intermediate testing procedures because they tend to balance excessive Type I and Type II errors.

The purpose of this report is to discuss various properties of each of several procedures separately, illustrating their application with an example, and discussing some of their differences. The Monte Carlo technique was used to compute power, the probability of not committing a Type II error, and protection, the probability of not committing a Type I error, for three test procedures; Fisher's LSD, the multiple-t test, and Duncan's new multiple range test. These are probably the most widely used multiple comparison procedures in experimental statistics.

GENERAL DISCUSSION

Before the properties and application of various tests are considered, it is helpful to discuss some of the general problems and some of the concepts related to Type I and Type II errors which are involved in multiple comparison test procedures.

The Multiple Decision Problem

For some experiments, it is not only necessary to determine equalities among means; but also, if inequalities exist to determine the magnitude of the inequalities. For a two-mean example the following decisions besides equality are possible:

$$\mu_1 < \mu_2$$

$$\mu_2 < \mu_1$$

For three means (n = 3)

$$\mu_1 < \mu_2 < \mu_3$$

$$\mu_1 < \mu_3 < \mu_2$$

$$\mu_2 < \mu_3 < \mu_1$$

$$\mu_2 < \mu_1 < \mu_3$$

$$\mu_3 < \mu_1 < \mu_2$$

$$\mu_3 < \mu_2 < \mu_1$$

$$\mu_1 < \mu_2 = \mu_3$$

$$\mu_2 < \mu_1 = \mu_3$$

$$\mu_3 < \mu_1 = \mu_2$$

$$\mu_1 = \mu_2 < \mu_3$$

$$\mu_1 = \mu_3 < u_2$$

$\mu_2 = \mu_3 < \mu_1$

$\mu_1 < \mu_2$, but μ_3 cannot be ranked relative to μ_1 or μ_2

$\mu_1 < \mu_3$, but μ_2 cannot be ranked relative to μ_1 or μ_3

$\mu_2 < \mu_1$, but μ_3 cannot be ranked relative to μ_2 or μ_1

$\mu_2 < \mu_3$, but μ_1 cannot be ranked relative to μ_2 or μ_3

$\mu_3 < \mu_1$, but μ_2 cannot be ranked relative to μ_3 or μ_1

$\mu_3 < \mu_2$, but μ_1 cannot be ranked relative to μ_3 or μ_2

A total of 19 possible decisions, including $\mu_1 = \mu_2 = \mu_3$, can be made. These 19 possible decisions are well explained and illustrated by Duncan (1) with his geometric method. In the case of three-mean comparisons he is able to represent all decisions in a two-dimensional sample space. Duncan shows properties of symmetry in comparisons among three means. As the number of treatment means increases, the number of decisions increases very rapidly, complicating the decision processes considerably. In the general case with n means there are n ! decisions of the form, $\mu_1 < \mu_2$, $\cdots < \mu_n$, $(n-1)n!/2!$ decisions of the form $\mu_1 = \mu_2 < \mu_3 < \mu_4 < \cdots < \mu_n$ with one pair of means equal $(n-2)n!/3!$ decisions of the form $\mu_1 = \mu_2 =$

$\mu_3 < \mu_4 \ldots < \mu_n$, with three equal means (n-2)n! decisions of the form $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ but $\mu_1 < \mu_3 < \mu_4 < \ldots < \mu_n$, with two overlapping pairs of equal means etc.

In making comparison among a set of means, two factors must be kept in mind: 1) Comparisons can be made on a per treatment basis or 2) on a group basis, i.e., some treatment means are grouped together. Hence it is very important that the procedural method be determined prior to running an experiment. Attempting to group treatments after an experiment has been performed might distort the probability statements one wishes to make from the experiment.

Protection Levels Against Type I Error

Several multiple comparison tests take different points of view toward committing Type I and Type II errors, and this again essentially is the reason for differences among the tests which have been proposed. The probability of a Type I error is called α, i.e., for a two mean comparison

$$\alpha = \Pr\left[\text{decision } (\mu_1 \neq \mu_2) \,\middle|\, \mu_1 = \mu_2\right]$$

The α also is called a significance level when two or more means are compared. The protection level is then defined as the probability of not committing a Type I error. In the Student-Newman-Keuls test protection level is kept at the same level, namely, $(1 - \alpha)$ for the two-mean, three-mean, ..., n-mean comparison. On the other hand, Duncan, Tukey, and Scheff'e, among others believe that the multiple-t test has too low a protection to be a satisfactory multiple comparison test. Duncan (1) taking an intermediate position, suggested the use of special protection levels. The special protection levels

involve degrees of freedom and are given as

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

where

$$\gamma_2 = (1 - \alpha)$$

α = significance level

p = number of means in the subset.

It is called "a p mean protection level and is the minimum probability of finding no wrong significant differences among p observed means." Duncan said, "first it should be noted that if a symmtric test with optimum power functions were constructed subject only to a restriction on the value γ_2, the higher order protection levels would almost invariably be too low to be satisfactory. .... The four-mean protection level of this multiple normal-deviate test, as it may be termed, will be seen later to be only $\gamma_4 = 79.7\%$. That is, the minimum probability of finding no wrong significant differences between the four means is only 79.7%. This is too low to be satisfactory. The three-mean protection levels in the same test have the value $\gamma_3 = 87.8\%$, which is also too low. On the other hand, it does not necessarily follow that all of the high order protection levels should be raised to the value γ_2 of the two-mean protection level as some writers have implicitly assumed. Any increases in the latter levels must necessarily be made at the expense of losses in power and it is most important that the levels be raised no more than is absolutely necessary. ..." This reasoning is developed with an experimental example (1).

Some of these numerical levels are shown below:

	Multiple-t test (4)	Duncan's NMRT
p = 3	87.8%	90.25%
p = 4	79.7%	85.7%

Power of a Test

The power of a test is the probability of rejecting H_o when H_o is false, i.e.,

$$\text{Power} = 1 - \beta,$$

where β = probability of a Type II error.

In studying the power of multiple comparison tests one is confronted with the difficulty pointed out by Duncan that none of the comparisons, even when two means are involved, is a two-decision procedure. The power function applied in this problem is Neyman and Pearson's (12), which however is defined as a power function strictly based on a two-decision concept.

When two means are compared, three decisions are possible: $\mu_1 = \mu_2$, $\mu_1 > \mu_2$, and $\mu_2 > \mu_1$. In order to avoid making separate decisions for the two inequalities, one tests the null hypothesis against the alternative $\mu_1 \neq \mu_2$. The test for a given α_o level, is two-sided to account for the two inequality statements. The power is the probability of the decision expressed as a function of the true difference $d = \mu_1 - \mu_2$. A power function for this case is illustrated by the dotted line in Fig. 1.

Duncan (1) criticized this procedure. He reasoned that, "by pooling the probability of the two decisions $(\mu_1 < \mu_2)$ and $(\mu_2 < \mu_1)$ for any given

value of the true difference, it combines the probability of the correct decision (that μ_1 or μ_2 is the higher means as the truth may be), with the probability of the most incorrect decision (that $\mu_1 > \mu_2$ when in fact $\mu_2 > \mu_1$, or $\mu_2 > \mu_1$ when in fact $\mu_1 > \mu_2$). A function which combines probabilities of serious errors in this way, is of no value in measuring desirable or undesirable properties."

To overcome this difficulty, Duncan (1) suggested that a three-decision test concerning two means can be changed to a joint application of two two-decision tests which would be tested by the hypotheses, $\mu_1 \leq \mu_2$ against the alternative $\mu_2 < \mu_1$, and $\mu_2 \leq \mu_1$ against the alternative $\mu_1 < \mu_2$. Therefore, in a two-mean comparison two Neyman-Pearson power functions are required for obtaining the power of a test. The power curve by this concept is illustrated by the sigmoid and reverse-sigmoid curves respectively in Fig. 1 where α for each one is $\frac{1}{2}\alpha_o$

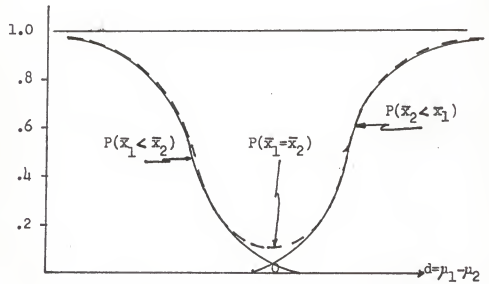

Fig. 1. Power Function for α_o-Level Symmetric Test

From his reasoning for the three-decision case, Duncan generalized to the case of n means. The number of power functions for n means is $n(n-1)/2$. However, Duncan also commented that instead of using all the power functions involved this number can be reduced because of symmetrical properties established in making comparison of the means. If this condition exists, only one of the $n(n-1)/2$ power functions need be investigated in order to investigate them all.

An increase in α will generally result in an increase in the power and consequently a decrease in β when the sample size is fixed. The selection of a particular test procedure depends upon the nature of the consequences of the decisions, i.e., whether one would rather tolerate a relatively high Type I error or a relatively high Type II error.

FISHER'S LSD TEST

In 1935 Fisher (5) proposed a very simple test procedure, later called the Least Significant Difference (or LSD) test, which has been widely used in experimental statistics. This test procedure is just a repeated use of Student's t-test for two mean comparisons after H_o has been rejected. A quantity

$$LSD = s_{\bar{x}} \sqrt{2}\ t_{\alpha,f} \quad ,$$

is computed, where

$$s_{\bar{x}} = \sqrt{\frac{\text{Error Mean Square}}{n}} = \text{the standard error of a mean,}$$

n = Number of observations for the mean

$t_{\alpha,f}$ = the α-level significance value of t with f degrees of freedom,

f = the number of degrees of freedom for the Error Mean Square in the analysis of variance.

Fisher suggested that this LSD be compared with all differences between pairs of sample means. If the difference between any pair of means is greater than the value of LSD, it can be said that μ_i and μ_j are different means. The number of such comparisons among n treatment means is

$$ {}_nC_2 = \frac{n(n-1)}{2} . $$

The following example is based on Keuls' (8) cabbage experiment. A suitable area was divided into 39 plots, grouped into 3 blocks of 13 plots each. In each block the 13 varieties to be investigated were planted in a randomized block design. Keuls stated, "The purpose was to learn which variety would give the highest gross yield per head of cabbage and which the lowest, in other words to find approximately the order of the varieties according to gross yield per cabbage." However, this report will use Keuls' example as illustrating the application of all the test procedures to be discussed. Table 1 shows the coded data and the means for each variety.

Table 1. Keuls' (8) Cabbage Experiment. Coded Gross Weights

Variety	Block A	Block B	Block C	Average	Rank
1	89	46	33	176.0	1
2	0	-5	-21	111.3	11
3	-25	-16	-26	97.7	13
4	22	7	-3	128.7	8
5	8	6	-12	120.7	10
6	31	10	-5	132.0	5
7	54	15	-4	141.7	4
8	-8	-20	-30	100.7	12
9	31	7	-5	131.0	6
10	26	14	-27	124.3	9
11	67	29	2	152.7	2
12	65	21	6	150.7	3
13	23	7	-3	129.0	7

In this example the null hypothesis would be that there are no variety effects on the size of the cabbage heads. Table 2 shows the results of the analysis of variance applied to Keuls' data.

Table 2. Analysis of Variance for Table 1. $\alpha = .05$

Source of Variations	D/F	Mean Square: Sample	Expected	F
Varieties	12	1392.8	$\sigma^2 + 0.25\Sigma T_i^2$	11.21*
Blocks	2	440.75	$\sigma^2 + 13\sigma^2_j$	35.46*
Error	24	124.29	σ^2	
	38			

The observed F-ratio for varieties is found to be significant so the null hypothesis is rejected. Keuls concluded that it was improbable that the variety observed means form a random sample from one and the same normal population. The next problem is to find all significant differences among variety means by the LSD procedure.

$$LSD = 2.064 \sqrt{124.29\ (2/3)} = 18.8 \ ,$$

where

$$t_{.05,24} = 2.064$$

Table 3 is constructed as an ordered array by listing all the variety means in rank order. All the differences between means have been computed. The differences which are less than LSD = 18.8 are underlined.

Table 3. Fisher's LSD Analysis of Keuls' Cabbage Data. $\alpha = .05$

Var	3	8	2	5	10	4	13	9	6	7	12	11	1
Var Mean	97.7	100.7	111.3	120.7	124.3	128.7	129.0	131.0	132.0	141.7	150.7	152.7	176.0
3		3.0	13.6	23.0	26.6	31.0	31.3	33.3	34.3	44.0	53.0	55.0	78.3
8			10.6	20.0	23.6	28.0	28.3	30.3	31.3	41.0	50.0	52.0	75.3
2				9.4	13.0	17.4	17.7	19.7	20.7	30.4	39.4	41.4	64.7
5					3.6	8.0	8.3	10.3	11.3	21.0	30.0	32.0	55.3
10						4.4	4.7	6.7	7.7	17.4	26.4	28.4	51.7
4							0.3	2.3	3.3	13.0	22.0	24.0	47.3
13								2.0	3.0	12.7	21.7	23.7	47.0
9									1.0	10.7	19.7	21.7	45.0
6										9.7	18.7	20.7	44.0
7											9.0	11.0	34.3
12												2.0	25.3
11													23.3

The hypotheses, $\mu_i = \mu_j$, which are accepted, $\alpha = .05$ are therefore

$\mu_2 = \mu_3$	$\mu_4 = \mu_{10}$	$\mu_7 = \mu_9$
$= \mu_4$	$= \mu_{13}$	$= \mu_{10}$
$= \mu_5$	$\mu_5 = \mu_6$	$= \mu_{11}$
$= \mu_8$	$= \mu_9$	$= \mu_{12}$
$= \mu_{10}$	$= \mu_{10}$	$= \mu_{13}$
$= \mu_{13}$	$= \mu_{13}$	$\mu_9 = \mu_{10}$
$\mu_3 = \mu_8$	$\mu_6 = \mu_7$	$= \mu_{13}$
$\mu_4 = \mu_5$	$= \mu_9$	$\mu_{10} = \mu_{13}$
$= \mu_6$	$= \mu_{10}$	$\mu_{11} = \mu_{12}$
$= \mu_7$	$= \mu_{12}$	
$= \mu_9$	$= \mu_{13}$	

and all other remaining hypotheses, $\mu_i = \mu_j$ are rejected.

MULTIPLE-t TEST

A test procedure very similar to the LSD test is the multiple-t. It does not require an F-test before applying the LSD procedure to all hypotheses of the form $\mu_i = \mu_j$, for all i, j. The multiple-t test will, in general, have lower probability of Type II error than the LSD test because even though the F-test leads to acceptance of the null hypothesis, the multiple-t test may detect differences among means.

This test could be applied to Keuls' cabbage example from the preceding section; but because F for varieties was significant, the results of this test are exactly the same as the LSD test.

LIMITATIONS OF LSD AND MULTIPLE-t TESTS

The LSD and multiple-t test procedures only test hypotheses $\mu_i = \mu_j$, i and j, 1 to n. However, in some experimental situations it is useful to make comparisons, not only of two-mean groups but also three-mean groups, . . ., and n-mean groups. Other forms of decisions also may be desired, such as certain linear combinations of treatment means. In addition the LSD and multiple-t tests make the probability of committing Type I errors somewhat greater than the specified significance level. Therefore several investigators (1, 14, 17) have proposed other test procedures. These are usually of two kinds: multiple range tests and multiple-F tests.

MULTIPLE RANGE TESTS

The Student-Newman-Keuls Test

This multiple range test differs from the multiple-t test in that the protection level for a group of n means is fixed at $(1 - \alpha)$ for all p - mean comparisons, p = 2, . . ., n. Student (16) first suggested using the quantity q = w/s, to determine differences among treatment means where w is the range in a sample of n observations from a normal population with standard deviation, σ, and s^2 is an independent estimate of σ^2. Later Newman (11) modified Student's idea presenting a table which was computed by quadrature from Pearson's (12) approximate probability law of the studentized range. Keuls (8) developed these ideas further. The Student-Newman-Keuls test is called a multiple range test because the over-all procedure involves the repeated use of range tests on the p-mean groups, p = 2, . . ., n. This method is summarized by the following rule: The difference

between any two means in a set of n means is significant provided the range of each and every subset which contains the given two means is significant in an α-level range test. Federer (4) lists the steps for following this rule:

"Step (I) Subdivide the treatment means into biological, physical, or sociological groups. Natural groupsings as prescribed by the choice of the particular set of treatments have meaning; it is doubtful if a ranked set of means from two or more natural groups has any practical significance.

Step (II) Choose a significance level, α, which usually will be the 5 or 1 per cent level.

Step (III) Compute the standard error of a treatment mean, $s_{\bar{x}}$, and the values $W_n = q_{\alpha,n} s_{\bar{x}}$, $W_{n-1} = q_{\alpha,n-1} s_{\bar{x}}$, . . ., $W_3 = q_{\alpha,3} s_{\bar{x}}$, and $W_2 = q_{\alpha,2} s_{\bar{x}} = t_{\alpha,f} \sqrt{2}\, s_{\bar{x}} = \text{LSD}$. Rank the treatment means from highest to lowest $\bar{x}_n$, $\bar{x}_{n-1}$, . . ., $\bar{x}_2$, $\bar{x}_1$.

Step (IV) Compare the range of n treatments, $\bar{x}_n - \bar{x}_1$ with the calculated W_n. If $\bar{x}_n - \bar{x}_1$ is less than W_n, the process stops, and the n-means are asserted to belong to a non-heterogeneous group. If $\bar{x}_n - \bar{x}_1 \geq W_n$ subdivide the means into two groups of n-1 means each, $\bar{x}_n$ to $\bar{x}_2$ and $\bar{x}_{n-1}$ to $\bar{x}_1$ and state that $\bar{x}_n$ is different from $\bar{x}_1$. Then, compare the range $\bar{x}_n - \bar{x}_2$ and $\bar{x}_{n-1} - \bar{x}_1$ with W_{n-1}. If either range is less than W_{n-1} the means in the group are said to belong to a single group. If either range exceeds W_{n-1} the n-1 means are divided into two groups

of n-2 means each and compared with W_{n-2}. The process continues until a subset of means is obtained which does not exceed the calculated value W_i. The process stops whenever the actual range of the subset is less than the calculated range. No subset of means is compared if the subset is included in a larger subset which is less than the calculated range W_i."

where

$$q_n = \frac{\bar{x}max - \bar{x}min}{S\bar{x}} = \frac{\text{range}}{\text{Standard deviation}}$$

As an example, consider Keuls' cabbage data. For the thirteen variety means, twelve W_i values are needed, namely,

$$s_{\bar{x}} = \sqrt{\frac{124.29}{3}} = \sqrt{41.43} = 6.437 \text{ with 24 degrees of freedom}$$

$$= .05$$

$W_{n=13} = 5.18\ (6.437) = 33.3$

$W_{n=12} = 5.10\ (6.437) = 32.8$

$W_{n=11} = 5.01\ (6.437) = 32.2$

$W_{n=10} = 4.92\ (6.437) = 31.7$

$W_{n=9} = 4.81\ (6.437) = 31.0$

$W_{n=8} = 4.68\ (6.437) = 30.1$

$W_{n=7} = 4.54\ (6.437) = 29.2$

$W_{n=6} = 4.37\ (6.437) = 28.1$

$W_{n=5} = 4.17\ (6.437) = 26.8$

$W_{n=4} = 3.90\ (6.437) = 25.1$

$W_{n=3} = 3.53\ (6.437) = 22.7$ and

$W_{n=2} = 2.92\ (6.437) = 18.8 = \text{lsd.}$

The thirteen variety means, in order of size, are:

Var.	3	8	2	5	10	4	13	9	6	7	12	11	1
Mean	97.7	100.7	111.3	120.7	124.3	128.7	129.0	131.0	132.0	141.7	150.7	152.7	176.0

The first step requires comparing the W_i with the range between the largest and smallest mean until the observed value does not exceed the calculated value:

$$\bar{x}_1 - \bar{x}_3 = 176.0 - 97.7 = 78.3 > W_{n=13} = 33.3$$

$$\bar{x}_1 - \bar{x}_8 = 176.0 - 100.7 = 75.3 > W_{n=12} = 32.8$$

$$\bar{x}_n - \bar{x}_3 = 152.7 - 97.7 = 55.0 > W_{n=12} = 32.8$$

$$\bar{x}_1 - \bar{x}_2 = 176.0 - 111.3 = 63.0 > W_{n=11} = 32.2$$

$$\bar{x}_{12} - \bar{x}_3 = 150.7 - 97.7 = 53.0 > W_{n=11} = 32.2$$

$$\bar{x}_{11} - \bar{x}_8 = 152.7 - 100.7 = 52.0 > W_{n=11} = 32.2$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\bar{x}_{12} - \bar{x}_2 = 150.7 - 111.3 = 39.4 > W_{n=9} = 31.0$$

$$\bar{x}_{12} - \bar{x}_5 = 150.7 - 120.7 = 30.0 < W_{n=8} = 30.1$$

The comparisons continue in this way. The result of comparisons of ranges between variety means leads to acceptance of the following hypotheses of the form $\mu_i = \mu_j$.

$\mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_8 = \mu_9 = \mu_{10} = \mu_{13}$

$\mu_3 = \mu_5 = \mu_8 = \mu_{10}$

$\mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_9 = \mu_{10}$

$\mu_4 = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10} = \mu_{13}$

$\mu_5 = \mu_{12} = \mu_{13}$

$\mu_6 = \mu_7 = \mu_9 = \mu_{10} = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_7 = \mu_9 = \mu_{10} = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_8 = \mu_{10} = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_9 = \mu_{10} = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_{10} = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_{11} = \mu_{12} = \mu_{13}$

$\mu_{12} = \mu_{13}$

and all other remaining hypotheses $\mu_i = \mu_j$ are rejected. From this example it can be noticed that this test accepts equalities of variety means for wider ranges of sample means than do the LSD and multiple-t tests.

Duncan's New Multiple Range Test

Duncan (1) has proposed a test called the new multiple range test (NMRT). He claims that this test is an optimum procedure. This procedure is intended to be a compromise between the Student-Newman-Keuls test and the multiple-t test. Duncan attempts to strike an optimum combination of

probabilities of Type I and II errors by introducing protection levels which vary with degrees of freedom. The values of his so-called shortest significant range, denoted as R_p, is smaller than the values used in any other range test except the LSD or multiple-t provided all the conditions are the same. If the difference between any two means exceeds the corresponding R_i, it is declared to be significant and $H_o(\mu_i = \mu_j)$ is rejected with one exception. The exception is that no difference between two means can be declared significant if the means concerned are contained in any subset between means in the ordered array which have a non-significant range. For example, consider five treatment means, $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$, and $\bar{x}_5$, and assume these means are in order of size. Then if it is found that $\bar{x}_5 - \bar{x}_1 < R_5$, $H_o(\mu_1 = \mu_5)$ is accepted; and no other differences among means between $\bar{x}_1$ and $\bar{x}_5$ can be considered significant, even though some difference might exceed the appropriate R_p. In other words, it is not possible to make decisions such as

$$\mu_1 < \mu_2, \quad \mu_2 < \mu_5, \text{ or } \mu_1 < \mu_3 < \mu_4, \text{ etc.}$$

if $H_o(\mu_1 = \mu_5)$ has been accepted and the sample means are in rank order from $\bar{x}_1$ to $\bar{x}_5$.

Duncan's shortest significant ranges are computed as follows:

$$R_p = s_{\bar{x}} q^*_{\alpha,p} ,$$

where

$s_{\bar{x}}$ = the standard error,

$q^*_{\alpha,p}$ = the tabular value (1) of Duncan's special significant studentized ranges with the α-level of significance and p the number of means in the subset.

The shortest significant range, which is $R_2 = s_{\bar{x}} q^*_{\alpha,2}$ is the same value as $W_{n=2}$ for Student-Newman-Keuls' test, which is the LSD. In the case of n means, the desired number of shortest significant ranges is (n-1). These R_p values increase at a somewhat slower rate than W_i in the Student-Newman-Keuls' test. Duncan's shortest significant ranges are computed so that the protection level for a group of n means is not fixed at $(1-\alpha)$, as for any subset of means in Student-Newman-Keuls' test, but is $(1-\alpha)^{p-1}$, where $p = 2, 3 \ldots, n$. It is Duncan's belief that this protection against Type I errors is adequate, and his NMRT maintains better power against Type II errors than does the Student-Newman-Keuls' test.

The data from Keuls' experiment are used to illustrate the calculations for Duncan's NMRT.

The standard error for the mean is

$$s_{\bar{x}} = \sqrt{124.29/3} = 6.437$$

with twenty-four degrees of freedom. The calculated least significant ranges are computed as follows:

$$R_{13} = 3.43\ (6.437) = 22.1$$

$$R_{12} = 3.41\ (6.437) = 22.0$$

$$R_{11} = 3.40\ (6.437) = 21.9$$

$$R_{10} = 3.38\ (6.437) = 21.8$$

$$R_{9} = 3.37\ (6.437) = 21.7$$

$$R_{8} = 3.34\ (6.437) = 21.5$$

$$R_{7} = 3.31\ (6.437) = 21.3$$

$$R_{6} = 3.28\ (6.437) = 21.1$$

$$R_{5} = 3.22\ (6.437) = 20.7$$

$$R_{4} = 3.15\ (6.437) = 20.3$$

$$R_{3} = 3.07\ (6.437) = 19.8$$

$$R_{2} = 2.92\ (6.437) = 18.8 = \text{LSD}\ .$$

The result of using this test leads to acceptance of the following hypotheses $\mu_i = \mu_j$ concerning variety means:

$\mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_8 = \mu_9 = \mu_{10} = \mu_{13}$

$\mu_3 = \mu_8$

$\mu_4 = \mu_5 = \mu_6 = \mu_7$

$\mu_4 = \mu_9 = \mu_{10} = \mu_{13}$

$\mu_5 = \mu_6 = \mu_7 = \mu_9 = \mu_{10} = \mu_{13}$

$\mu_6 = \mu_7 = \mu_9 = \mu_{10} = \mu_{12}$

$\mu_6 = \mu_{13}$

$\mu_7 = \mu_9 = \mu_{10} = \mu_{11} = \mu_{12} = \mu_{13}$

$\mu_9 = \mu_{10} = \mu_{12} = \mu_{13}$

$\mu_{10} = \mu_{13}$

$\mu_{11} = \mu_{12}$

and all other remaining hypotheses, $\mu_i = \mu_j$ are rejected.

Tukey's Test Based on Allowances

Several multiple comparison procedures have been introduced by J. W. Tukey (17). However, this report deals with only one of his procedures based on "allowances". When there are only two treatments, an "allowance" is the same as the LSD = $ts_{\bar{x}}\sqrt{2}$. However, when there are more than two treatments, the test based on allowances becomes a multiple range test and an "allowance" is equal to the value of $W_{n=n}$ obtained in the Student-Newman-Keuls' multiple range test. This value is called an hsd (honestly significant difference) and if two means (or groups of means) differ by more than hsd they are said to differ significantly. This procedure may also be used for finding confidence interval for the difference between any two means.

In his paper Tukey (17) discussed the experimenter's desire to examine

all contrasts between treatment means; not only simple comparisons, i.e., differences between pairs of treatment means. His error rate is on a per experiment basis rather than on a per decision basis for these comparisons. He felt that it would be impractical to set an α-level significance test for each of the comparisons because the accumulated total errors over all comparisons among the n treatments would be too high. This is the objection to the LSD test most often found in the literature.

The value of hsd is 33.3 for testing differences between variety means in Keuls' cabbage experiment. There are a fewer number of rejections of the hypothesis $\mu_i = \mu_j$ for Tukey's procedure than for any other test so far discussed.

In 1953, Tukey (18) proposed another multiple range test procedure with a less conservative attitude towards Type I error than in his previous test. In this procedure the significant ranges are midway between the ones required by the test based on allowances and those by the Student-Newman-Keuls' test.

MULTIPLE-F TESTS

The multiple-F test consists of the combined use of range tests and results of significant F tests. Duncan's (2) multiple comparison test and Scheffé's (13) test are generally recognized as representative of this procedure.

Duncan's Multiple Comparison Test

Duncan (2) proposed a multiple comparison test in 1951, which he described as a "Multiple-F Test". It was a compromise between two rules. These he defined as: "Rule 1 is the difference between any two means in a set of n means is significant provided the variance of each and every subset which contains the given means is significant according to an α_p -level F test where $\alpha_p = 1 - \gamma_p$, where $\gamma_p = (1 - \alpha)^{p-1}$, and p is the number of means in the subset concerned. Rule 2 is any comparison of the form $c = \sum_{i=1}^{n} k_i \bar{x}_i$ is significantly different from zero provided the variance of each and every subset which contains all of the means involved in c is significant according to an α_p -level F test; and provided, also, that c differs significantly from zero according to an α -level t-test, where $c = \sum_{i=1}^{n}(k_i \bar{x}_i)$ and $k_1, k_2, \ldots, k_n$ is any set of arbitrary constants such that $\sum_{i=1}^{n} k_i = 0$." Rule 1 is similar to the method described for Fisher's (5) LSD test except that Duncan used an α_p -level instead of an α -level. Duncan's compromise should be interpreted so that as many significant differences as possible are found by Rule 1.

Rule 2 is then used to test any comparisons within subsets of means already found to contain significant differences by Rule 1.

Duncan's (2) multiple comparison test can be summarized in the following four steps:

Step 1 List treatment means ranked in order, e.g.,

$$\bar{x}_1 < \bar{x}_2 < \ldots < \bar{x}_n \ .$$

Step 2 Determine significant ranges from

$$R'_p = s_{\bar{x}} q_{\alpha;p,f} ,$$

where

$s_{\bar{x}}$ = estimated standard error of a mean

$q_{\alpha;p,f}$ = the tabular value (4) with α-level significance. These values are different from $q^*_{\alpha;p,f}$ for the range test.

p = the number of means in a subset

f = degrees of freedom associated with $s_{\bar{x}}$.

Step 3 Determine a set of least significant sums of squares and of the sumes of squares among certain combinations of means

$$ss_p = 1/2\ R'^{2}_{p} .$$

These values are compared with sums of squares among means. For example if there are three means: $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$, compute the sums of squares

$$ss'_{1,2,3} = \bar{x}_1^2 + \bar{x}_2^2 + \bar{x}_3^2 - \frac{(\bar{x}_1 + \bar{x}_2 + \bar{x}_3)^2}{3}$$

and compare with $1/2\ R'_3 = ss_3$.

Step 4 This step is used only in certain cases when the sample range for all means in the group is less than R'_2, and the observed sum of squares among means is larger than the computed sum of squares for the number in the group. The n-1 degrees of freedom

are partitioned into single degree of freedom contrasts. The comparison or comparisons contributing to the significant sum of squares are segregated as in Step 3. This is useful when testing the significance of a comparison involving groups of means.

The first step is conducted ordinarily as a range test. In order to compute the significant range for n means in Step 1, Duncan uses the relation

$$S_{\bar{x}} \sqrt{2(n-1)\ F_{(n-1,\ f)}} = q_n\ s_{\bar{x}} = R'_n \quad .$$

Duncan (3) recently published a new procedure. It is called the Minimum Average-Weight-Risk Analysis, and is based on Bayes' theorem and some recent ideas of Lehmann (9). This new procedure is similar in concept of the multiple-t test. Duncan is still conducting research on this problem.

Scheffé's Test

In 1953, Henry Scheffé (13) introduced a test procedure based on linear contrasts which include a wide variety of treatment comparisons. This test procedure may be described as an F-test analogue of Tukey's (18) test based on allowances. This multiple comparison test is defined by Scheffé as: "A kind of simultaneous interval estimation and multiple significance test. For testing the hypothesis

$$H_o:\ \mu_1 = \mu_2 = \cdots = \mu_n$$

a contrast among the parameters $\mu_1, \mu_2, \cdots, \mu_n$, is defined to be a linear function of the μ_i's: $\sum_{i=1}^{n} c_i = 0$. This statement is similar to Rule 2 in Duncan's multiple comparison test because both are linear combinations

of the means.

Scheffé (14) stated the following theorem; "The probability is $(1-\alpha)$ that the values of all contrasts simultaneously satisfy the inequalities

$$\hat{\psi} - S\hat{\sigma}_{\hat{\psi}} \leq \psi \leq \hat{\psi} + S\hat{\sigma}_{\hat{\psi}}"$$

where $\psi = \sum_{i=1}^{n} c_i \mu_i$,

$\hat{\psi} = \sum_{i=1}^{n} c_i \bar{x}_i$ (unbiased estimate of ψ),

$\hat{\sigma}^2_{\hat{\psi}}$ = variance of $\hat{\psi}$

$S = \sqrt{(n-1) F_{\alpha;n-1,f}}$,

$n-1$ = degrees of freedom for parameters, and

f = degrees of freedom for the error variance.

In the general case for any linear functions, i.e., no restriction on $\sum c_i$, the same confidence interval is used as in the above case, and the same probability statement is applicable, but the value of S becomes

$$S = \sqrt{q F_{\alpha;q,f}}$$

where q is generally the number of means.

Scheffé himself admitted this method is undesirable for contrasts of the type $\mu_i = \mu_j$ because it gives rather wide intervals compared to other methods. Therefore, this test is capable of accepting the null hypothesis $H_o(\mu_i = \mu_j)$ too often.

A careful study of Scheffé's test will show that it is built similar to Tukey's test with allowances, which also is based on confidence intervals.

GRAPHICAL COMPARISON OF TEST PROCEDURES

The purpose of showing Fig. 2 is to summarize schematically the comparison of the results of various test procedures applied to the Keuls' cabbage experiment. Here only five procedures are compared namely: Tueky's test, Duncan's new multiple range test, Student-Newman-Keuls' test, LSD test, and multiple t-test. Among these tests Tukey's test and S-N-K test are much alike in that they accept more equalities among variety means than the three other tests. Tukey's test is the only one that decalres Variety 1 equal to Varieties 11 and 12. The LSD test and Duncan's NMRT detect differences and accept equality among the means in reasonably the same way except for a little disagreement in the middle of the range of variety means.

EMPIRICAL STUDY OF MULTIPLE COMPARISON TESTS

Results of Monte Carlo Study

The purpose of this part of the study is to attain a practical evaluation of three of the multiple comparison test procedures described above. Specifically, interest is focused on determining the power and the protection level of three test procedures; namely, Fisher's LSD, multiple-t and Duncan's new multiple range test. In this study, the power and the protection level were determined separately by using different combinations of means with different variances.

In order to conduct this study the following three stages were required:

Stage 1 Random samples of size n = 10 were drawn from populations with known means and variances. The population means ranged from 5.0 to 13.0 and their variances were either 4 or 16.

Stage 2 Step 1 was repeated 100 to 500 times depending on the number of decisions desired for chosen sets of means and variance. An analysis of variance, F-test, and Duncan's shortest significant difference for each set was obtained. The specific combination of means and variances are shown in Tables 4 and 5. A high-speed computer was used draw samples and to perform computations for Stage 1 and 2.

Stage 3 All the differences between means were computed, and these differences were compared to corresponding values of LSD and the Duncan's Rp*s. The results are illustrated by an example, given in analysis of data.

For the first situation in Table 5, a set of means: 5,5,5,7,7 with variance equal to 4 was used. The total number of decisions made was 4600, since

$$\binom{5}{2} = \frac{5(5-1)}{2} = 10 \text{ decisions for each of 460 sets of samples.}$$

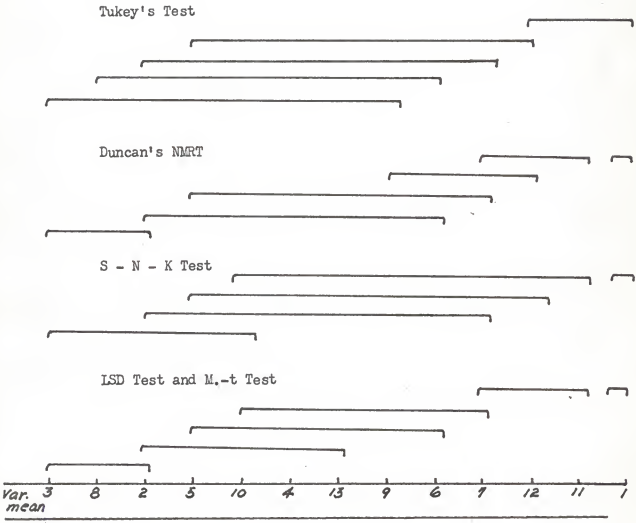

Fig. 2. The variety means within a bracket are asserted to be not heterogeneous, and means not bracketed together are asserted to be different.

Analysis of Data

Case N(5,5, 6,6, 7,7, 8,8, 9,9;4)

$F_{.05,9,90} = 1.98$; $F = 6.70^{*}$

$Rp^{*} = \frac{1.69}{lsd}$, 1.77, 1.83, 1.87, 1.90, 1.93, 1.95, 1.98, 1.99

Ordered Array of Means

	1	2	5	6	3	4	7	8	9	10
	5.20	5.40	5.80	6.53	6.88	7.18	8.16	8.47	8.58	9.92
1					1.68	1.98	2.96	3.27	3.38	4.72
2						1.78	2.76	3.07	3.18	4.52
5						1.38	2.36	2.67	2.78	4.12
3							1.28	1.94	2.05	2.39
4								1.29	1.70	3.04
7									1.40	2.76
8										1.76
9										1.34
10										

DECISIONS

H_o	LSD	NMRT	MT	H_o	FLSD	NMRT	MT	H_o	LSD	NMRT	MT
$\mu_1=\mu_2$				$\mu_2=\mu_9$	R	R	R	$\mu_5=\mu_6$			
$\mu_1=\mu_3$				$\mu_2=\mu_{10}$	R	R	R	$\mu_5=\mu_7$	R	R	R
$\mu_1=\mu_4$	R	R	R	$\mu_3=\mu_4$				$\mu_5=\mu_8$	R	R	R
$\mu_1=\mu_5$				$\mu_3=\mu_5$				$\mu_5=\mu_9$	R	R	R
$\mu_1=\mu_6$				$\mu_3=\mu_6$				$\mu_5=\mu_{10}$	R	R	R
$\mu_1=\mu_7$	R	R	R	$\mu_3=\mu_7$				$\mu_6=\mu_7$			
$\mu_1=\mu_8$	R	R	R	$\mu_3=\mu_8$				$\mu_6=\mu_8$	R	R	R
$\mu_1=\mu_9$	R	R	R	$\mu_3=\mu_9$	R		R	$\mu_6=\mu_9$	R	R	R
$\mu_1=\mu_{10}$	R	R	R	$\mu_3=\mu_{10}$	R	R	R	$\mu_6=\mu_{10}$	R	R	R
$\mu_2=\mu_3$				$\mu_4=\mu_5$				$\mu_7=\mu_8$			
$\mu_2=\mu_4$	R		R	$\mu_4=\mu_6$				$\mu_7=\mu_9$			
$\mu_2=\mu_5$				$\mu_4=\mu_7$				$\mu_7=\mu_{10}$	R		R
$\mu_2=\mu_6$				$\mu_4=\mu_8$				$\mu_8=\mu_9$			
$\mu_2=\mu_7$	R	R	R	$\mu_4=\mu_9$				$\mu_8=\mu_{10}$			
$\mu_2=\mu_8$	R	R	R	$\mu_4=\mu_{10}$	R	R	R	$\mu_9=\mu_{10}$			

NO. CORRECT EQUALITIES 5-5-5 NO CORRECT INEQUALITIES 21-18-21

Table 4. Summary of protection levels for indicated sampling situations, as obtained by Monte Carlo studies. All samples were of size n=10 from each population.

Experimental Situation	Number of Decisions	%Correct Decisions When Equality True: FLSD	Mt	NMRT
N(5,5,5,7,7; 4)	1840	97	97	98
N(5,5,5.5,5.5,6,6,6.5,6.5, 7,7; 4)	1500	96	96	97
N(5,5,5,5,5,5,7,7,7,7; 4)	7035	95	95	98
N(5,5,6,6,7,7,8,8,9,9; 4)	1500	94	94	96
N(5,5,5,5.5,5.5,5,5,5,5.5,6, 6,6,6,6.5,6.5,6.5,6.5,7, 7,7,7; 4)	1500	94	94	98
N(5,5,7,7,9,9,11,11,13,13; 4)	1500	96	96	96
N(5,5,5,7,7; 16)	1128	98	92	92
N(5,5,5.5,5.5,6,6,6.5,6.5, 7,7; 16)	1370	98	95	97
N(5,5,5,5,5,5,7,7,7,7; 16)	2100	97	94	94
N(5,5,6,6,7,7,8,8,9,9; 16)	500	93	92	95

Table 5. Summary of powers for indicated sampling situations, as obtained by Monte Carlo studies. All samples were of size n=10 from each population.

Experimental Situation	Number of Decisions	%Correct Decisions If Inequality True FLSD	Mt	NMRT
N(5,5,5,7,7,; 4)	2760	51	55	53
N(5,5,5.5,5.5,6,6,6.5,6.5, 7,7; 4)	12000	18	22	16
N(5,5,5,5,5,5,7,7,7,7; 4)	8040	56	58	49
N(5,5,6,6,7,7,8,8,9,9; 4)	12000	54	54	49
N(5,5,5,5,5.5,5.5,5.5,5.5, 6,6,6,6,6.5,6.5,6.5,6.5, 7,7,7,7; 4)	8000	25	25	15
N(5,5.5,6,6.5,7.7.5,8,8.5, 9,9.5; 4)	8910	48	48	44
N(5,6,7,8,9; 4)	7030	52	52	51
N(5,7,9,11,13; 4)	3960	83	83	83
N(5,5,7,7,9,9,11,11,13,13; 4)	12000	83	83	81
N(5,5,5,7,7; 16)	1692	14	19	17
N(5,5,5.5,5.5,6,6,6.5,6.5, 7,7; 16)	10960	3	9	5
N(5,5,5,5,5,5,7,7,7,7; 16)	2400	11	21	13
N(5,5,6,6,7,7,8,8,9,9; 16)	4000	20	24	18
N(5,6,7,8,9; 16)	4510	16	22	18
N(5,7,9,11,13; 16)	3850	74	74	74

Within this set of 10 comparisons there are 4 equalities and 6 inequalities so that total equalities = 4 x 460 = 1840 and total inequalities = 6 x 460 = 2760.

The protection levels (Table 4) and powers (Table 5) are computed in the following way, as for example was done for N(5, 5, 7, 7, 9, 9, 11, 11, 13, 13; 4). A total of 300 sets were used, hence

total number of decisions = 45 x 300 = 13500

total number of true equalities = 1500

total number of true inequalities = 12000

The following are the numbers of incorrect decisions among the 1500 decisions possible on equalities:

	LSD	MT	NMRT
Total No.	64	62	64

Percentage: 64/1500 = 4.27 62/1500 = 4.13 64/1500 = 4.27

Therefore, the per cent of correct decisions for the situation is 96 per cent, to the nearest whole per cent.

The numbers of correct decisions for inequalities are

	LSD	MT	NMRT
Total No.:	9943	9765	9943
% = Total No./13500	82.86	81.37	82.86

Comments on the Empirical Results

From Table (4) and Table (5) the multiple-t test has power superiority as predicted over the other two tests. Fisher's LSD test has a power advantage, in most situations, over Duncan's NMRT. For the Type I error, the multiple-t

test gives a slightly higher percentage of errors than the other two tests, which is not serious. For Fisher's LSD test and Duncan's NMRT, the probabilities of committing Type I errors are variable from one situation to another.

Empirically, therefore, the multiple-t test has better power than either of the other tests, and is simpler and more convenient to apply. If one fears that its Type I error rate is too high--which is not confirmed herein--one can use Fisher's LSD and maintain $(1-\alpha)$ protection against Type I errors on the n-mean decisions.

ACKNOWLEDGMENT

The writer deeply wishes to express his sincere gratitude to his major professor, Dr. Holly C. Fryer, for his many helpful suggestions and kind encouragement given, not only during the preparation of this report, but also during the past two years while this writer was pursuing graduate studies.

The writer is also indebted to Mr. Larry Janssen, Department of Statistics, for his help on the high-speed computer.

REFERENCES

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

2. Duncan, D. B.
"On the Properties of the Multiple Comparisons Test". Virginia Journal of Science, 3:49-67, 1952.

3. Duncan, D. B.
"Bayes Rules for a Common Multiple Comparison Problem and Related Student-t Problems". Annals of Mathematical Statistics, 32,: 1013-1033, 1961

4. Federer, W. T.
Experimental Design. The Macmillan Company, New York, 1955.

5. Fisher, R. A.
The Design of Experiments, Sixth Edition. Oliver and Boyd, London, 1935-1951.

6. Fryer, H. C.
"Statistical Methods I and II". Unpublished teaching notes, Kansas State University, Manhattan, 1961.

7. Hartley, H. O.
"Some Significance Test Procedure for Multiple Comparisons". (Abstract) Annals of Mathematical Statistics, 25:176, 1954.

8. Keuls, M.
"The Use of the 'Studentized Range' in Connection with an Analysis of Variance". Euphytica, 1:112-122, 1952.

9. Lehmann, E. L.
"A Theory of Some Multiple Decisions Problems, 1". Annals of Mathematical Statistics, 28:1-25, 1957.

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

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

12. Pearson, E. S., and H. O. Hartley.
"Tables of the Probability Integral of the 'Studentized Ranges'". Biometrika, 33:89-99, 1943.

13. Scheffe, H.
"A Method of Judging All Contrasts in the Analysis of Variance". Biometrika, 40:87-104, 1953.

14. Scheffe, H.
The Analysis of Variance. John Wiley and Sons, New York, 1959.

15. Snedecor, G. W.
Statistical Methods. Fifth Edition. Iowa State College Press, Ames. 1956.

16. Student
"Errors of Routine Analysis". Biometrika, 19:151-164, 1957.

17. Tukey, J. W.
"Comparing Individual Means in the Analysis of Variance". Biometrics, 5:99-114, 1949.

18. Tukey, J. W.
"The Problem of Multiple Comparisons". Unpublished dittoed notes, Princeton University, Princeton, 1953.

PROPERTIES OF MULTIPLE COMPARISON TEST PROCEDURES

by

DONG HYUN KIM

B. S., Washburn University, Topeka, Kansas, 1960

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirement for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

This report is an attempt to compare and contrast various multiple comparison tests developed by statisticians and mathematicians since 1927. Not only are the test procedures illustrated, but also a discussion of the procedures, emphasizing their individual advantages and limitations, has been attempted.

Whenever a statistical experiment is conducted which is intended to compare treatment effects of various sorts, the experimenter hopes to determine which treatments are equal and which are unequal, on the average with respect to the measurement taken. Essentially, the experiment would decide whether or not the samples came from the same population. Usually the population parameter of most interest is the mean. If the treatment means are different from one another, it is of interest to know which means differ, and what are the magnitudes of these differences. These questions can be answered by some of the methods of multiple comparison.

A discussion of the concept of multiple decisions, protection level against Type I error, and power of a test precedes the descriptions of the various testing procedures.

The following test procedures are discussed: Fisher's LSD test, the multiple-t test, the Student-Newman-Keuls' test, Duncan's new multiple range test, Tukey's test based on allowances, Duncan's multiple comparison test, and Scheffe's test. These test procedures were felt to be representative in their method of attacking the problem of multiple comparisons. They differ from each other primarily in the relative importance assumed for errors of the first and second kinds. The underlying assumptions are usually normality and homogeneity of variance.

Some results of some Monte Carlo studies of three multiple comparison

test procedures are reported. The three test procedures considered were: Fisher's LSD test, the multiple-t test and Duncan's new multiple range test. These tests were compared for protection against Type I error, and with respect to their powers against Type II error for a number of known sampling situations in which differences among the population means were known to exist. Most discussions in the literature seem to overemphasize avoidance of Type I error, when, in fact, most experiments are conducted after an attempt has been made to create real differences.

It was found that:

a) No test, even Fisher's LSD, had poor protection against the sort of Type I error studied.

b) The powers of the three tests studied generally were in the order:

Multiple-t Fisher's LSD Duncan's NMRT.

The latter conclusion only verifies Duncan's own statements but also the conclusion a) seems to be contrary to the fears usually expressed in the literature.