MULTIPLE RANGE AND MULTIPLE F TESTS*

DAVID B. DUNCAN**

Virginia Polytechnic Institute
Blacksburg, Virginia

1. INTRODUCTION

The common practice for testing the homogeneity of a set of n treatment means in an analysis of variance is to use an F (or z) test. This procedure has special desirable properties for testing the homogeneity hypothesis that the n population means concerned are equal. An F test alone, however, generally falls short of satisfying all of the practical requirements involved. When it rejects the homogeneity hypothesis, it gives no decisions as to which of the differences among the treatment means may be considered significant and which may not.

To illustrate, Table I shows results of a barley grain yield experiment conducted by E. Shulkcum of this Institute at Accomac, Virginia, in 1951. Seven varieties, A, B, $\cdots$, G, were replicated six times in a randomized block design. The F ratio (in section b) for testing the homogeneity of the varietal means is highly significant. This indicates that one or more of the differences among the means are significant but it does not specify which ones.

TABLE I. BARLEY GRAIN YIELDS IN BUSHELS PER ACRE

a) *Varietal Means Ranked in Order*

A	F	G	D	C	B	E
49.6	58.1	61.0	61.5	67.6	71.2	71.3

b) *Analysis of Variance*

Source	*d.f.*	*m.s.*	*F*
Between varieties	6	366.97	4.61**
Between blocks	5	141.95	
Error	30	79.64	

c) *Standard Error of a Varietal Mean*

$$s_m = \sqrt{79.64/6} = 3.643 \qquad (n_2 = 30)$$

The problem we wish to consider is that of testing these differences more specifically. Several test procedures have been proposed for

*Sponsored by the Office of Ordnance Research, U. S. Army, under Contract DA-36-034-ORD-1477, Technical Reports Nos. 3 (June 1953), 6 (September 1953) and 9 (July 1954).

**Now at the University of Florida.

answering this problem. The simplest of these is one which is often termed the *least-significant-difference* (or *L.S.D.*) *test*. This has developed from a brief discussion of the problem by R. A. Fisher (9, section 24) and is described in detail by several authors, for example, Paterson (14, pp. 38-42) and Davies (4, section 5.28). In this test, the difference between any two means is declared significant, at the 5% level, say, if it exceeds a so-called *least significant difference* $\sqrt{2}\ ts_m$ (t being the 5% level significant value from the t distribution), and provided also that the F test for the homogeneity of the n means involved is significant. If the F test is not significant, none of the differences is significant irrespective of its magnitude relative to the least significant difference.

Many other tests have also been proposed for solving this problem, including several put forward within the last year or two. Further tests are being developed at the present time. Originators of these, not to mention all, include D. T. Sawkins (18), D. Newman (12), D. B. Duncan (5-8), J. W. Tukey (21-23), H. Scheffé (19), M. Keuls (10), S. N. Roy, R. C. Bose (17), H. O. Hartley (25), and J. Cornfield, M. Halperin, S. Greenhouse (3). Unfortunately, these tests vary considerably and it is difficult for the user to decide which one to choose for any given problem.

One objective of this paper is to consider several of the procedures which have been proposed and to illustrate their basic points of difference, using a geometric method with simple cases involving only three means. A second objective is to present certain simple extensions of the concepts of power and significance which are useful in analyzing these procedures. The development of the simple case examples and the latter general concepts will point the way to a clearer evaluation of the relative properties and merits of the procedures in general and should help the user in making a choice among the available procedures. The final objective is to present a new multiple range test (8) which combines the features considered to be the best from the previously proposed tests.

2. THE NEW MULTIPLE RANGE TEST

Before discussing the general problem in more detail, it may be helpful to look ahead at an example of the application of one of the tests. An example of the proposed new test will be used for this purpose. This *new multiple range test*, as it will be termed, combines the simplicity and speed of application of a test proposed by Newman (12) and Keuls (10) with most of the power advantages of the multiple comparisons test previously proposed by the author (6, 7). For the example, we shall consider the application of a 5% level test to the varietal yield means in Table I.

TABLE II. SIGNIFICANT STUDENTIZED RANGES FOR A 5% LEVEL NEW* MULTIPLE RANGE TEST

n_2 \ p	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
1	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0
2	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09
3	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
4	3.93	4.01	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
5	3.64	3.74	3.79	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
6	3.46	3.58	3.64	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68
7	3.35	3.47	3.54	3.58	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.61
8	3.26	3.39	3.47	3.52	3.55	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
9	3.20	3.34	3.41	3.47	3.50	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52
10	3.15	3.30	3.37	3.43	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47	3.47	3.48	3.48	3.48
11	3.11	3.27	3.35	3.39	3.43	3.44	3.45	3.46	3.46	3.46	3.46	3.46	3.47	3.48	3.48	3.48
12	3.08	3.23	3.33	3.36	3.40	3.42	3.44	3.44	3.46	3.46	3.46	3.46	3.47	3.48	3.48	3.48
13	3.06	3.21	3.30	3.35	3.38	3.41	3.42	3.44	3.45	3.45	3.46	3.46	3.47	3.47	3.47	3.47
14	3.03	3.18	3.27	3.33	3.37	3.39	3.41	3.42	3.44	3.45	3.46	3.46	3.47	3.47	3.47	3.47
15	3.01	3.16	3.25	3.31	3.36	3.38	3.40	3.42	3.43	3.44	3.45	3.46	3.47	3.47	3.47	3.47
16	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.44	3.45	3.46	3.47	3.47	3.47	3.47
17	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.44	3.45	3.46	3.47	3.47	3.47	3.47
18	2.97	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.43	3.45	3.46	3.47	3.47	3.47	3.47
19	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.43	3.44	3.46	3.47	3.47	3.47	3.47
20	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.43	3.44	3.46	3.46	3.47	3.47	3.47
22	2.93	3.08	3.17	3.24	3.29	3.32	3.35	3.37	3.39	3.42	3.44	3.45	3.46	3.47	3.47	3.47
24	2.92	3.07	3.15	3.22	3.28	3.31	3.34	3.37	3.38	3.41	3.44	3.45	3.46	3.47	3.47	3.47
26	2.91	3.06	3.14	3.21	3.27	3.30	3.34	3.36	3.38	3.41	3.43	3.45	3.46	3.47	3.47	3.47
28	2.90	3.04	3.13	3.20	3.26	3.30	3.33	3.35	3.37	3.40	3.43	3.45	3.46	3.47	3.47	3.47
30	2.89	3.04	3.12	3.20	3.25	3.29	3.32	3.35	3.37	3.40	3.43	3.44	3.46	3.47	3.47	3.47
40	2.86	3.01	3.10	3.17	3.22	3.27	3.30	3.33	3.35	3.39	3.42	3.44	3.46	3.47	3.47	3.47
60	2.83	2.98	3.08	3.14	3.20	3.24	3.28	3.31	3.33	3.37	3.40	3.43	3.45	3.47	3.48	3.48
100	2.80	2.95	3.05	3.12	3.18	3.22	3.26	3.29	3.32	3.36	3.40	3.42	3.45	3.47	3.53	3.53
∞	2.77	2.92	3.02	3.09	3.15	3.19	3.23	3.26	3.29	3.34	3.38	3.41	3.44	3.47	3.61	3.67

*Using special protection levels based on degrees of freedom.

TABLE III. SIGNIFICANT STUDENTIZED RANGES FOR A 1% LEVEL NEW* MULTIPLE RANGE TEST

n_2 \ p	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
1	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
2	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0
3	8.26	8.5	8.6	8.7	8.8	8.9	8.9	9.0	9.0	9.0	9.1	9.2	9.3	9.3	9.3	9.3
4	6.51	6.8	6.9	7.0	7.1	7.1	7.2	7.2	7.3	7.3	7.4	7.4	7.5	7.5	7.5	7.5
5	5.70	5.96	6.11	6.18	6.26	6.33	6.40	6.44	6.5	6.6	6.6	6.7	6.7	6.8	6.8	6.8
6	5.24	5.51	5.65	5.73	5.81	5.88	5.95	6.00	6.0	6.1	6.2	6.2	6.3	6.3	6.3	6.3
7	4.95	5.22	5.37	5.45	5.53	5.61	5.69	5.73	5.8	5.8	5.9	5.9	6.0	6.0	6.0	6.0
8	4.74	5.00	5.14	5.23	5.32	5.40	5.47	5.51	5.5	5.6	5.7	5.7	5.8	5.8	5.8	5.8
9	4.60	4.86	4.99	5.08	5.17	5.25	5.32	5.36	5.4	5.5	5.5	5.6	5.7	5.7	5.7	5.7
10	4.48	4.73	4.88	4.96	5.06	5.13	5.20	5.24	5.28	5.36	5.42	5.48	5.54	5.55	5.55	5.55
11	4.39	4.63	4.77	4.86	4.94	5.01	5.06	5.12	5.15	5.24	5.28	5.34	5.38	5.39	5.39	5.39
12	4.32	4.55	4.68	4.76	4.84	4.92	4.96	5.02	5.07	5.13	5.17	5.22	5.24	5.26	5.26	5.26
13	4.26	4.48	4.62	4.69	4.74	4.84	4.88	4.94	4.98	5.04	5.08	5.13	5.14	5.15	5.15	5.15
14	4.21	4.42	4.55	4.63	4.70	4.78	4.83	4.87	4.91	4.96	5.00	5.04	5.06	5.07	5.07	5.07
15	4.17	4.37	4.50	4.58	4.64	4.72	4.77	4.81	4.84	4.90	4.94	4.97	4.99	5.00	5.00	5.00
16	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.84	4.88	4.91	4.93	4.94	4.94	4.94
17	4.10	4.30	4.41	4.50	4.56	4.63	4.68	4.72	4.75	4.80	4.83	4.86	4.88	4.89	4.89	4.89
18	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.76	4.79	4.82	4.84	4.85	4.85	4.85
19	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.72	4.76	4.79	4.81	4.82	4.82	4.82
20	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.69	4.73	4.76	4.78	4.79	4.79	4.79
22	3.99	4.17	4.28	4.36	4.42	4.48	4.53	4.57	4.60	4.65	4.68	4.71	4.74	4.75	4.75	4.75
24	3.96	4.14	4.24	4.33	4.39	4.44	4.49	4.53	4.57	4.62	4.64	4.67	4.70	4.72	4.74	4.74
26	3.93	4.11	4.21	4.30	4.36	4.41	4.46	4.50	4.53	4.58	4.62	4.65	4.67	4.69	4.73	4.73
28	3.91	4.08	4.18	4.28	4.34	4.39	4.43	4.47	4.51	4.56	4.60	4.62	4.65	4.67	4.72	4.72
30	3.89	4.06	4.16	4.22	4.32	4.36	4.41	4.45	4.48	4.54	4.58	4.61	4.63	4.65	4.71	4.71
40	3.82	3.99	4.10	4.17	4.24	4.30	4.34	4.37	4.41	4.46	4.51	4.54	4.57	4.59	4.69	4.69
60	3.76	3.92	4.03	4.12	4.17	4.23	4.27	4.31	4.34	4.39	4.44	4.47	4.50	4.53	4.66	4.66
100	3.71	3.86	3.98	4.06	4.11	4.17	4.21	4.25	4.29	4.35	4.38	4.42	4.45	4.48	4.64	4.65
∞	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20	4.26	4.31	4.34	4.38	4.41	4.60	4.68

*Using special protection levels based on degrees of freedom.

The data necessary to perform the test are: (a) the means as shown in Table I; (b) the standard error of each mean, $s_m = 3.643$ and (c) the degrees of freedom on which this standard error is based, $n_2 = 30$.

First, a table (Table II) of special *significant studentized ranges* for a 5% level test is entered at the row for $n_2 = 30$ degrees of freedom, and significant studentized ranges are extracted for samples of sizes $p = 2, 3, 4, 5, 6$ and 7. The values obtained in this way are 2.89, 3.04, 3.12, 3.20, 3.25 and 3.29 respectively. (Table III shows the significant studentized ranges which would be used for a 1% level test.)

The significant studentized ranges are then each multiplied by the standard error, $s_m = 3.643$, to form what may be called *shortest significant ranges*. The shortest significant ranges R_2, R_3, $\cdots$, R_7 are recorded at the top of a worksheet as shown in Table IV.

As a final preparatory step it is convenient to display the means in ranked order from left to right, spaced so that the distances between them are very roughly proportional to their numerical differences. This may be done on the worksheet immediately under the shortest significant ranges as in Table IV. The lines underscoring the means indicate the results and are added as the test proceeds.

TABLE IV. WORKSHEET

a) *Shortest Significant Ranges*

p:	(2)	(3)	(4)	(5)	(6)	(7)
R_p :	10.53	11.07	11.37	11.66	11.84	11.99

b) *Results*

Varieties:	A	F	G	D	C	B	E
Means:	49.6	58.1	61.0	61.5	67.6	71.2	71.3
	───	───	───	───	───	───	───
		───	───	───	───		

Note: Any two means *not underscored* by the same line are *significantly different.*

Any two means *underscored* by the same line are *not significantly different.*

We now set out to test the differences in the following order: the largest minus the smallest, the largest minus the second smallest, up to the largest minus the second largest; then the second largest minus the smallest, the second largest minus the second smallest, and so on, finishing with the second smallest minus the smallest. Thus, in the case of this example the order for testing is: $E - A$, $E - F$, $E - G$, $E - D$, $E - C$, $E - B$; $B - A$, $B - F$, $B - G$, $B - D$, $B - C$; $C - A$, $C - F$, $C - G$, $C - D$; $D - A$, $D - F$, $D - G$; $G - A$, $G - F$; and finally $F - A$.

With only one exception, given below, *each difference is significant if it exceeds the corresponding shortest significant range*; *otherwise it is not significant.* Because $E - A$ is the range of seven means, it must exceed $R_7 = 11.99$, the shortest significant range of seven means, to be significant; because $E - F$ is the range of six means, it must exceed $R_6 = 11.84$, the shortest significant range for six means, to be significant; and so on. *Exception*: The sole exception to this rule is that *no difference between two means can be declared significant if the two means concerned are both contained in a subset* of the means which has a non-significant range.*

Because of this exception, as soon as a non-significant difference is found between two means, it is convenient to group these two means and all of the intervening means together by underscoring them with a line, as shown for the means $\{G, D, C, B, E\}$, for example, in Table IV. The remaining differences between all members of a subset underscored in this way are not significant according to the exception rule. Thus they need not, and should not, be tested against shortest significant ranges.

The details of the test are as follows:

1) $E - A = 21.7 > 11.99$; thus $E - A$ is significant.

2) $E - F = 13.2 > 11.84$; thus $E - F$ is significant.

3) $E - G = 10.3 < 11.66$; thus $E - G$ is not significant, and hence $E - D, E - C, E - B$; $B - G, B - D, B - C$; $C - G, C - D$; and $D - G$ are not significant by the exception rule. These results are all denoted by drawing the line under the subset $\{G, D, C, B, E\}$.

4) $B - A = 21.6 > 11.84$; thus $B - A$ is significant.

5) $B - F = 13.1 > 11.66$; thus $B - F$ is significant.

6) $B - G, B - D, B - C$; $C - G, C - D$; and $D - G$ are not significant from step 3. No line need be added to show this because of the line under $\{G, D, C, B, E\}$ already.

7) $C - A = 18.0 > 11.66$; thus $C - A$ is significant.

8) $C - F - 9.5 < 11.37$; thus $C - F$ is not significant; and $C - G$, $C - D$; $D - F, D - G$; and $G - F$ are not significant by the exception rule. These results are all denoted by drawing the line under the subset $\{F, G, D, C\}$.

9) $D - A = 11.9 > 11.37$; thus $D - A$ is significant.

10) $D - F$ is not significant from step 8 and $D - G$ is not significant from step 3 or 8.

11) $G - A = 11.4 > 11.07$; thus $G - A$ is significant.

12) $G - F$ is not significant from step 8.

*The term *subset* will be used to include the complete set where necessary, as is the case here.

13) $F - A = 8.5 < 10.53$; thus $F - A$ is not significant. The result is denoted by drawing the line under $\{A, F\}$.

Each of the steps can be done almost by inspection and the complete test takes very little time. All that is necessary for a complete recording of the result is the array of means with the lines underneath, together with the brief statement giving their interpretation, as shown in section b of Table IV.

In practice there is a short cut which can be used repeatedly to good advantage, especially when the number of means is large. Instead of starting by finding the difference $E - A$, subtract the shortest significant range for seven means from the top mean E. This gives $71.3 - 11.99 = 59.31$. Since A and F are each less than 59.31, it follows that $E - A$ and $E - F$ are both significant. This is so because the shortest significant ranges R_p become smaller with decreases in the subset size p. This takes care of steps 1 and 2 in one operation. The same idea can be used repeatedly throughout the complete application and may often eliminate many steps at a time especially in a case with a large number of means.

The foregoing provides a brief introduction to many of the features of the problem involved as well as an illustration of the proposed new multiple range test. We now begin afresh considering matters in more detail.

3. GENERAL ASSUMPTIONS AND DECISIONS

In the general problem we are given a sample of observed means, $m_1, m_2, \cdots, m_n$, which are assumed to have been drawn independently from n normal populations with "true" means, $\mu_1, \mu_2, \cdots, \mu_n$, respectively, and a common standard error σ_m. This standard error is unknown, but there is available the usual estimate s_m, which is independent of the observed means and is based on a number of degrees of freedom, denoted by n_2. (More precisely, s_m has the property that $n_2 s_m^2/\sigma_m^2$ is distributed as χ^2 with n_2 degrees of freedom, independently of $m_1, m_2, \cdots, m_n$.)

In the simplest case, with only two means m_1 and m_2, there are three possible decisions. These are:

1) *m_1 is significantly less than m_2;*

2) *m_1 and m_2 are not significantly different;*

3) *m_2 is significantly less than m_1.*

It is convenient to denote these decisions by (1, 2), $(\underline{1, 2})$, and (2, 1), respectively. The order of the numbers in each pair of parentheses indicates the ranking of the means except when underscored, in which case the means are not ranked.

In passing it should be noted that we do not intend to restrict consideration, as some writers have done, for example R. E. Bechhofer (1), to problems in which the middle decision $(\underline{1, 2})$ is eliminated and the investigator is obliged to make one of the two positive decisions (1, 2) or (2, 1). Problems of this type and their extensions to cases involving more than two means may be regarded as special cases of the problems treated here in which the significance level is fixed at 100% instead of the usual 5% or 1% level.

In the case $n = 3$, with three means, m_1 , m_2 , and m_3 , there are 19 possible decisions. These comprise:

a) Six decisions of the form: "*m_1 is significantly less than m_2 , m_2 is significantly less than m_3 , and m_1 is significantly less than m_3 .*" This joint decision may be conveniently denoted by (1, 2, 3). The remaining five denoted in the same way are (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), and (3, 2, 1).

b) Three decisions of the form: "*m_1 is significantly less than m_2 and m_3 , but m_2 and m_3 are not significantly different from one another.*" This joint decision may be denoted by $(1, \underline{2, 3})$. The remaining two denoted in the same way are $(2, \underline{1, 3})$ and $(3, \underline{1, 2})$.

c) Three decisions of the form: "*m_1 and m_2 are significantly less than m_3 , but m_1 and m_2 are not significantly different from one another.*" This one may be denoted by $(\underline{1, 2}, 3)$ and the remaining two in a similar way by $(\underline{1, 3}, 2)$ and $(\underline{2, 3}, 1)$.

d) Six decisions of the form: "*m_1 is significantly less than m_3 , but m_1 and m_2 are not significantly different from one another, and m_2 and m_3 are not significantly different from one another.*" This decision may be denoted by $(\underline{1, \underline{2}}\underline{, 3})$ and the remainder by $(\underline{1, \underline{3}}\underline{, 2})$, $(\underline{2, \underline{1}}\underline{, 3})$, $(\underline{2, \underline{3}}\underline{, 1})$, $(\underline{3, \underline{1}}\underline{, 2})$, and $(\underline{3, \underline{2}}\underline{, 1})$.

e) One decision stating: "*m_1 , m_2 , and m_3 are not significantly different from one another,*" which may be denoted by $(\underline{1, 2, 3})$.

The number of decisions increases very rapidly as n increases. In the general case with n means there are $n!$ decisions of the form $(1, 2, \cdots , n)$ with no underscoring, $(n - 1)n!/2$ decisions of the form $(\underline{1, 2}, 3, \cdots , n)$ with one pair of means underscored, $(n - 2)n!/3!$ decisions of the form $(\underline{1, 2, 3}, 4, \cdots , n)$ with three means underscored, $\cdots$, $(n - 2)n!$ decisions of the form $(\underline{1, \underline{2}}\underline{, 3}, 4, \cdots , n)$ with two overlapping pair of means underscored, and so on through often large numbers of many forms finishing with one decision of the form $(\underline{1, 2, \cdots , n})$ in which all means are underscored with the one line. The underscoring has the same interpretation as before, for example $(\underline{1, 2, \cdots , n})$ is the decision that the means m_1 , m_2 , $\cdots$, m_n are not significantly different from one another.

The statements of the respective decisions may alternatively be made in terms of the true means, $\mu_1, \mu_2, \cdots, \mu_n$. The statement, "$m_i$ is significantly less than m_j," is equivalent to the statement, "μ_i is less than μ_j." Thus, the decision (1, 2, 3), for example, implies the acceptance of the hypothesis that $\mu_1 < \mu_2 < \mu_3$. The statement, "m_i and m_j are not significantly different," is equivalent to the statement "*μ_i is unranked relative to μ_j*," where this is taken to mean that there is insufficient evidence to tell whether μ_i is less than, equal to, or greater than μ_j. Thus the decision $(2, \underline{1, 3})$, for example, consists of accepting the hypothesis that "$\mu_2 < \mu_1$, $\mu_2 < \mu_3$, but μ_1 is unranked relative to μ_3."

4. CONCEPTS OF POWER AND SIGNIFICANCE

4.1 *Power Functions.*

In analysing the power of these tests we are first faced with the difficulty that none of them, not even in the simplest case involving only two means, is a two-decision procedure, whereas a power function as defined by Neyman and Pearson (13) is strictly a two-decision-test concept.

In the three-decision test in the simplest case of two means, one way of avoiding this difficulty is to group the decisions (1, 2) and (2, 1) together as the decision that m_1 and m_2 are significantly different, or in other words as acceptance of the hypothesis $\mu_1 \neq \mu_2$. A convenient notation for this decision is $(1 \neq 2)$. The given three-decision test is reduced in this way to a two-decision procedure with decisions $(\underline{1, 2})$ and $(1 \neq 2)$ and as such may be analysed as an α-level test of $\mu_1 = \mu_2$ against the two-sided alternative $\mu_1 \neq \mu_2$. The power function obtained in this way is given by the probability of the decision $(1 \neq 2)$ expressed as a function of the true difference $\epsilon = \mu_1 - \mu_2$. This may be conveniently denoted by $p(1 \neq 2)$, thus

$$p(1 \neq 2) = P[\text{dec. } (1 \neq 2) \mid \epsilon, \sigma^2].$$

An example of $p(1 \neq 2)$ is illustrated by the familiar curve shown by the dotted line in Figure 1b.

Although $p(1 \neq 2)$ is a most desirable function for measuring the properties of a test of $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$ it has a serious weakness for measuring the properties of a three-decision test of two means. By pooling the probabilities of the two decisions (1, 2) and (2, 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 mean as the truth may be), with the probability of the most incorrect decision (that μ_1 is the higher mean when in fact μ_2 is, or that μ_2 is the higher mean when in fact μ_1 is). A function which combines probabilities of correct decisions with probabilities of serious errors in this way, is of no value in measuring desirable or undesirable properties. For this reason $p(1 \neq 2)$ will not be used as a measure of power in this problem. It has been discussed only because this function is so familiar that otherwise readers might have expected to have seen it used.

A more useful analysis of a three-decision test of two means is one which treats it as the joint application of two two-decision tests, namely, a test of the hypothesis, $\mu_1 \leq \mu_2$ against the alternative $\mu_2 < \mu_1$, and a test of the hypothesis $\mu_2 \leq \mu_1$ against the alternative $\mu_1 < \mu_2$. This type of analysis, which is suggested in a more general form by Lehmann (11, section 11), avoids the difficulties inherent in the $p(1 \neq 2)$ function, and extends readily to cases with more than two means.

From this point of view, a three-decision test has two power functions

$$p(2, 1) = P[\text{dec. } (2, 1) \mid \epsilon, \sigma^2]$$

and

$$p(1, 2) = P[\text{dec. } (1, 2) \mid \epsilon, \sigma^2],$$

which are the Neyman-Pearson power functions of the tests of $\mu_1 \leq \mu_2$ and $\mu_2 \leq \mu_1$ respectively. Examples of these functions are illustrated by the sigmoid and the reverse-sigmoid curves respectively in Figure 1b. Each of these functions has the merit that for any given value of the true difference ϵ, the function gives the probability of a correct *or* incorrect decision, and it is therefore clear whether the function should be as high or as low as possible. For example, $p(2, 1)$ represents the probability of deciding that μ_1 is the higher mean. Clearly then, it will be desirable for $p(2, 1)$ to be as high as possible for $\epsilon = \mu_1 - \mu_2 > 0$, and to be as low as possible for $\epsilon \leq 0$.

In the general case of n means we shall use ${}_nP_2$ power functions of the form

$$p(i, j) = P[\text{dec. } (i, j) \mid \mu_1, \mu_2, \cdots, \mu_n, \sigma^2],$$

where decision (i, j) includes all decisions which rank μ_i lower than μ_j; and $i, j = 1, 2, \cdots, n$; $i \neq j$. Each function $p(i, j)$ is the Neyman-Pearson power function of the test of the hypothesis $\mu_j \leq \mu_i$ against the alternative $\mu_i < \mu_j$. In general, therefore, $p(i, j)$ measures the probability of a correct decision with respect to μ_i and μ_j, over all values of the true means for which $\mu_i < \mu_j$, and the probability of a wrong decision over all values of the means for which $\mu_j \leq \mu_i$.

This approach is greatly simplified in all tests we wish to consider as a result of the reasonable symmetry restriction that all test properties be invariant under all $n!$ permutations of the true means. In other words any test we consider must have the same properties for any set of values of the means irrespective of the identification of (the varieties represented by) the given means. Under these conditions it is necessary to investigate only one of the power functions $p(i, j)$ in order to investigate them all. An example of this is shown by the symmetry of $p(2, 1)$ and $p(1, 2)$ in Figure 1b.

4.2 *Significance Levels.*

So far as joint test properties are concerned only a relatively small number of significance levels need be considered. These are chosen so as to be as few in number as possible and yet have the property that once they are fixed at appropriate values, the merits of a test can then be judged solely in terms of its individual power functions.

In the simplest case involving only two means the significance levels or maximum type 1 error probabilities of the tests of $\mu_1 \leq \mu_2$ and $\mu_2 \leq \mu_1$ considered individually both occur when $\mu_1 = \mu_2$ and, by symmetry, these levels are equal. Because of this, only one significance level need be considered for the joint test, and this level may be taken as

$$\alpha = P[\text{dec. } (1 \neq 2) \mid \mu_1 = \mu_2],$$

which is the familiar significance level of the Neyman-Pearson test of $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$. Given that α is fixed at α_0 the significance levels of the individual tests must be $\frac{1}{2}\alpha_0$ each.

In further discussion a type 1 error in a test of $\mu_i \leq \mu_j$, namely the decision (j, i) in cases where $\mu_i \leq \mu_j$, may be usefully termed an *error of wrong ranking* or the finding of a *wrong significant difference.* The importance of fixing α at α_0 may then be said to rest, not so much on the fact that the probability of a wrong ranking when $\mu_1 - \mu_2 = 0$ has been fixed at α_0, but on the fact that the probability of a wrong ranking at any value of the difference $\mu_1 - \mu_2$ cannot exceed α_0.

Any test for the case of three means may be regarded as having four significance levels of a nature similar to the significance level of a two-mean test. Three of these are of the form

$$\alpha(1, 2) = \text{maximum } P[\text{dec. } (1 \neq 2) \mid \mu_1 = \mu_2],$$

where the decision $(1 \neq 2)$ includes all decisions which rank μ_1 above or below μ_2 and the maximization is taken over all possible values of the true means μ_1, μ_2 and μ_3 for which $\mu_1 = \mu_2$. The level $\alpha(1, 2)$ is, moreover, the maximum value of the probability of making a wrong ranking of μ_1 and μ_2 over all possible values of the true means. The remaining two levels of this same form are

$$\alpha(1, 3) = \text{maximum } P[\text{dec. } (1 \neq 3) \mid \mu_1 = \mu_3],$$

$$\alpha(2, 3) = \text{maximum } P[\text{dec. } (2 \neq 3) \mid \mu_2 = \mu_3],$$

and are the maximum probabilities of making a wrong ranking between μ_1 and μ_3 and between μ_2 and μ_3 in a similar way.

The fourth significance level involves all three means and is defined as

$$\alpha(1, 2, 3) = P[\text{dec. } \overline{(1, 2, 3)} \mid \mu_1 = \mu_2 = \mu_3],$$

where the decision $\overline{(1, 2, 3)}$ includes all decisions which rank at least one pair of the means relative to one another. In other words, decision $\overline{(1, 2, 3)}$ includes all the 19 decisions previously listed except decision $(\underline{1,\ 2,\ 3})$. This three-mean significance level is simply the probability of finding at least one wrong significant difference between m_1, m_2 and m_3, that is, of making at least one wrong ranking of any pair of the true means μ_1, μ_2, and μ_3.

In the case of four means there are eleven significance levels which may be defined in a similar way. Six of these are two-mean significance levels of the form

$$\alpha(1, 2) = \text{maximum } P[\text{dec. } (1 \neq 2) \mid \mu_1 = \mu_2],$$

where, as before, the decision $(1 \neq 2)$ includes all decisions ranking μ_1 and μ_2 relative to one another, and the maximization is taken over all values of the means μ_1, μ_2, μ_3 and μ_4 for which $\mu_1 = \mu_2$. The remaining five two-mean significance levels defined in a similar way are $\alpha(1, 3)$, $\alpha(1, 4)$, $\alpha(2, 3)$, $\alpha(2, 4)$ and $\alpha(3, 4)$.

Four of the levels in this case are three-mean significance levels of the form

$$\alpha(1, 2, 3) = \text{maximum } P[\text{dec. } \overline{(1, 2, 3)} \mid \mu_1 = \mu_2 = \mu_3],$$

where the decision $\overline{(1, 2, 3)}$ includes all decisions which rank at least one pair of the means μ_1, μ_2 and μ_3 relative to one another, and where the maximization is taken over all values of the true means for which $\mu_1 = \mu_2 = \mu_3$. The remaining three three-mean significance levels similarly defined are $\alpha(1, 2, 4)$, $\alpha(1, 3, 4)$ and $\alpha(2, 3, 4)$.

Finally there is a single four-mean significance level defined as

$$\alpha(1, 2, 3, 4) = P[\text{dec. } \overline{(1, 2, 3, 4)} \mid \mu_1 = \mu_2 = \mu_3 = \mu_4],$$

where decision $\overline{(1, 2, 3, 4)}$ represents all decisions which rank at least one pair of the four means relative to one another. In other words decision $\overline{(1, 2, 3, 4)}$ includes all decisions except decision $(\underline{1,\ 2,\ 3,\ 4})$, which, following the previous pattern, is the decision that none of the differences among the four means is significant.

In a general test of n means, there are ${}_nC_2$ two-mean significance levels, ${}_nC_3$ three-mean significance levels, and so on up to ${}_nC_n = 1$ n-mean significance level. A p-mean significance level in general represents the maximum probability of finding at least one wrong significant difference among p observed means.

On careful consideration it appears that all* errors of wrong ranking in a test of n means can be adequately controlled by fixing these significance levels at appropriate values. The problem of finding a good test is then reduced to finding a procedure which optimizes the power functions $p(i, j)$ given that these significance levels are fixed at the chosen values.

4.3 *Protection Levels.*

The complement of any p-mean significance level may be termed a *p-mean protection level,* and is the minimum probability of finding no wrong significant differences among p observed means. The name "protection level" is suitable in that the level measures protection against finding wrong significant differences.

Thus, in a two-mean test, there is one protection level

$$\gamma = P[\text{dec. } (\underline{1, 2}) \mid \mu_1 = \mu_2] = 1 - \alpha.$$

If the significance level is 5%, for example, the protection level is 95%.

In a three-mean test, there are three two-mean protection levels $\gamma(1, 2)$, $\gamma(1, 3)$ and $\gamma(2, 3)$, where, for example,

$$\gamma(1, 2) = \text{minimum } P[\text{dec. } (\underline{1, 2}) \mid \mu_1 = \mu_2] = 1 - \alpha(1, 2)$$

and decision $(\underline{1, 2})$ includes all decisions for which μ_1 and μ_2 are not ranked relative to one another. In addition there is one three-mean protection level

$$\gamma(1, 2, 3) = P[\text{dec. } (\underline{1, 2, 3}) \mid \mu_1 = \mu_2 = \mu_3] = 1 - \alpha(1, 2, 3).$$

In a general test of n means there are ${}_nC_p$ p-mean protection levels of the form

$$\gamma(a_1, a_2, \cdots, a_p)$$
$$= \text{minimum } P[\text{dec. } (\underline{a_1, a_2, \cdots, a_p}) \mid \mu_{a_1} = \mu_{a_2} = \cdots \mu_{a_p}]$$

where $p = 2, 3, \cdots, n$, each one being the complement of the corresponding significance level. The symbols a_1, a_2, $\cdot\cdot$, a_p stand for the subscripts identifying the particular set of p means concerned.

*See also comments on class 2 protection levels in section 5.4.4.

(Thus decision $(\underline{a_1 , a_2 , \cdots , a_p})$ represents the decision that there are no significant differences between the observed means m_{a_1}, $m_{a_2}, \cdots , m_{a_p}$).

In further discussion of the controlling of errors of wrong ranking it will be somewhat more convenient to think in terms of fixing the protection levels of a test rather than in terms of fixing the significance levels.

4.4 *Consistent Protection Levels.*

We now consider the important question: In any test of n means, given that γ_2 is an appropriate value for the two-mean protection levels, what values $\gamma_3 , \gamma_4 , \cdots , \gamma_n$ should be regarded as satisfactory for the three-mean, four-mean, etc., protection levels, and for the n-mean protection level?

First it should be noted that if a symmetric 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. For example in the case of four means when $n_2 = \infty$, a test of this type with $\gamma_2 = 95\%$ would be obtained by applying six 5% level symmetric normal-deviate tests to each of the six differences between the four means. 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 higher 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 (that is, of increases in probabilities of type 2 errors), and it is most important that the levels be raised no more than is absolutely necessary. We shall now show that there are good reasons* for raising the higher order protection levels only part of the way towards the value of the two-mean protection levels.

Suppose, for the sake of an example, that a randomized block experiment were designed for the purpose of testing (a) the difference between two varieties V_1 and V_2, (b) the difference between two fertilizers F_1 and F_2 and (c) the difference between two insect control

*See also (5, section 6) and (6, p. 177).

spray methods S_1 and S_2 . If interactions could be assumed to be zero, as might well be reasonable, a good design would be obtained by randomizing the four treatment combinations $V_1F_1S_1$, $V_1F_2S_2$, $V_2F_1S_2$ and $V_2F_2S_1$ within each block, where $V_1F_1S_1$, for example, denotes the application of fertilizer F_1 and spray method S_1 in a plot sown with variety V_1 . If the observed means of these combinations are denoted respectively by m_1 , m_2 , m_3 and m_4 , the varietal, fertilizer and spray differences would be measured respectively by the independent differences:

$$d_1 = (m_1 + m_2) - (m_3 + m_4) = m_1 + m_2 - m_3 - m_4$$

$$d_2 = (m_1 + m_3) - (m_2 + m_4) = m_1 - m_2 + m_3 - m_4$$

$$d_3 = (m_1 + m_4) - (m_2 + m_3) = m_1 - m_2 - m_3 + m_4$$

Now, provided that the number, r, of replications and hence the number of error degrees of freedom, $n_2 = 3r$, were large enough, it would be possible to make independent tests of the three given differences. Under these circumstances, if, say, a 5% level test of each difference were desired, no reasonable objection could be raised to the joint unmodified application of three 5% level tests. The joint use of these tests would be just as valid as if the differences were tested in three independent and separate experiments. In this joint test, it is clear that if the three null hypotheses in the individual tests were simultaneously true, which would imply that the true means μ_1 , μ_2 , μ_3 , and μ_4 of the four combinations were all equal, the probability of not rejecting this joint hypothesis would be $(.95)^3 = 85.7\%$. Although this value is lower than 95%, it is clearly an implicitly unobjectionable result of having chosen a 95% protection level for each of the independent tests.

Now, the error of wrongly rejecting the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$ in this type of test is no less serious than the error of rejecting the same hypothesis in the type of test under consideration, and a four-mean protection level is the probability of not making an error of this kind. Hence, it is argued that the objections to the low four-mean protection level $\gamma_4 = 79.7\%$ of the 5% level multiple normal-deviate test above would be appropriately remedied if the level were raised to $\gamma_4 = 85.7\%$.

A similar analogy with two independent 5% level tests of two independent differences among three means can be invoked for choosing an appropriate value for the three-mean protection levels in the same test. This leads to the conclusion that the objection to the low value $\gamma_3 = 87.8\%$ for these levels would be removed if they were increased to $(.95)^2 = 90.25\%$.

The same argument readily generalizes to give the result that the value $\gamma_p = \gamma_2^{p-1}$ for any p-mean protection level is appropriate in association with the value γ_2 for a two-mean protection level. The exponent $p - 1$ in these levels is given by the number of independent comparisons which can be specified, or the degrees of freedom, among the p means. For this reason the levels $\gamma_p = \gamma_2^{p-1}$ may be termed *protection levels based on degrees of freedom.*

Protection levels of this type have been used in constructing the multiple comparisons test (6, 7) and the new multiple range test. In the example of section 2 giving a 5% level new multiple range test of the seven barley variety means, the values of the protection levels are: $\gamma_2 = 95\%$, $\gamma_3 = 90.25\%$, $\gamma_4 = 85.7\%$, $\gamma_5 = 81.5\%$, $\gamma_6 = 77.4\%$ and $\gamma_7 = 73.5\%$. Since $\gamma_2 = 95\%$, we know that the probability of finding a significant difference between any two means when the corresponding true means are equal is definitely less than or equal to 5%. The higher order protection level values are in accord with this property.

In a similar 5% level test of 101 means, the first seven protection level values would be the same and the remainder would get progressively smaller down to $\gamma_{101} = (.95)^{100} = 0.6\%$ for the 101-mean protection level. Despite the independent tests analogy already given, the higher order protection levels may appear unduly low unless their progressively diminishing importance is fully realized. The appropriateness of these higher order protection levels in general will be emphasized by a further discussion of the independent tests analogy with particular reference to the justification of the 101-mean level $\gamma_{101} = 0.6\%$.

To take a corresponding analogy, suppose that in the course of a year's work, an experimenter has tested 100 separate null hypotheses $H_1, H_2, \cdots, H_{100}$ in 100 independent experiments, and that he has chosen a 5% level test in each case. Should he be alarmed over the obvious fact that *if* the 100 null hypotheses were simultaneously true there has been only a 0.6% chance of not rejecting this joint hypothesis? Clearly the answer is no, because it would be illogical to alter any given individual test for reasons entirely independent of that test.

In choosing a 5% level of significance in each test the experimenter has implicitly expressed the opinion that there is some *a priori* chance that the respective null hypothesis is not true. It can be stated as a general rule that the more one can argue against the truth of a null hypothesis on *a priori* grounds the lower, other things being equal, should be the protection level of the test, in order not to waste power in detecting the truth of the alternative hypothesis. In choosing a 5% level test which has a 95% protection level the experimenter is implicitly prepared to assume that the *a priori* probability of the null hypothesis is less than unity and lower than if, for example, he had chosen a 1% level test which has a 99% protection level.

Now, if the individual null hypotheses are independent in the sense that their *a priori* probabilities are independent, and if these probabilities are each appreciably less than unity as is implied by the choice of 5% levels of significance, the joint *a priori* probability for p such null hypotheses will be the product of the individual probabilities and will get less and less as p increases. Hence in the interests of not wasting power in detecting the truth of alternatives, it can well be appropriate to have lower and lower protection levels for each joint null hypothesis as p increases. In the case of the joint null hypothesis that all of the 100 individual null hypotheses are simultaneously true, for example, the *a priori* probability would be so small that it may be wasteful to use more than a very low protection level.

On extending this line of argument to a full average-weighted-risk analysis (24) including considerations of error weight functions and more complete Bayes (*a priori* probability) functions, the appropriateness of the overall joint test can be fully substantiated. In the full analysis the result is found to depend not directly on the independence of the Bayes functions of the individual tests, but on a closely related property, namely, the additivity of the error weight functions of the individual tests. An interesting more general form of this result, the proof and discussion of which will be presented subsequently as a separate paper, may be summarized as follows:

> Let T represent the joint test formed by k individual tests T_1, T_2, $\cdots$, T_k. Suppose that the error weight functions of the individual tests are additive in the sense that the error weight or loss for any joint decision D given any joint hypothesis H in the joint test T is equal to the sum of the error weights or losses for the decisions D_1, D_2, $\cdots$, D_k given the respective hypotheses H_1, H_2, $\cdots$, H_k, where the latter are individual test decisions and hypotheses forming D and H respectively.
>
> Then it follows, that if each individual test T_i is an optimum procedure from the point of view of minimizing average weighted risk, the joint test T is also an optimum procedure in the same sense.

Applying this to our example with 100 independent 5% level tests, we can say that since the error losses from one test to the next are additive, which is reasonable to assume because of the independent nature of the tests, and if each 5% level has been chosen as the best level to use for each test considered individually, then all features of the joint test are optimum including, among many others, the low 0.6% protection level under special consideration.

A corresponding argument may be developed concerning the higher order protection levels in a test of the differences between n means. The larger the number of means involved, the less the *a priori* chance that the means will be homogeneous and the less, therefore, the need for a high protection level. The 101-mean protection level value of 0.6% in a 5% level multiple range test of 101 means, for example, may well be an optimum value for this level because of the remoteness of the possibility that all of the 101 true means are equal.

Owing to added complexities, it has not been possible thus far to prove in complete detail that protection levels based on degrees of freedom are exactly optimum in these tests also. However, since such protection levels are optimum in sets of independent tests, and since their functions are so similar in these tests, it is safe to conclude at least that they are close to optimum, and far closer than their only proposed rivals, namely, levels which are all equal to the two-mean protection level. It therefore seems sound practice to use these levels until they can be further improved by a more thorough minimum average risk analysis.

Having defined a set of relations among the values of the p-mean protection levels of a test, we therefore need to specify only one of these values and the remainder are fixed accordingly. From a practical point of view it is most pertinent and useful to define the levels in the way adopted in the multiple comparisons test (6, 7) and retained in the new multiple range test. The example given for the latter test in section 2 is a 5% level test in the sense that its two-mean significance levels are 5% and the protection levels are $\gamma_p = (.95)^{p-1}, p = 2, 3, \cdots, 7$. Likewise in a general test of n means, an α-level test denotes a procedure in which the two-mean significance levels are α and the protection levels are $\gamma_p = (1 - \alpha)^{p-1}$, $p = 2, 3, \cdots, n$. With the significance level of a test defined in this way, all that is necessary in choosing a level for a test of a given set of n means is to choose the level which would be considered appropriate for a test of the difference between any two of the means *assuming that the remaining means were not present.* Provided an appropriate value is chosen for this level, the remaining levels in the test are automatically fixed at their correspondingly appropriate values.

5. REVIEW OF SEVERAL TESTS

Comparisons will now be made between several test procedures which have been proposed for the given problem. In most of the detailed discussion, consideration will be restricted to the following special simplifying conditions: The degrees of freedom for error will be assumed to be infinite, i.e., $n_2 = \infty$; the standard error of a mean will be assumed to be unity, i.e., $\sigma_m = 1$; and the significance level α of each test will be 5%, i.e., $\alpha = .05$. These will be referred to briefly as the special conditions $n_2 = \infty$, $\sigma_m = 1$ and $\alpha = .05$. This will provide a simple and familiar context for bringing out the main points of difference between the tests as clearly as possible. These main points are essentially unaltered when the special conditions are removed.

5.1 *The Symmetric Three-Decision t Test of Two Means.*

In the case of two means, the best test for choosing between the three possible decisions is the following familiar rule, which may be termed an *α-level symmetric three-decision t test*: *Make the decision* (1, 2) *if* $m_1 - m_2 < -\sqrt{2}t_\alpha s_m$, *the decision* $(\underline{1,\, 2})$ *if* $| m_1 - m_2 | \le \sqrt{2}t_\alpha s_m$, *or the decision* (2, 1) *if* $m_1 - m_2 > \sqrt{2}t_\alpha s_m$; where t_α is the two-tail α-level significant value of t.

Under the special conditions $n_2 = \infty$, $\sigma_m = 1$, $\alpha = .05$, the test reduces to a 5% *level symmetric three-decision normal-deviate test* and the significant difference $\sqrt{2}t_\alpha s_m = \sqrt{2}u_\alpha \sigma_m$ is the familiar value $1.960\sqrt{2} = 2.77$.

This test is satisfactory for the case of two means, and it is only when we pass on to consider tests involving more than two means that the differences arise in proposed test procedures. It is worthwhile, however, to consider various special details of an analysis of the three-decision normal-deviate test as an introduction to methods of analysing the more complex tests.

(i) *Sample Space.* A common useful method for representing this test graphically is shown in Figure 1a. In this figure, the horizontal straight line provides an example of a one-dimensional *sample space* and is used for plotting the observed difference $x = m_1 - m_2$. Any point on this line representing an observed value of x is called a *sample point.* The line is divided into three intervals, $x < -2.77$, $-2.77 \le x \le 2.77$, and $2.77 < x$. These represent the respective sets of points for which the decisions (1, 2), $(\underline{1,\, 2})$ and (2, 1) are made and are termed *decision regions.* It is convenient to denote each region by the same symbol, (1, 2), $(\underline{1,\, 2})$ or (2, 1), that is used for the corresponding decision.

(ii) *Parameter Space.* The straight line in Figure 1a may also be used for plotting values of the "true" difference, $\epsilon = \mu_1 - \mu_2$, between the true means involved. When used in this way, the line provides an example of a *parameter space*, as distinct from its function as a *sample* *space* when used for plotting x. Any point on the line representing a given value of ϵ is called a *parameter point.*

(iii) *Probability Density.* In the special case we are considering, the probability distribution function $f(x;\ \epsilon)$ of a sample point x (observed difference) about a given parameter point ϵ (given true difference) is given by a normal probability density function with mean ϵ and variance 2. For example, when $\epsilon = 0$ this function may be represented by the familiar curve shown in Figure 1a. The curve for any other value

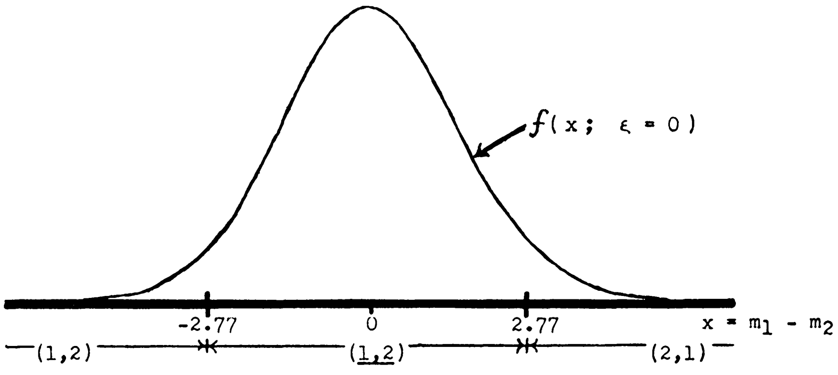


FIGURE 1a

Regions for a 5%-level symmetric three-decision normal-deviate test ($\sigma_x = \sqrt{2}$)

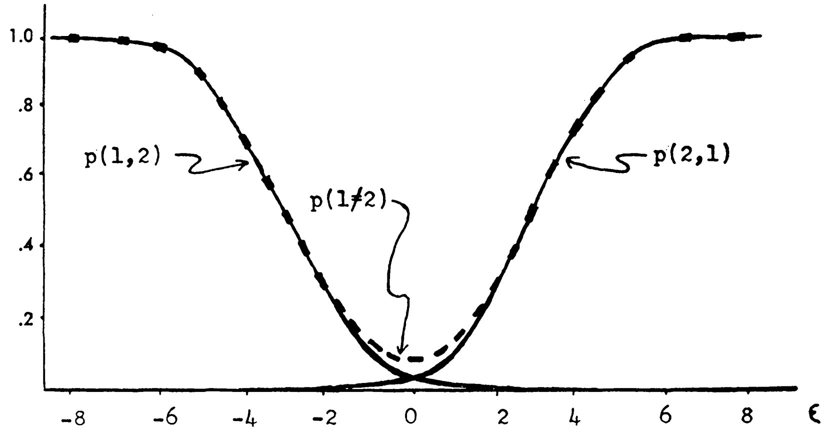


FIGURE 1b

Power Functions for 5% Level Symmetric Three-Decision Normal-Deviate Test ($\sigma_x = \sqrt{2}$)

of ϵ has the same shape and is located with its center over the given ϵ value.

(iv) *Power Functions.* The power function $p(1, 2)$ representing the probability of decision (1, 2) for any given value of ϵ is given by the area under the probability density curve for the given ϵ, over the region (1, 2). Likewise the power function $p(2, 1)$ for the same ϵ value is given by the area under the same curve and over the region (2, 1). The functions $p(1, 2)$ and $p(2, 1)$ are represented by the reverse-sigmoid and the sigmoid curves in Figure 1b.

(v) *Significance and Protection Levels.* The significance level, $\alpha = 5\%$, of this test is represented by the sum of the ordinates of the power curves in Figure 1b at $\epsilon = 0$, each of which is $2\frac{1}{2}\%$. The protection level is $1 - \alpha = 95\%$. In Figure 1a, the significance level is the sum of the areas under the dotted curve for $\epsilon = 0$, over the regions (1, 2) and (2, 1). The protection level is the area of the same curve over the region ($\underline{1, 2}$). Extensions of these familiar ideas will be useful in illustrations of corresponding features in tests of more than two means.

The virtues of the 5% level normal-deviate three-decision test can be summarized most usefully as follows: The minimum protection against making a wrong ranking of the two means is 95%, and, for all procedures for which this is true, the power curves of this test are uniformly maximized over all values of ϵ for which they measure probabilities of correct decisions, and are uniformly minimized over all values of ϵ for which they measure probabilities of incorrect decisions. This provides a good example of the general usefulness of the new multiple power function analysis which we have adopted for this and for the more complex procedures.

5.2 *Tests of Three Means. General Details.*

(i) *Sample Space.* To represent a test involving three means, m_1, m_2, and m_3, a two-dimensional sample space or plane is required in place of the one-dimensional sample space or line used above for a two-mean test. In this two-dimensional space it is convenient to plot the difference $x_1 = m_1 - m_2$ on the horizontal axis and the comparison $x_2 = (m_1 + m_2 - 2m_3)/\sqrt{3}$ on the vertical axis as rectangular Cartesian coordinates. Figures 2, 2a, 2b and 2c, and all subsequent sample space illustrations use these particular coordinates. It will be noted that x_2 is distributed independently of x_1 and has the same variance, $\sigma_x^2 = 2\sigma_m^2$. This leads to certain helpful features of symmetry which will become evident as we proceed.

Any set of values for the three differences $m_1 - m_2$, $m_1 - m_3$, and $m_2 - m_3$, between the three means, can be represented by a sample point $(x_1\ , x_2)$ in this two-dimensional sample space. For example, the set of differences $m_1 - m_2 = 4, m_1 - m_3 = -1$, and $m_2 - m_3 = -5$, found in the sample of means $m_2 = 10, m_1 = 14, m_3 = 15$, gives $x_1 = 4$ and $x_2 = -2\sqrt{3}$. These differences would thus be represented by the point $(4, -2\sqrt{3})$ located 4 units to the right of and $2\sqrt{3}$ units below the center of the space. The inverse relations by which the differences can be obtained from a sample point are $m_1 - m_2 = x_1\ , m_1 - m_3 = (x_1 + \sqrt{3}x_2)/2$, and $m_2 - m_3 = (-x_1 + \sqrt{3}x_2)/2$. Thus a point $(-2, 1)$ represents the set of differences $m_1 - m_2 = -2, m_1 - m_3 = -(2 - \sqrt{3})/2$, and $m_2 - m_3 = (2 + \sqrt{3})/2$.

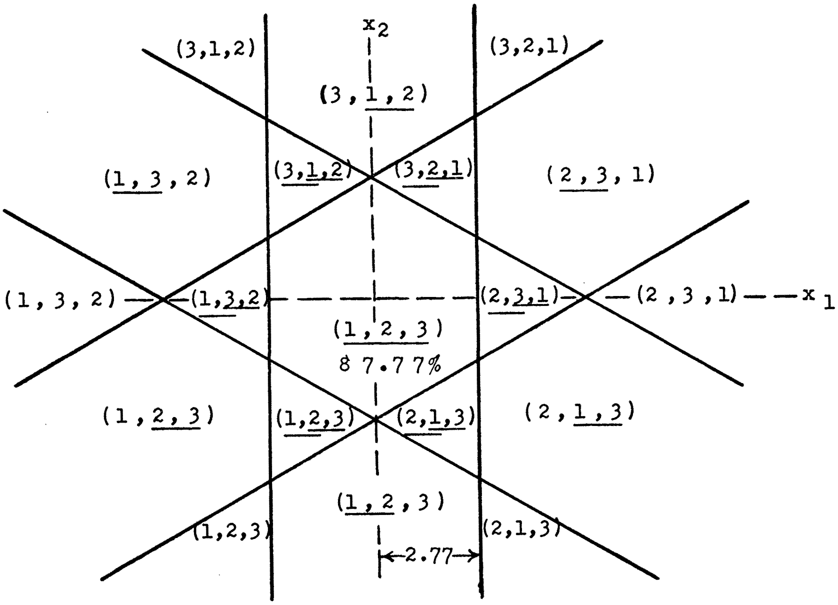
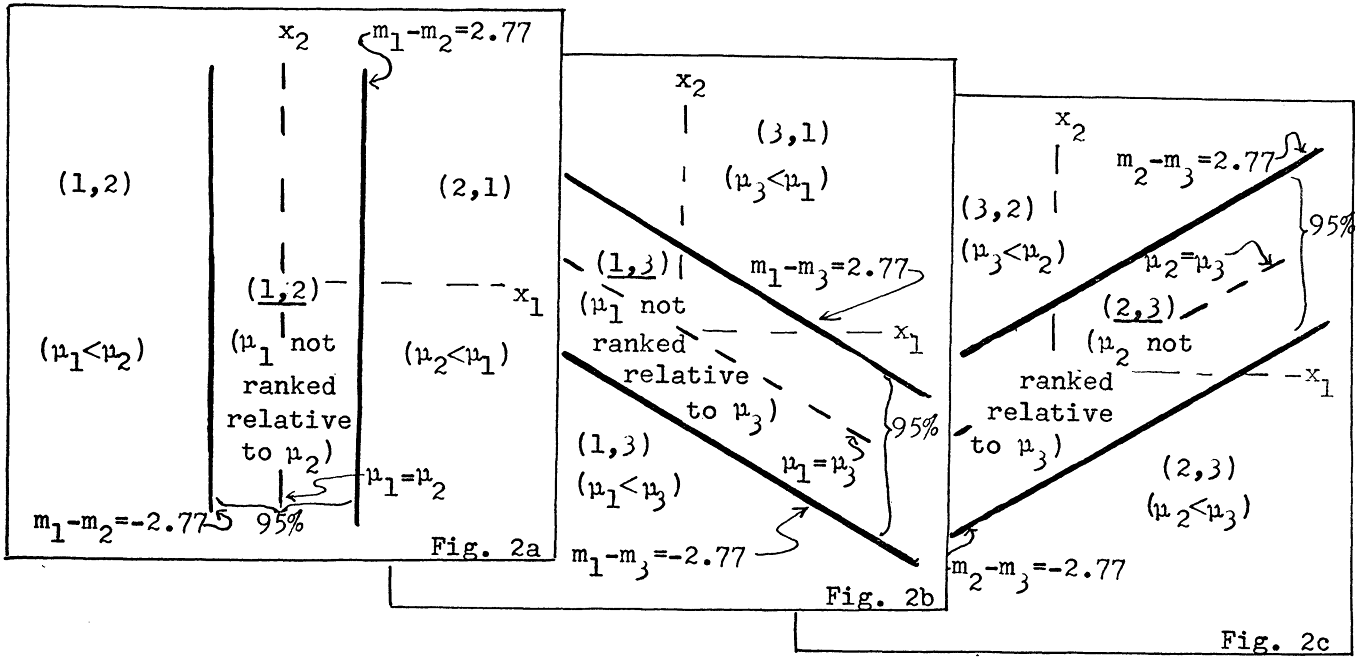


FIGURE 2

Regions of 5% Level Multiple Normal-Deviate Test ($n_2 = \infty, \sigma_m = 1$)

(ii) *Parameter Space.* The plane used as a sample space in these figures may also be used for plotting values of the "true" comparisons $\epsilon_1 = \mu_1 - \mu_2$ and $\epsilon_2 = (\mu_1 + \mu_2 - 2\mu_3)/\sqrt{3}$ between the true means involved. When used in this way it is termed a *parameter space,* and values for ϵ_1 and ϵ_2 constitute a *parameter point* $(\epsilon_1\ , \epsilon_2)$. In the parameter space we shall need to make frequent references to the parameter point $(\epsilon_1\ , \epsilon_2) = (0, 0)$, the origin, at which all true means are equal, i.e., at which $\mu_1 = \mu_2 = \mu_3$. Similarly we shall need to refer to the



FIGURES 2a, 2b, 2c

Development of Regions in Figure 2

dotted lines labelled $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, and $\mu_2 = \mu_3$ in Figures 2a, 2b, and 2c, representing all points for which $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, and $\mu_2 = \mu_3$, respectively. The position of a parameter point on any one of the lines depends on the magnitude of the third mean relative to the two equal means represented by the line.

(iii) *Probability Density.* The probability distribution of a sample point (x_1, x_2) depends only on (ϵ_1, ϵ_2) and from the definition of x_1 and x_2 it is readily seen that the distribution function $f(x_1, x_2; \epsilon_1, \epsilon_2)$ is a bivariate normal one. Each x_i is distributed normally and independently about ϵ_i as mean and with a variance of 2. The distribution for any parameter point (ϵ_1, ϵ_2) can be visualized geometrically as a bell-shaped surface standing on the sample space plane with its center located over the given parameter point.

5.3 *The Multiple t Test.*

To illustrate the way in which a test can be represented in the sample space, we shall consider a previously mentioned special case of the procedure obtained by applying an α-level symmetric three-decision t test separately to each of the hypotheses, $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, and $\mu_2 = \mu_3$. This may be termed an *α-level multiple t test,* and readily generalizes to the case of n means in which the individual t tests are applied to all ${}_nC_2$ hypotheses of the form $\mu_i = \mu_j$ which equate the means considered in all possible pairs.

As has been pointed out, this procedure does not provide a satisfactory test for our problem, and it is definitely not recommended for this purpose. We use it here and at other points in the discussion because of the excellent introduction it affords to better but more complex procedures.

Under the special conditions $n_2 = \infty$, $\sigma_m = 1$, $\alpha = .05$, the α-level multiple t test reduces to the 5% level multiple normal-deviate test. The 19 regions of this test are as shown in Figure 2.

(i) *Decision Regions.* The regions of the joint test are formed by the symmetrical intersection of three sets of two-mean test regions as shown in Figures 2a, 2b, and 2c. In Figure 2a the lines $m_1 - m_2 = -2.77$ and $m_1 - m_2 = 2.77$ divide the sample space into three regions (1, 2), $(\underline{1, 2})$, and (2, 1). The region $(\underline{1, 2})$ consists of the entire vertical strip passing down the center of the plane between the lines $m_1 - m_2 = -2.77$ and $m_1 - m_2 = 2.77$. The regions (1, 2) and (2, 1) are the remainders of the sample space plane lying to the left and right of $(\underline{1, 2})$, respectively. These are the regions of the test of $\mu_1 = \mu_2$ and are two-dimensional extensions of the corresponding one-dimensional regions in Figure 1a. The notation has the same meaning as before;

for example, if a point falls in (1, 2) the decision (1, 2) is made, namely that m_1 is significantly less than m_2 .

Likewise, the lines $m_1 - m_3 = \pm 2.77$ in Figure 2b divide the sample plane into the three regions (1, 3), $(\underline{1,\ 3})$, (3, 1) for the test of $\mu_1 = \mu_3$, and the lines $m_2 - m_3 = \pm 2.77$ in Figure 2c divide the sample plane into the three regions (2, 3), $(\underline{2,\ 3})$, (3, 2) for the test of $\mu_2 = \mu_3$. The sets of regions for each of these tests are identical with those for the test of $\mu_1 = \mu_2$, except for a rotation about the origin which is 60° counterclockwise for the first and 60° clockwise for the second.

Each of the 19 product regions for the joint test in Figure 2 corresponds to one of the 19 decisions previously listed for the case of three means. For example, in the intersection of (1, 2), (3, 1), and (3, 2) in the top left-hand corner of the figure, the associated decisions (1, 2), (3, 1), and (3, 2) constitute the joint decision (3, 1, 2). This, it will be recalled, is the decision that m_1 is significantly less than m_2 , m_3 is significantly less than m_1 , and m_3 is significantly less than m_2 . The region involved may be thus conveniently denoted as the region (3, 1, 2). Likewise the intersection of the regions $(\underline{1,\ 2})$, $(\underline{1,\ 3})$, and $(\underline{2,\ 3})$ is the hexagonal region at the center in which the decision $(\underline{1,\ 2,\ 3})$ is made. This may accordingly be denoted as the region $(\underline{1,\ 2,\ 3})$.

(ii) *Power Functions.* The power function $p(1,\ 2)$, to take one of the six power functions involved, may be visualized as a *power surface* $P[\text{dec. } (1,\ 2) \mid \epsilon_1 ,\ \epsilon_2]$ above the parameter space. The ordinate of the surface at any point $(\epsilon_1 ,\ \epsilon_2)$ is given by the integral over the region (1, 2) of the bell-shaped distribution for that point. Since the boundary of region (1, 2) is parallel to the ϵ_2 axis it is clear that sections of the power surface for different values of ϵ_2 are identical. Each section is depicted by the reverse-sigmoid $p(1,\ 2)$ curve shown for the two-mean test in Figure 1b.

The remaining power functions $p(1,\ 3)$, $p(2,\ 3)$, $p(2,\ 1)$, $p(3,\ 1)$ and $p(3,\ 2)$ may be visualized as power surfaces, identical with the surface for $p(1, 2)$, except that the one for $p(1, 3)$ is rotated 60° counterclockwise about the origin, the one for $p(2,\ 3)$ is rotated a further 60° counterclockwise about the origin, and so on.

(iii) *Protection Levels.* The two-mean protection level $\gamma(1,\ 2) =$ minimum P [dec. $(\underline{1,\ 2}) \mid \mu_1 = \mu_2]$ is the minimum integral over the strip-region $(\underline{1,\ 2})$, of any of the normal bivariate distributions centered on the line $\mu_1 = \mu_2$. Since the boundaries of $(\underline{1,\ 2})$ are parallel to the line $\mu_1 = \mu_2$, the minimum is given by the integral for any one parameter point $(0,\ \epsilon_2)$, and is 95%. The remaining two-mean protection levels $\gamma(1, 3)$ and $\gamma(2, 3)$ can be seen to be 95% in the same way.

The only remaining protection level is the three-mean level

$\gamma(1, 2, 3) = P[\text{dec. } (\underline{1, 2, 3}) \mid \mu_1 = \mu_2 = \mu_3]$. This is given by the integral over the hexagonal region $(\underline{1, 2, 3})$ of the bell-shaped bivariate normal distribution centered at the origin (0, 0). Since this region is the locus of all points for samples in which the range is less than 2.77, it follows that the integral is the probability $P[q_3 < 2.77]$, where q_p is the standardized range of a sample of p independent observations from a normal population. Tables for these probabilities are given by Pearson and Hartley (15), and from these a value of 87.8% is found for this three-mean protection level. According to the principle of protection levels based on degrees of freedom, the three-mean protection level should be 90.25%.

In the test of four means the twelve power functions are similar to those of the simpler cases in that $p(1, 2)$, for example, can be expressed as a function of $\mu_1 - \mu_2$ alone. In the reduced form $p(1, 2)$ is identical with the $p(1, 2)$ function of the two-mean test illustrated in Figure 1b. The six two-mean and four three-mean protection levels in this test are readily seen to be $P[q_2 \leq 2.77] = 95\%$ and $P[q_3 \leq 2.77] = 87.8\%$ as for the corresponding levels in the three-mean test. The four-mean protection level is similarly found to be $P[q_4 \leq 2.77] = 79.7\%$.

As has been mentioned previously, it is the lowness of the three-mean and four-mean protection levels in these tests which invalidates them as satisfactory 5% level procedures. On the other hand their power functions considered individually have all of the optimum properties of those of the two-mean test. Similar properties are possessed by α-level multiple t tests in general.

The general problem of finding a satisfactory test may be regarded as that of raising the higher order protection level values of an α-level multiple t test to acceptable values, by methods which interfere as little as possible with its optimum power functions.

5.4 *Multiple Range Tests.*

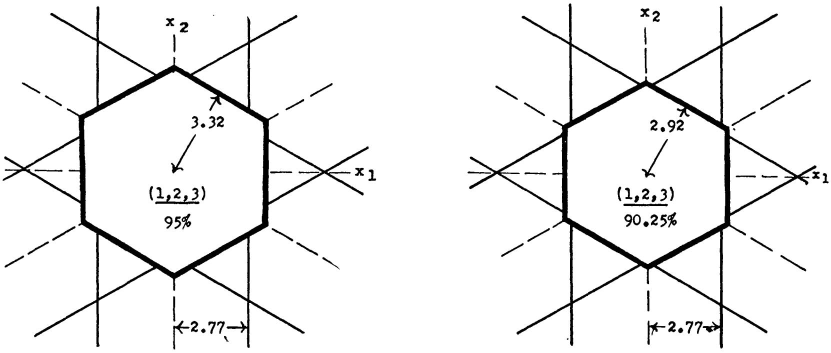
5.4.1 *The Newman-Keuls Test.*

A test proposed by Newman* (12) in 1939 and again by Keuls (10) in 1952 succeeds very simply in raising all of the low protection levels of the multiple t test. This test is equivalent to a multiple t test preceded by several preliminary range tests. Since the t tests of which the multiple t test is composed may be regarded as range tests of

*Newman mentions that the principle of this test was initially suggested to him by "Student."

subsets of two means each, the overall procedure is composed entirely of range tests and may be usefully termed a *multiple range test.*

An α-level Newman-Keuls multiple range test is given by the 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 according to an α-level range test.* Thus in the case of three means under the special conditions $n_2 = \infty$, $\sigma_m = 1$, $\alpha = .05$, the difference $m_1 - m_2$, for instance, is significant when the range of m_1, m_2, m_3 exceeds 3.32 (the 5% level value of the range of three means) and $m_1 - m_2$ exceeds 2.77. In the case of four means, $m_1 - m_2$ is significant when the range of m_1, m_2, m_3, m_4 exceeds 3.63 (the 5% level value of the range of four means), the ranges of m_1, m_2, m_3 and m_1, m_2, m_4 each exceed 3.32, and $m_1 - m_2$ exceeds 2.77.



NEWMAN-KEULS TEST (WITH CONSTANT PROTECTION LEVELS)

NEW TEST (WITH SPECIAL PROTECTION LEVELS)

FIGURE 3

5% level multiple range tests ($n_2 = \infty$, $\sigma_m = 1$)

The regions of the three-mean test are shown in Figure 3. These are the same as those of the corresponding multiple normal-deviate test except for the changes caused by the expansion of the region $(\underline{1, 2, 3})$ from a regular hexagon with radius* 2.77 to a regular hexagon with radius 3.32. This raises the three-mean protection level from 87.8% to 95%. On the other hand, the two-mean protection levels remain unaltered at 95%. For example, the level $\gamma(1, 2)$, which is the minimum integral over the modified strip region $(\underline{1, 2})$ of any distribution

*The radius of a hexagon will be used as short for the radius of the *inscribed* circle of the hexagon.

centered on the line $\epsilon_1 = \mu_1 - \mu_2 = 0$, is unchanged because the region $(\underline{1, 2})$ is unaltered away from the origin $(\epsilon_1, \epsilon_2) = (0, 0)$. The integrals are larger than 95% at the origin but drop to 95% as $|\epsilon_2|$ increases.

The six power functions are readily seen to be similar to those of the corresponding multiple normal-deviate test except for a general lowering in the area around the origin. For example, $p(1, 2)$ which is the integral over the region (1, 2) of the distribution centered at any point (ϵ_1, ϵ_2) is reduced by an amount equal to the integral over the trapezium shaped region which has been taken from (1, 2) and added to $(\underline{1, 2})$. This reduction is greatest for a distribution centered at $(\epsilon_1, \epsilon_2) = (-3.04, 0)$ (the center of the trapezium) and gets less as the distance from this point increases.

In the test of four means, the four-mean and three-mean protection levels are raised from 87.8% and 79.7% respectively to 95%, and corresponding reductions in the power functions accompany these changes.

5.4.2 *The New Multiple Range Test.*

The new multiple range test applied to the barley yield data in section 2 is a multiple range test like the Newman-Keuls procedure, except that, as has already been emphasized, it employs the special protection levels system based on degrees of freedom. A general α-level multiple range test of this type is given by the 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 means is significant according to an α_p-level range test where $\alpha_p = 1 - \gamma_p$, $\gamma_p = (1 - \alpha)^{p-1}$, and p is the number of means in the subset concerned.*

Figure 3 shows the regions of this test applied to three means under the same special conditions as before. These regions are identical with those of the corresponding Newman-Keuls test, also shown in Figure 3, except that the center hexagon has a radius of 2.92 instead of 3.32 and the adjacent regions are changed accordingly. This is sufficient to give the test a three-mean protection level of 90.25%. The two-mean protection levels remain unaltered at 95%, the same as in the Newman-Keuls test.

The power functions of this test are similar to those of the Newman-Keuls test except that the reductions relative to the multiple normal deviate test are uniformly smaller, making the test uniformly more powerful. The reductions in $p(1, 2)$, for example, are given as before by integrals over the trapezium formed by the intersection of the center hexagon $(\underline{1, 2, 3})$ with the original (1, 2) region in Figure 2a. Since the hexagon is smaller than in the previous test, the trapezium is smaller, and the reduction integrals are therefore uniformly decreased. The difference in power is greatest at a point near the center $(-3.04, 0)$ of the bigger trapezium and diminishes towards zero with increase of distance away from this point.

In the case of four means, this test raises the four-mean protection level from 79.7% to 85.7% and the three-mean levels from 87.8% to 90.25% in a similar way. The two-mean protection levels remain unaltered at 95%. Likewise the power functions are uniformly lower than those of the corresponding multiple t test but uniformly higher than those of the corresponding Newman-Keuls test.

The gains in power in the new multiple range test are quite appreciable, expecially for some parameter points and are entirely due to use of protection levels based on degrees of freedom. In passing, the independent tests analogy used in support of these new levels may be illustrated for purposes of comparison by the regions of the test shown in Figure 4. These are the regions of two 5% level independent normal deviate tests of $x_1 = m_1 - m_2$ and $x_2 = (m_1 + m_2 - 2m_3)/\sqrt{3}$ respectively, assuming $n_2 = \infty$ and $\sigma_m = 1$ as before. Tests like these would be needed, for example, if m_1 and m_2 were grain yields from two strains of one barley variety (A) and m_3 were the yield of another variety (B). Attention under these circumstances might well be restricted to testing the difference x_1 between the two strains of variety A and the difference x_2 between the two varieties A versus B.

The case for protection levels based on degrees of freedom may be put very briefly in terms of the tests illustrated in Figures 3 and 4, as follows: Because of the independence of its two component tests, the joint test in Figure 4 is a valid and acceptable joint procedure. The square region $(\underline{1, 2, 3})$ at the center of this test has the same function as the hexagonal region at the center of a multiple range test in that it is the locus of all points which do not lead to the rejection of the hypothesis $\mu_1 = \mu_2 = \mu_3$ (which implies $(\epsilon_1 , \epsilon_2) = (0, 0)$). It is adequate, therefore, to increase the dimensions of the hexagonal region in a multiple range test only so far as is needed to make the integral of the distribution at origin $(0, 0)$ over this region equal to the integral of the same distribution over the square region in Figure 4. The latter integral is 90.25% and the hexagonal region of the new multiple range test in Figure 3 has been constructed in this way.

5.4.3 *Tukey's Test Based on "Allowances."*

In 1951 Tukey (22) introduced a procedure for estimating confidence intervals, or "allowances" as he called them, for the differences $\mu_i - \mu_j$ which we have been considering. He defined a confidence coefficient

β for the joint procedure as the probability that all intervals simultaneously contain the values of the corresponding true differences. This method can be used to give, among other things, a significance test for our general problem. If, in a procedure with confidence coefficient β, the confidence interval for $\mu_i - \mu_j$ is denoted by $I_{ij}(\beta)$ this test may be expressed as the following rule: *Make the decision* (i, j) *if* $I_{ij}(\beta)$ *lies to the left of zero, the decision* $(\underline{i, j})$ *if* $I_{ij}(\beta)$ *includes zero, or the decision* (j, i) *if* $I_{ij}(\beta)$ *lies to the right of zero.* An α-level test, by the originator's definition, is obtained by putting $\beta = 1 - \alpha$.

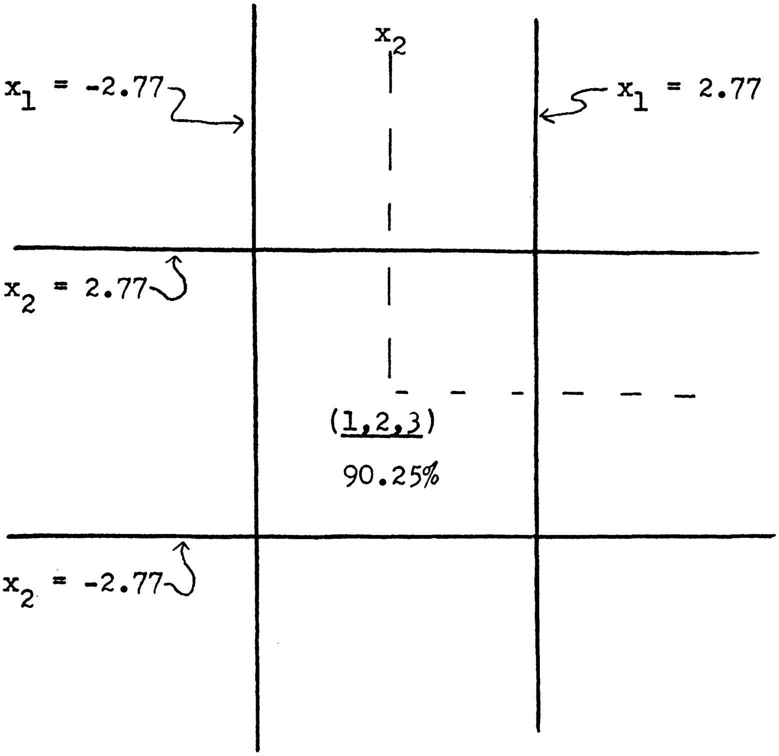


FIGURE 4

Regions for 5% Level Joint Normal-Deviate Tests of Two Independent Comparisons ($n_2 = \infty, \sigma_x = \sqrt{2}$)

The test given in this way for three means, under the special conditions $n_2 = \infty$, $\sigma_m = 1$, $\alpha = .05$, is identical with the multiple normal-deviate test shown in Figure 2 except that the width of each of the strips $(\underline{1, 2})$, $(\underline{1, 3})$, $(\underline{2, 3})$ is increased from 2 × 2.77 to 2 × 3.32. The method of derivation from confidence intervals implicitly imposes the restriction that the boundaries of $(\underline{1, 2})$, $(\underline{1, 3})$, and $(\underline{2, 3})$ must be parallel straight lines. The distance between the lines is widened until the dimensions of the center hexagon $(\underline{1, 2, 3})$ are as large as those of the Newman-Keuls test, thus making the three-mean protection level $1 - \alpha = 95\%$. At the same time the two-mean protection levels are increased uniformly from 95% to 98.1%. This test is readily seen to be more conservative and uniformly less powerful than any of the previous procedures.

5.4.4 *Tukey's 1953 Multiple Range Test.*

In 1953 Tukey (23) relaxed the conservatism of the previous test somewhat by proposing a multiple range procedure in which the significant ranges are each midway between the ones required by the test based on allowances and those required by the Newman-Keuls test. In the case of three means, under the same special conditions as before, the regions of this test are the same as those of the Newman-Keuls procedure except that the widths between the parallel lines are increased from 2.77 to $\frac{1}{2}(2.77 + 3.32) = 3.04$. The hexagon radius is 3.32 in both tests.

In suggesting this test, Tukey drew attention to an important point which may be illustrated by the following example. Suppose that in a 5% level Newman-Keuls test of four means, again assuming $n_2 = \infty$ and $\sigma_m = 1$, the values of the true means are $\mu_1 = \mu_2 = \mu$ and $\mu_3 = \mu_4 = \mu + \delta$. Suppose the difference δ between the two groups of means is so large that the preliminary range tests are practically certain to be significant, then the probability of jointly deciding that both $| m_1 - m_2 |$ and $| m_3 - m_4 |$ are not significant is $P[| m_1 - m_2 | \leq 2.77] \times P[| m_3 - m_4 | \leq 2.77] = 90.25\%$. This is an example of a whole set of levels, which we may call class 2 protection levels, which are not raised to $(1 - \alpha)$ in an α-level Newman-Keuls test and are more akin to levels based on degrees of freedom. Both of Tukey's procedures have been designed with the objective of raising these class 2 protection levels along with the others to at least $(1 - \alpha)$. The 1953 test is a modification of the test based on allowances which is uniformly more powerful than the later but which, Tukey judges, still meets his given objective.

When protection levels based on degrees of freedom are adopted, as in the new multiple range test, the class 2 levels are automatically fixed at, or slightly above (when n_2 is small), their appropriate values and need no special attention.

In the case of the Newman-Keuls procedure it is not clear whether either one of the authors was aware of the presence of these lower levels and whether he would wish to defend them as this writer does or not.

5.5 *Multiple F Tests.*

A series of tests paralleling the above multiple range tests can be defined using F tests instead of range tests. These may conveniently be termed multiple F tests. Thus, corresponding to the new multiple range test, an *α-level multiple F test with protection levels based on degrees of freedom* may be defined by the following rule: Rule 1. *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 $\alpha_p = 1 - \gamma_p$, $\gamma_p = (1 - \alpha)^{p-1}$, and p is the number of means in the subset concerned.*

In the case of three means under the special conditions $n_2 = \infty$, $\sigma_m = 1$, $\alpha = .05$, the regions of this test are as shown in Figure 5. These regions are the same as those of the corresponding multiple normal-deviate test except that the strip-regions $(\underline{1, 2})$, $(\underline{1, 3})$, $(\underline{2, 3})$ have their boundaries expanded to those of the circle centered at the origin, with radius 3.05. This radius 3.05 is calculated as $\sqrt{4F}$, where* F is the 9.75% significant value of an F ratio with degrees of freedom 2 and ∞. If the center region $(\underline{1, 2, 3})$ were comprised of the circle alone, this would raise the three-mean protection level to just 90.25%

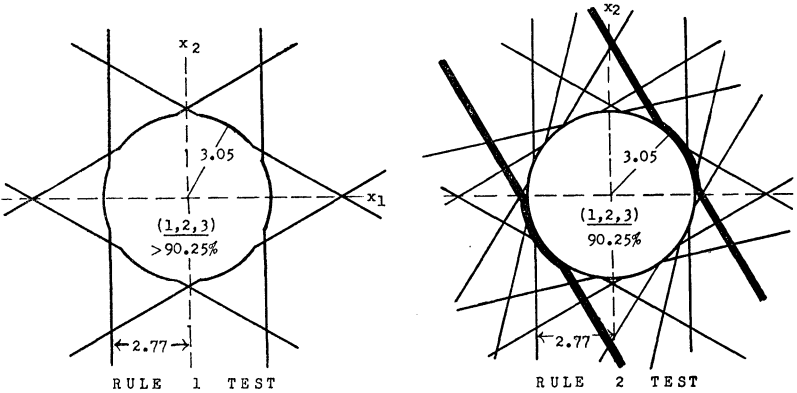


FIGURE 5

5% level multiple F tests with special protection levels ($n_2 = \infty$, $\sigma_m = 1$)

*This test requires special F tables or equivalent tables as given in (6), Tables 1 and 2.

as desired. The six small areas outside the circle but inside $(\underline{1, 2, 3})$ give the test a slightly higher protection level than 90.25%, which is not necessary and makes some modification of Rule 1 desirable.

The multiple F test can be generalized to test the significance of all linear comparisons of the form $c = \sum_{i=1}^{n} k_i m_i$, where $k_1, k_2, \cdots, k_n$ is any set of arbitrary constants such that $\sum_{i=1}^{n} k_i = 0$. (Each linear function of this form can be regarded as the difference between weighted means of two subsets of the full set of means.) The general rule is: Rule 2. *Any comparison of the form $c = \sum_{i=1}^{n} k_i m_i$ 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 $\alpha_p = 1 - \gamma_p$, $\gamma_p = (1 - \alpha)^{p-1}$, and p is the number of means in the subset concerned.* By "all of the means involved in c" is meant all means which have non-zero coefficients in the linear function $c = \sum_{i=1}^{n} k_i m_i$.

The regions of this more general test, under the same special conditions, are also shown in Figure 5. The three intersecting strip regions given by Rule 1 are now replaced by an infinity of strips, all of which pass symmetrically through the center of the sample space and intersect each other at all angles. Each strip and the areas to either side of it represent the test regions for the comparison measured at right angles to the axis of the strip. For example, the strip region between the heavy lines in the illustration contains points for samples in which the comparison $c = \frac{1}{2}m_2 + \frac{1}{2}m_3 - m_1$ is not significantly different from zero. The areas to either side of this region contain points for samples in which the comparison is significantly positive or negative.

5.5.1 *The Multiple Comparisons Test.*

The *multiple comparisons test* proposed by the author in 1951 (6, 7) is a multiple F test which consists of a compromise between Rule 1 and Rule 2. As many significant differences as possible are found by the Rule 1 test. Rule 2 is then used to test any comparisons of interest within subsets of means not already found to contain significant differences by Rule 1.

Figure 6 shows the regions of this test under the same special conditions as before. These regions are identical with those of the Rule 1 test in Figure 5 except for the additional six regions lying outside the circle and inside the original hexagon. These represent regions in which comparisons involving all three means are found to be significant. In the small region at the top of the circle, for example, various weighted means of m_1 and m_2 are significantly larger than m_3.

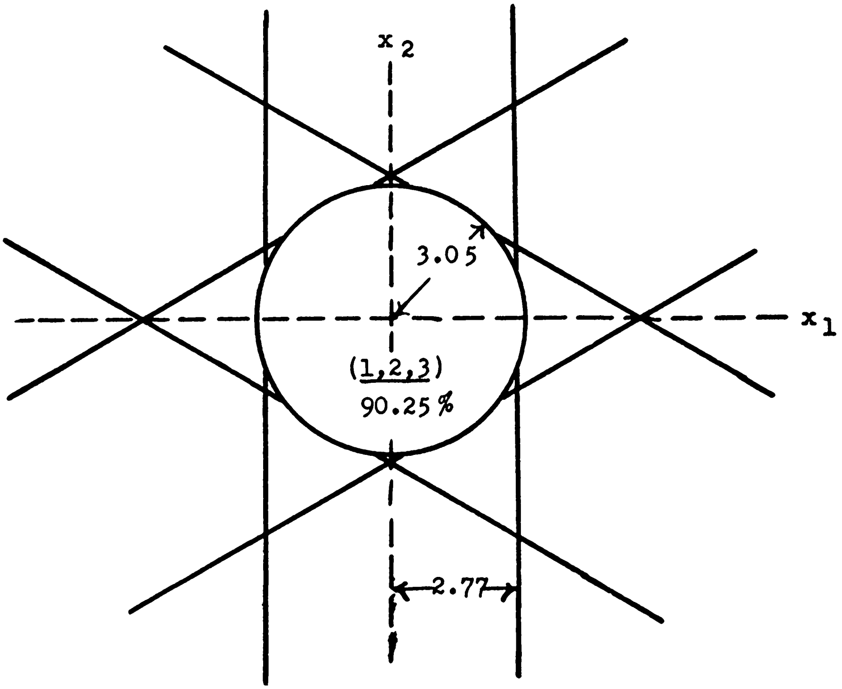


FIGURE 6

5% Level Multiple Comparisons Test ($n_2 = \infty, \sigma_m = 1$)

5.5.2 *The Least-Significant-Difference Test.*

The basic principle of using a preliminary homogeneity of means test to raise a low protection level was first proposed by R. A. Fisher (9). A test which has arisen out of his discussion is the *least-significant-difference test* already mentioned in the introduction.

A general α-level test of this type is given by the rule: *The difference between any two means in a set of n is significant provided that the difference is significant according to an α-level t test and provided also that the variance of the whole set is significant according to an α-level F test.*

In the case of three means, this is identical with an α-level Rule 1 multiple F test with constant levels. The regions of the test under the same special conditions as before are the same as those of the Rule 1 multiple F test with special levels in Figure 5 or the multiple comparisons test in Figure 6 except that the radius of the circle is increased to $\sqrt{4F} = 3.46$, F now being the 5% level value of the F ratio with degrees of freedom 2 and ∞, In the more general case with n means, $n > 3$, the least significant difference test does not use all of the F tests prescribed by a multiple F test and fails to fix adequate values for all of the protection levels involved. For example in a test of four means, assuming $n_2 = \infty$, $\sigma_m = 1, \alpha = .05$ as before, we find $\gamma_2 = 95\%, \gamma_3 = 87.8\%$, and $\gamma_4 = 95\%$. The value γ_3 of the three-mean protection levels is as low as that of the corresponding multiple normal deviate test. In general, the value γ_p of any p-mean protection level in an α-level least significant difference test is as low as the γ_p value in the corresponding α-level multiple t test with the one exception that γ_n is raised to $1 - \alpha$.

Thus while this test is more conservative than the new multiple range test or the multiple comparisons test for the case of three means, it is less conservative in cases with more than three means.

5.5.3 *Scheffé's Test Based on Contrasts.*

A recent procedure proposed by Scheffé (19) may be described as the F test analogue of Tukey's test based on allowances.

In the case of three means under the same conditions as before, the regions of this test are generated by the symmetrical intersection of strip regions with straight boundaries like those of the multiple normal-deviate test except that (i) the width of the strips is 2×3.46 instead of 2×2.77, and (ii) the strips are infinite in number as in the Rule 2 multiple F test. The intersections of these strips form a circle of radius 3.46 at the center and this gives the test a three-mean protection level of 95%. At the same time the strip-region protection levels are raised, by the increases in strip-widths, from 95% to 98.6%.

5.6 *Other Decision Procedures.*

As mentioned previously several writers including Bechhofer (1) have dealt with a problem which may be regarded as a special case of the general one with which we have been concerned, and procedures have been proposed which may be regarded as degenerate multiple range or multiple F tests. The decision procedures proposed in the given reference, for example, are for deciding that the t largest means in a sample of n means $m_1, m_2, \cdots, m_n$ are all significantly larger than all of the remaining $n - t$ means. In one procedure the true means corresponding to the t largest observed means are not ranked relative to one another; in another procedure they are. In both cases the true means in the remaining subgroup are left unranked relative to one another. To take a simple illustration, in a procedure for choosing the largest mean among four, that is, $t = 1$ and $n = 4$, the decisions in terms of our previous notation are $(\underline{1, 2, 3}, 4)$, $(\underline{1, 2, 4}, 3)$, $(\underline{1, 3, 4}, 2)$ and $(\underline{2, 3, 4}, 1)$, where $(\underline{1, 2, 3}, 4)$, for example, is the decision that μ_4 is larger than each of the remaining means, which are left unranked relative to another.

One very restrictive result of eliminating the missing decisions is that all of the protection levels of the procedure are forced to zero, or in other words all of the significance levels are forced to 100%. For example, in a procedure involving only two means, the experimenter is forced to make the decision (1, 2) or (2, 1). Thus, if it so happens that $\mu_1 = \mu_2$ the probability of making a wrong decision is 100%. The power curves of this test are similar to the $p(1, 2)$ and $p(2, 1)$ curves illustrated for the 5% level test in Figure 1b except that each curve is forced to pass through the 50% power value at $\epsilon = \mu_1 - \mu_2 = 0$. The usefulness of these procedures is therefore restricted to problems in which the experimenter feels impelled to choose a best mean from the results of the given experiment alone.

By limiting themselves to procedures with zero protection levels at the outset, the authors of these tests have been able to avoid the controversial problem of consistent protection levels and to concentrate on other problems such as the tabulation of relations between power functions and sample sizes, (Bechhofer, 1), and the optimum choice of the size of an experiment based on minimax considerations, (Somerville, 20).

6. CONCLUDING REMARKS

Most of the foregoing procedures can be classified usefully according to three basic characteristics:

1. *Type of significant differences*: separating a procedure such as the Newman-Keuls test having a set of significant differences which decrease as the test proceeds, from a procedure such as Tukey's test based on allowances which has one constant significant difference.
2. *Type of protection levels*: separating a procedure such as the Newman-Keuls test having constant values (or lower limits) of $(1 - \alpha)$ for its protection levels*, from a test such as the new multiple range test having protection levels based on degrees of freedom.
3. *Type of component tests*: separating procedures into several categories according to whether they employ range tests, F tests, or component tests of another type.

*excluding class 2 protection levels.

Table V shows the allocation of several procedures in a classification of this kind.

The most important of these characteristics is the first, separating tests 1a, with decreasing significant differences, from tests 1b, with constant significant differences. The nature of the confidence interval methods from which the 1b tests are derived is such that in an application of one of these tests there is only one single significant value against which all differences or linear comparisons are tested. This makes for considerable simplicity. However, the single significant value has to be so high that the power functions are severely reduced.

TABLE V. CLASSIFICATION OF TEST PROCEDURES ACCORDING TO THREE BASIC CHARACTERISTICS

2. Type of Protection Levels	1. Type of Significant Differences			
	1a) Decreasing 3. Component Tests		*1b) Constant* 3. Component Tests	
	3a) Range	3b) F	3a) Range	3b) F
2a) None less than constant values $\gamma_p = (1 - \alpha)$	Newman-Keuls Test		Tukey's Test Based on Allowances	Scheffé's Test
2b) Protection Levels Based on degrees of freedom $\gamma_p = (1 - \alpha)^{p-1}$	New Multiple Range Test	Multiple Comparisons Test		

For example, in a 5% level Tukey test based on allowances for a case with 20 means (again assuming $n_2 = \infty$, $\sigma_m = 1$), the significant ranges all have the same value 5.01, as shown in Table VI. This value 5.01 is equal to the largest of the significant ranges of the corresponding 1a test, a 5% level Newman-Keuls test, for which the significant ranges, also shown in Table VI, decrease with subset size from 5.01 down to 2.77. In the 1a test, a difference between two means which exceeds only 2.77 can be significant depending on the disposition of the other means. In the 1b test no difference can be significant without exceeding 5.01.

Comparing these two tests further, consider two true means in particular, say μ_1 and μ_2, and suppose that μ_1 is smaller than μ_2. Let

μ_1 and μ_2 on one hand be well separated from the remaining true means $\mu_3, \mu_4, \cdots, \mu_{20}$ on the other. For example, suppose $\frac{1}{2}(\mu_1 + \mu_2) = 120$ and $\mu_3 = \mu_4 = \cdots = \mu_{20} = 100$. Under these circumstances, recalling that $\sigma_m = 1$, the observed means m_1 and m_2 will be well separated from the remaining observed means $m_3, m_4, \cdots, m_{20}$. Because of this, the ranges of all subsets of three or more of the observed means which include m_1 and m_2 are practically certain to be significant. Thus in the 1a test the probability of correctly deciding that μ_1 is less than μ^{2} will be virtually the same as if the remaining means were not present, that is,

TABLE VI. COMPARISON OF SIGNIFICANT RANGES FOR 5% LEVEL TESTS OF 20 MEANS

Test	Subset Sizes								
	2	3	4	5	6	8	10	14	20
Tukey's Test Based on Allowances	5.01	5.01	5.01	5.01	5.01	5.01	5.01	5.01	5.01
Tukey's 1953 Test	3.89	4.16	4.32	4.44	4.52	4.65	4.74	4.88	5.01
Newman-Keuls Test	2.77	3.32	3.63	3.86	4.03	4.29	4.47	4.74	5.01
New Multiple Range Test	2.77	2.92	3.02	3.09	3.15	3.23	3.29	3.38	3.47

$$p_{1a}(1, 2) = P[\text{dec. } (1, 2) \mid \mu_2 - \mu_1] = P[m_1 - m_2 < -2.77 \mid \mu_2 - \mu_1].$$

For the 1b test, however, the corresponding power is given by

$$p_{1b}(1, 2) = P[\text{dec. } (1, 2) \mid \mu_2 - \mu_1] = P[m_1 - m_2 \leq -5.01 \mid \mu_2 - \mu_1].$$

Table VII shows the values of these two functions and their differences for various values of $\mu_2 - \mu_1$. The differences represent the losses in power in the 1b test relative to the 1a test and some of these can be seen to be very large.

At other parameter values in a 20-mean test, with other arrangements of the true means, the relative losses in power will not be as great. However, it is clear that losses will occur at all values of the parameters and many will be considerable. For tests involving more than 20 means the differences in power will be even greater, increasing as the number of means increases.

TABLE VII. SEVEREST POWER LOSSES OF 1b TEST RELATIVE TO 1a TEST (5% LEVEL TESTS OF 20 MEANS)

$\mu_2 - \mu_1$	1a Test	1b Test	Loss
0	.0250	.0002	.0248
1	.1056	.0023	.1033
2	.2946	.0166	.2780
3	.5636	.0778	.4858
4	.8078	.2389	.5689
5	.9429	.4960	.4469
6	.9887	.7580	.2307
7	.9986	.9207	.0779
8	.9999	.9826	.0173
∞	1.0000	1.0000	0.0000

Similar decreases in power must occur in all 1b tests using constant significant differences. These losses appear unnecessary and tests of this type are therefore not recommended.

A partial concession to this point of view is made by Tukey (23) in his 1953 test already mentioned. The significant ranges for this test lie midway between those of the corresponding 1a and 1b tests. An example of these under the conditions already used for the previous 20-mean test examples is also given in Table VI. A test of this type, however, still suffers considerable losses in power probabilities relative to the Newman-Keuls procedure and is also considered to be unnecessarily conservative.

The second most important characteristic is the one concerning protection levels. This separates tests 2a, using constant values (or lower limits) for protection levels, from tests 2b, using the special lower limits based on degrees of freedom.

As has already been mentioned, the power functions of the 2a tests are uniformly lower than those of the corresponding 2b tests. Some further idea of this may be obtained from Table VI by comparing the Newman-Keuls significant ranges, discussed above, with those of the corresponding new multiple range test, which have been taken from Table II, row $n_2 = \infty$.

Each of these tests requires that a difference between any two means must exceed 2.77 before it can be significant and each thus has two-mean protection levels of 95%. The significant ranges for subsets of more than two means, however, are larger in the 2a test. As a result of this, some differences which may not be significant in the 2a test may be significant in the 2b test. It can be seen that the amounts by which the power functions of the 2b test exceed those of the 2a test are greatest around the origin $\mu_1 = \mu_2 = \cdots = \mu_{20}$ and decrease toward zero in certain directions away from this point. The same holds for any 2b test, relative to the corresponding 2a test.

There appears to be no sound reason for not using protection levels based on degrees of freedom thereby gaining considerably in power to detect real differences.

Finally, there is the subdivision of the test procedures according to the type of component tests employed. In this paper we have considered only procedures based on range tests (3a) and F tests (3b). However, other types of component tests, for example, extreme deviate tests and gap tests, have been proposed and one procedure given by Tukey (21) is based on a combination of three types of component tests.

The problem of deciding the relative merits of various types of component tests is complex, and much work needs to be done in this direction. At present, it appears that the best choice lies between range tests and F tests. The relative merits of these depend on the objectives involved.

Under some circumstances (i), interest may lie in testing linear comparisons involving several means as well as differences between single means; under others (ii), interest may be restricted to testing only differences between single means.

Under circumstances (i) additional power functions are needed to measure the power of the test with respect to the additional comparisons involved. When these are all included it seems safe to assume that multiple F tests are more powerful in some average sense than multiple range tests. Under circumstances (ii), however, the relations are more obscure. The preliminary tests in a multiple F test with decreasing significant differences (1a tests) may cause a little less general interference* with subsequent tests than do the preliminary range tests in a corresponding multiple range test. In this event, the multiple F tests may still be more powerful in an average sense but only slightly so.

The important deciding factor under circumstances (ii) will often be the difference in time and effort required in applying the two types of tests. The application of a multiple range test is much easier and a test of this type will generally be preferred for this reason.

To summarize, the features recommended in each classification are:

1. Decreasing significant differences, as used in tests 1a;

*This does not apply of course in 1b tests with *constant* significant differences, in which case the use of range tests gives more powerful procedures. Thus, for example, under circumstances (ii), Tukey's test based on allowances is uniformly more powerful than Scheffé's test

2. Protection levels based on degrees of freedom, as used in tests 2b; and
3. Range tests as used in tests 2a, unless one is interested in linear comparison other than differences between single means, in which case F tests are recommended, as used in tests 3b.

The new multiple range test and the multiple comparisons test have been designed to include these recommended features.

Computation of Tables II and III for New Multiple Range Test.

Let $Q(p, n_2, \alpha)$ represent the entry for given values of p, n_2, and α given in Tables II and III for $\alpha = .05$ and $.01$, respectively. Put $R(p, n_2, \gamma_{p,\alpha})$ for the $100\gamma_{p,\alpha}$ percentage point of the studentized range where $\gamma_{p,\alpha} = (1 - \alpha)^{p-1}$. Then the tabled values have been computed from the relation $Q(p, n_2, \alpha) = R(p, n_2, \gamma_{p,\alpha})$ for $p = 2$, and from $Q(p, n_2, \alpha) = R(p, n_2, \gamma_{p,\alpha})$ or $Q(p - 1, n_2, \alpha)$, whichever is the larger, for all other values of p. This ensures that each p-mean protection level in the new multiple range test is $\gamma_{p,\alpha}$ for all values of p.

The studentized range values $R(p, n_2, \gamma_{p,\alpha})$ for $2 \leq p \leq 20$ and $10 \leq n_2 \leq \infty$ used in this process have been obtained from Pearson and Hartley's Tables (16). The remainder of the $R(p, n_2, \gamma_{p,\alpha})$ values involved have been obtained by new methods (see Beyer, 2) specially developed for this purpose.

Acknowledgment. The author is indebted to W. H. Beyer for much of the theoretical developments and the computational work involved in getting the values $R(p, n_2, \gamma_{p,\alpha})$ of the studentized range required for Tables II and III as explained above.

7. REFERENCES

(1) Bechhofer, Robert E., "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances," *Annals of Mathematical Statistics, 25*, 16–39, 1954.

(2) Beyer, William H., "Certain Percentage Points of the Distribution of the Studentized Range of Large Samples," Virginia Polytechnic Institute M.S. thesis, 56 pp., 1953.

(3) Cornfield, J., Halperin, M. and Greenhouse, S., "Simultaneous Tests of Significance and Simultaneous Confidence Intervals for Comparisons of Many Means," unpublished mimeographed notes, National Institutes of Health, Public Health Service, Department of Health, Education and Welfare, Bethesda, Maryland, 18 pp., 1953.

(4) Davies, Owen L., "Statistical Methods in Research and Production," 2nd ed., London, Oliver and Boyd, 1949.

(5) Duncan, D. B., "Significance Tests for Differences between Ranked Variates Drawn from Normal Populations," Iowa State College Ph.D. thesis, 117 pp., 1947.

(6) Duncan, D. B., "A Significance Test for Differences between Ranked Treatments in an Analysis of Variance," *Virginia Journal of Science*, *2*, 171–189, 1951.

(7) Duncan, D. B., "On the Properties of the Multiple Comparisons Test," *Virginia Journal of Science*, *3*, 49–67, 1952.

(8) Duncan, D. B., "Multiple Range Tests and the Multiple Comparisons Test," (Preliminary Report), *Biometrics*, *9*, Abstract 220, 1953.

(9) Fisher, R. A., "The Design of Experiments," six eds., London, Oliver and Boyd, 1935–1951.

(10) Keuls, M., "The Use of the 'Studentized Range' in Connection with an Analysis of Variance," *Euphytica*, *1*, 112–122, 1952.

(11) Lehmann, E. L., "Some Principles of the Theory of Testing Hypotheses," *Annals of Mathematical Statistics*, *21*, 1–26, 1950.

(12) Newman, D., "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, 1939.

(13) Neyman, J., "First Course in Probability and Statistics," Henry Holt and Company, Inc., New York, 1950.

(14) Paterson, D. D., "Statistical Technique in Agricultural Research," McGraw-Hill Book Company, Inc., New York, 1939.

(15) Pearson, E. S., and Hartley, H. O., "The Probability Integral of the Range in Samples of n Observations from a Normal Population," *Biometrika*, *32*, 301–310, 1942.

(16) Pearson, E. S., and Hartley, H. O., "Tables of the Probability Integral of the 'Studentized' Range," *Biometrika*, *33*, 89–99, 1943.

(17) Roy, S. N. and Bose, R. C., "Simultaneous Confidence Interval Estimation," *Annals of Mathematical Statistics*, *24*, 513–536, 1953.

(18) Sawkins, D. T., unpublished work, University of Sydney, 1938.

(19) Scheffé, H., "A Method for Judging All Contrasts in the Analysis of Variance," *Biometrika*, *40*, 87–104, 1953.

(20) Somerville, P. N., "Some Problems of Optimum Sampling," *Biometrika*, *41*, 420–429, 1954.

(21) Tukey, J. W., "Comparing Individual Means in the Analysis of Variance," *Biometrics*, *5*, 99–114, 1949.

(22) Tukey, J. W., "Quick and Dirty Methods in Statistics," part II, Simple Analyses for Standard Designs, *Proceedings Fifth Annual Convention, American Society for Quality Control*, 189–197, 1951.

(23) Tukey, J. W., "The Problem of Multiple Comparisons," unpublished dittoed notes, Princeton University, 396 pp., 1953.

(24) Wald, A., "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *Annals of Mathematical Statistics*, *10*, 299–326, 1939.

(25) Hartley, H. O., "Some Significance Test Procedures for Multiple Comparisons," *Annals of Mathematical Statistics*, *25*, Abstract 19, 1954.