DOCUMENT RESUME

ED 109 185 TM 004 635

AUTHOR Carlson, James E.; And Others
TITLE The Distribution of the Test Statistic Used in the Newman-Keuls Multiple Comparison Technique.
PUB DATE [Apr 75]
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS *Analysis of Variance; *Hypothesis Testing; Sampling; Statistical Analysis; *Tests of Significance
IDENTIFIERS *Newman Keuls Analyses

ABSTRACT

Researchers often use the analysis of variance to test hypotheses about the means, followed by a multiple comparison technique when the F-test is significant. The technique is this study was developed by Newman (1939) and Keuls (1952). A flaw in the rationale underlying this technique was evaluated to determine whether the flaw is sufficiently serious to advise against the use of the Newman-Keuls method. A monte carlo study of the sampling distribution of the test statistic used in the Newman-Keuls method was carried out and implications are discussed. (Author)

THE DISTRIBUTION OF THE TEST STATISTIC USED IN THE NEWMAN-KEULS MULTIPLE COMPARISON TECHNIQUE

James E. Carlson
Charles E. Stegman
Wayne R. Applebaum

University of Pittsburgh

Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April, 1975.

THE DISTRIBUTION OF THE TEST STATISTIC USED IN THE NEWMAN-KEULS MULTIPLE COMPARISON TECHNIQUE

James E. Carlson, Charles E. Stegman, and Wayne R. Applebaum
University of Pittsburgh

1. Introduction

The purpose of this study was to investigate the distribution of the test statistic used in the Tukey, Newman-Keuls and Duncan methods of performing post-hoc multiple comparisons following a significant analysis of variance F-test. There is a problem in this distribution that is not directly discussed in the literature we have surveyed but is mentioned in a personal communication from Rupert G. Miller (1971), and is alluded to by Spjøtvoll(1974). Because different Studentized range distributions are used directly in determining critical values for the Newman-Keuls method our discussion will center around this particular method.

2. The Newman-Keuls Technique

Following a significant analysis of variance F-test, among the many available post hoc multiple comparison techniques is one first mentioned by "Student" (Newman, 1939), developed by Newman (1939) and rediscovered by Keuls (1952). This procedure is now known as the Newman-Keuls method.

Given the set of k population means, $\mu_1, \mu_2, \ldots, \mu_k$, estimated by the means of k independent random sample, $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$, the first step in the Newman-Keuls procedure is to rank these sample means. Denoting the j^{th} smallest mean as $\overline{Y}_{(j)}$ so that the smallest is $\overline{Y}_{(1)}$ and the largest is $\overline{Y}_{(k)}$, the ordered set of sample means is $\overline{Y}_{(1)}, \overline{Y}_{(2)}, \ldots, \overline{Y}_{(k)}$.

The second step is to compute the following ratio: $_kQ_{(k)-(1)} = \frac{\overline{Y}_{(k)} - \overline{Y}_{(1)}}{\sqrt{MS_{e/n}}}$ where n is the size of each of the k samples and MS_e is the error mean square from the analysis of variance (note that the techniique assumes that the k samples are all of the same size, n). This ratio is then compared to $q_{k,v,1-\alpha}$, the 100(1-α) percentile of the Studentized range distribution with parameters k and v, the number of degrees of freedom associated with the error sum of squares in the analysis of variance.

If

$$_kQ_{(k)-(1)} > q_{k,v,1-\alpha} \qquad [1]$$

the hypothesis of no difference between the means of groups (k) and (1) is rejected and the researcher proceeds to test other pairs of means. If, on the other hand,

$$_kQ_{(k)-(1)} \leq q_{k,v,1-\alpha} \qquad [2]$$

the researcher concludes that the two most extreme group means are not different. Since the two most extreme group means are judged not to be different, the researcher also concludes that all other, less extreme, pairs are not different. In this case, no further comparisons are made.

If the inequality in [1] is true, the next comparisons to be made are those repreaented in the inequalities.

$$_kQ_{(k-1)-(1)} = \frac{\overline{Y}_{(k-1)} - \overline{Y}_{(1)}}{\sqrt{MS_e / n}} > q_{k-1,v,1-\alpha} \qquad [3]$$

and

$$_kQ_{(k)-(2)} = \frac{\overline{Y}_{(k)} - \overline{Y}_{(2)}}{\sqrt{MS_e / n}} > q_{k-1,v,1-\alpha} \qquad [4]$$

If [3] is true the researcher proceeds to examine $\overline{Y}_{(k-2)} - \overline{Y}_{(1)}$, but if [3] is not true then the two means $\mu_{(k-2)}$ and $\mu_{(1)}$ are judged not to be different. Similarly, if [4] is true $\overline{Y}_{(k-1)} - \overline{Y}_{(2)}$ and $\overline{Y}_{(k)} - \overline{Y}_{(3)}$ are examined, but if [4] is not true the analogous population means are judged not to be different. If we construct a table of differences between pairs of ordered means, as illustrated in Table 1, it can be seen that this procedure results in the following rules: (a) we start by examining the difference in the upper right corner and if it is significant we proceed to examine the next difference in the same row and the next difference in the same column, (b) when a nonsignificant difference is found, no further differences in the same row and the same column are examined.

Table 1

Mean Difference Table for Newman-Keuls Technique

$\overline{Y}_{(2)} - \overline{Y}_{(1)}$	$\overline{Y}_{(3)} - \overline{Y}_{(1)}$	....	$\overline{Y}_{(k-1)} - \overline{Y}_{(1)}$	$\overline{Y}_{(k)} - \overline{Y}_{(1)}$
	$\overline{Y}_{(3)} - \overline{Y}_{(2)}$	...	$\overline{Y}_{(k-1)} - \overline{Y}_{(2)}$	$\overline{Y}_{(k)} - \overline{Y}_{(2)}$
			$\vdots$	$\vdots$
			$\overline{Y}_{(k-1)} - \overline{Y}_{(k-2)}$	$\overline{Y}_{(k)} - \overline{Y}_{(k-2)}$
				$Y_{(k)} - Y_{(k-1)}$

3. The Problem

The use of the $q_{k,v}$ distribution for the distribution of the test statistic, ${}_kQ_{(k)-(1)}$, as a test of

$$H_{o1}: \mu_1 = \mu_2 = \ldots = \mu_k$$

is valid when the usual ANOVA assumptions are met (Spjøtvoll, 1974). That is, given normality, independence, homogeneity of variance, and the fact that the null hypothesis is true then the variable ${}_kQ_{(k)-(1)}$ (in formula 1) is distributed as the Studentized range with k and v degrees of freedom. As Spjøtvoll notes, if we reject H_{o1} using the decision rule in [1] then we "conclude" that $\mu_k \neq \mu_1$. In this case, it is tenable that it is population k that is different from the others or that it is population 1. Therefore, we have two possible simplifying hypotheses:

$$H_{o2}: \mu_1 = \mu_2 = \ldots = \mu_{k-1}$$

and

$$H_{o3}: \mu_2 = \mu_3 = \ldots = \mu_k$$

These correspond respectively to the cases where μ_k and μ_1 lead to the rejection of H_{o1}.

Now if it is <u>true</u> that H_{o1} is false then in H_{o2} and H_{o3} we have chosen to remove a population mean (either μ_k or μ_1) that is indeed different from the others. In this case, the removal of this extreme value should not effect the order statistics of the remaining k-1 means and we can use the test statistics ${}_kQ_{(k-1)-(1)}$ and ${}_kQ_{(k)-(2)}$ to test H_{02} and H_{03} respectively. Again, if we assume H_{o2} and H_{o3} are true and the ANOVA assumptions are met then the variables ${}_kQ_{(k-1)-(1)}$ and ${}_kQ_{(k)-(2)}$ should be distributed as the Studentized range with k-1 and v degrees of freedom.

However, if we made a Type I error in "concluding" that H_{o1} is false then the distributions of the variables ${}_kQ_{(k-1)-(1)}$ and ${}_kQ_{(k)-(2)}$ are not the results of k-1 independent samples. Rather they represent k-1 samples remaining after discarding one extreme observation. This has the effect of truncating or restricting the range of values for the test statistics. It was hypothesized that this would reduce the actual mean and variance of the distributions and consequently will lead to a more conservative test of H_{o2} and H_{o3}. The same argument applies to H_{o4}, H_{o5}, etc. if we rejected H_{o2} or H_{o3} on the basis of ${}_kQ_{(k-1)-(1)}$ or ${}_kQ_{(k)-(2)}$.

Of course, if we knew that H_{o1} were actually true then following the logic of the testing procedure we would not test H_{o2}, H_{o3}, etc. However, in practice we make the original decision on the basis of ${}_kQ_{(k)-(1)}$ and indeed Type I errors will be made. Since this is a problem for all such comparisons, a natural question is, "How different are the true distributions of the Newman-Keuls test statistics from the Studentized range distributions that are actually used?"

Since the mathematical derivations and subsequent numerical integrations to determine the critical points for the actual distributions are quite complex, a Monte Carlo sampling study was designed to estimate how conservative the test procedures are. The design of the study is discussed in Section 5.

4. Related Literature

Although we were unable to find any literature relating directly to the problem of this study there are several sources that provide information about the Newman-Keuls technique and its use.

As mentioned earlier the technique examined in this study was first proposed by Newman (1939) and later Keuls (1952), apparently unaware of Newman's paper, proposed the same technique. Newman, and Miller (1966) both indicate that the basic idea was an outgrowth of the work of "Student."

The most comprehensive textbook discussions of the Newman-Keuls technique are those of Miller (1966, pp. 81-90) and Winer (1971, pp. 191-196). Miller presents most of the underlying theory and compares the technique to alternative techniques, in particular to the Duncan multiple range test. Winer presents the computational procedures and also presents a comparison of seven different multiple comparison methods, including the Newman-Keuls method. Other texts that discuss the technique include those of Kirk (1968) and Mendenhall (1968).

Besides the above-mentioned textbook discussions there are several journal articles that include discussions of the Newman-Keuls technique. The most comprehensive of these is the article by Spjøtvoll (1974) in which the bases of several procedures and comparisons of them are discussed. O'Neill and Wetherill (1971) briefly mention the Newman-Keuls technique in an interesting "state of the art" discussion, and also provide an extensive bibliography on multiple comparison procedures. Two studies (Petrinovich & Hardyck, 1969; Carmer & Swanson, 1973) have been conducted in which Monte Carlo methods were used to compare the Type I and Type II error rates of several multiple comparison technqiues. Although the Newman-Keuls technique was investigated in both of these studies the conclusions were quite different. Petrinovich and Hardyck argue against using the Newman-Keuls technqiue because they found it had a high experimentwise Type I error rate under conditions other than the case where all populations have identical means. Carmer and Swanson, on the other hand reject the Newman-Keuls procedure because of its Type II error rate. Games (1971) provides some additional discussion in a criticism of the Petrinovich and Hardyck article. Interestingly, Carmer and Swanson generated their data according to a randomized block design with zero block effects, and analyzed the data according to that model. With zero block effects the correct model would be the completely randomized model and, in general, analysis according to an incorrect model leads to a loss in precision. The authors do not indicate why they chose the blocking design.

5. Design of the Study

In order to investigate the sampling distribution of the test statistic used in the Newman-Keuls technique a total of 9 experiments were simulated, each with 10,000 replications. The first three experiments each had three groups and used samples of size 5, 10, and 15 respectively. The next three experiments involved four groups again using sample sizes of 5, 10 and 15. The final three experiments were explicitly designed to generate empirical sampling distributions of the specific Studentized range distributions that would be needed in some of the first six experiments. The descriptions of the nine experiments shown in Table 2, include the error degrees of freedom in order to indicate in which experiments the distributions are comparable.

Table 2

Descriptions of the Simulated Experiments

Experiment No.	1	2	3	4	5	6	7	8	9
No. of Populations	3	3	3	4	4	4	2	2	3
Size of each Sample	5	10	15	5	10	15	7	22	13
Error Degrees of Freedom	12	27	42	16	36	56	12	42	36

For each of the possible contrasts in each experiment the frequency with which the value of the test statistic fell in each of seven intervals was tabulated. The seven intervals were defined by using percentiles of the particular Studentized range distribution that would be used for the Newman-Keuls technique on each contrast. The percentiles (.900, .950, .975, .990, .995, .999) were taken from Harter's (1960) tables or calculated by linear harmonic interpolation of values from those tables as suggested by Harter. The upper tail percentiles were chosen because they are the critical areas of the distribution used by the Newman-Keuls technique. Thus, using Q to represent the value of the test statistic and P_j to represent the j^{th} percentile of the appropriate Studentized range distribution, the seven intervals that were used were: $Q \leq P_{90}$; $P_{90} < Q \leq P_{95}$; $P_{95} < Q \leq P_{97.5}$; $P_{97.5} < Q \leq P_{99}$; $P_{99} < Q \leq P_{99.5}$; $P_{99.5} < Q \leq P_{99.9}$; and $P_{99.9} < Q$.

Estimates of the mean, variance, skewness and kurtosis were also calculated for each distribution, using the formulas given by Bennett and Franklin (1954, pp. 81-82). Using the method described by Newman (1939) expected values of each theoretical distribution were computed and comparisons were made between these expectations and the means of the distributions that were generated.

It was anticipated that the distributions for the test statistics would show less variability than the theoretical distributions for contrasts involving pairs of means other than the two most extreme for a particular experiment.

6. Generation Techniques

Values of the dependent variable were pseudo-random standard normal deviates generated by using the log-and-trig transformation method (Box & Muller, 1958) on pseudo-random uniform values generated by the multiplicative congruential method. The multiplier for the congruential method was 131075, which was selected because it has been demonstrated to result in variates having desirable characteristics (Pingel, 1975; MacLaren & Marsaglia, 1965). All computations were done on a PDP10 computer at the University of Pittsburgh Computer Center. This machine stores an integer in a location having 35 bits plus sign, and thus residues mod $(2^{35}-1)$ were used by allowing integer multiplications to overflow. For more details of this technique the interested reader is referred to Newman and Odell (1971) or Hammersley and Handscomb (1964).

In order to check the computer programs, samples of the pseudo-random uniform and normal variates were generated and unbaised estimates of the means and variances computed. A sample of 10,000 uniform values had a mean of .49945 and variance of .08238 as compared to the theoretical values of .5 and .08333, respectively. Three samples, each of 10,000 normal values were generated and had means of -.01254, -.00344, and .01052 and variances of 1.00182, .99960 and .99350 as compared to the theoretical values of zero and one. Finally, using the final version of the computer program, a total sample of 3000 normal values was generated and the mean and variance estimates were .00858 and .99565 respectively.

7. Results

Tables 3 and 4 show the frequencies with which values of the Newman-Keuls test statistic were distributed across the seven intervals defined by using percentiles of the theoretical Studentized range distribution.

For rows 1, 4, 5, 8, and 11 the Studentized range distribution is theoretically correct. As can be seen by examining these rows the observed frequencies correspond closely to the expected frequencies. Row 1 corresponds to the distribution of ${}_3Q_{(3)-(1)}$ and the Studentized range distribution is the appropriate distribution. As outlined before, if a Type I error is made in this test then we would proceed to incorrectly test ${}_3Q_{(2)-(1)}$ and ${}_3Q_{(3)-(2)}$. The distribution of these statistics do not theoretically have Studentized range distributions. The observed frequencies in rows 2 and 3 clearly indicate that these two tests deviate markedly from the expected frequencies. The deviations are in the predicted direction, that is, the tests would be conservative. For comparison purposes we generated the data in row 4 which has the same degrees of freedom as rows 2 and 3. However, this distribution is theoretically correct because it involves the contrast between the two extreme means in a two population case. Recall from Table 2 that in order to get comparable degrees of freedom the sample size in experiment 7 were equal to seven. Further examination of Table 3 for the cases where n=10 and n=15 supplies further evidence for the conclusion that if a Type I error is made for the first contrast tested by the Newman-Keuls technique then the significance level for the next two tests is much smaller than the nominal level the researcher may think he is using when he performs the tests. That is, when k=3 the sample sizes, and hence the number of error degrees of freedom, appear to make little difference in the conservativeness of the tests.

Similar conclusions are reached after examining rows (1,2,3; 7,8,9,10; 14,15,16) of Table 4. These cases represent the situations where k=4 and n=5, 10, and 15. It is easy to see again that when a Type I error is made in the first test then the resultant tests based on $_4Q_{(4)-(2)}$ and $_4Q_{(3)-(1)}$ are conservative. Next, consider rows (4,5,6; 11,12,13; 17,18,19) of Table 4. Examination of these rows shows, as might be expected, that the distribution of the test statistic is even more drastically affected when two or three Type I errors are made. In this case we are conducting tests of means receiving adjacent ranks for four-population experiments.

In all cases considered here, the Newman-Keuls technique becomes very conservative with respect to Type I errors once one or more Type I errors have been committed. This finding is in agreement with the author's a priori hypothesis that the distribution of the test statistic would have a smaller mean and variance under these conditions than when no Type I error was made.

As a further illustration of the departure of the actual test statistic distribution from that used to obtain the critical values we present, in Figures 1, 2, and 3, plots of the empirically generated distributions for $_2Q_{(2)-(1)}$, $_3Q_{(2)-(1)}$ and $_3Q_{(3)-(2)}$. These three distributions were chosen because it was possible to design, for example (Figure 1) a two-population experiment with 12 degrees of freedom, for which $q_{2,12}$ is the theoretically correct test statistic distribution, and a three-population experiment which would result in the use of $q_{2,12}$ after having made a Type I error.

Table 3

Upper Tail Area Frequencies: Contrasts from 3-Group Experiments and Comparison Experiments with 2 Groups

Row No.	Statistic	v	Interval in Studentized Range Distribution							No. of Previous Type I Frrors
			$Q \leq P_{90}$	$P_{90} < Q \leq P_{95}$	$P_{95} < Q \leq P_{97.5}$	$P_{97.5} < Q \leq P_{99}$	$P_{99} < Q \leq P_{99.5}$	$P_{99.5} < Q \leq P_{99.9}$	$P_{99.9} < Q$	
			Expected Frequencies							
			9000	500	250	150	50	40	10	
1	$_3Q(3)-(1)$	12	8996	502	232	154	65	36	15	0
2	$_3Q(3)-(2)$	12	9576	224	119	56	13	10	2	1
3	$_3Q(2)-(1)$	12	9576	239	103	47	15	18	2	1
4	$_2Q(2)-(1)$	12	8980	525	251	158	43	35	8	0
5	$_3Q(3)-(1)$	27	8960	514	258	160	53	45	10	0
6	$_3Q(3)-(2)$	27	9599	231	101	47	15	5	2	1
7	$_3Q(2)-(1)$	27	9613	231	103	33	10	8	2	1
8	$_3Q(3)-(1)$	42	8975	514	260	146	59	40	6	0
9	$_3Q(3)-(2)$	42	9586	251	103	45	12	3	0	1
10	$_3Q(2)-(1)$	42	9654	208	82	38	9	9	0	1
11	$_2Q(2)-(1)$	42	9000	491	252	133	66	52	6	0
12	$_3Q(3)-(2)$	36	9633	210	94	43	12	7	1	1
13	$_3Q(2)-(1)$	36	9627	231	75	41	16	9	1	1

Table 4

Upper Tail Area Frequencies: Contrasts from 4-Group Experiments and Comparison Experiments with 3 Groups

Row No.	Statistic	v	Interval in Studentized Range Distribution							No. of Previous Type I Errors
			$Q \leq P_{90}$	$P_{90} < Q \leq P_{95}$	$P_{95} < Q \leq P_{97.5}$	$P_{97.5} < Q \leq P_{99}$	$P_{99} < Q \leq P_{99:5}$	$P_{99.5} < Q \leq P_{99.9}$	$P_{99.9} < Q$	
			Expected Frequencies							
			9000	500	250	150	50	40	10	
1	$_4Q(4)-(1)$	16	9003	490	248	151	52	49	7	0
2	$_4Q(4)-(2)$	16	9627	230	89	38	5	8	3	1
3	$_4Q(3)-(1)$	16	9633	211	81	51	14	9	0	1
4	$_4Q(4)-(3)$	16	9766	152	53	21	6	2	0	2
5	$_4Q(2)-(1)$	16	9772	144	51	28	4	1	0	2
6	$_4Q(3)-(2)$	16	9908	61	21	7	1	2	0	3
7	$_4Q(4)-(1)$	36	9010	499	267	136	46	31	11	0
8	$_4Q(4)-(2)$	36	9668	201	88	31	4	6	2	1
9	$_4Q(3)-(1)$	36	9693	198	63	33	7	5	1	1
10	$_3Q(3)-(1)$	36	8990	524	232	149	46	48	11	0
11	$_4Q(4)-(3)$	36	9782	135	50	23	6	3	1	2
12	$_4Q(2)-(1)$	36	9803	136	43	13	1	3	1	2
13	$_4Q(3)-(2)$	36	9940	47	11	1	1	0	0	3
14	$_4Q(4)-(1)$	56	9036	470	242	158	47	33	14	0
15	$_4Q(4)-(2)$	56	9697	177	79	32	9	3	3	1
16	$_4Q(3)-(1)$	56	9673	198	78	42	8	0	1	1
17	$_4Q(4)-(3)$	56	9806	128	45	15	3	2	1	2
18	$_4Q(2)-(1)$	56	9773	150	45	27	3	1	1	2
19	$_4Q(3)-(2)$	56	9934	47	13	6	0	0	0	3

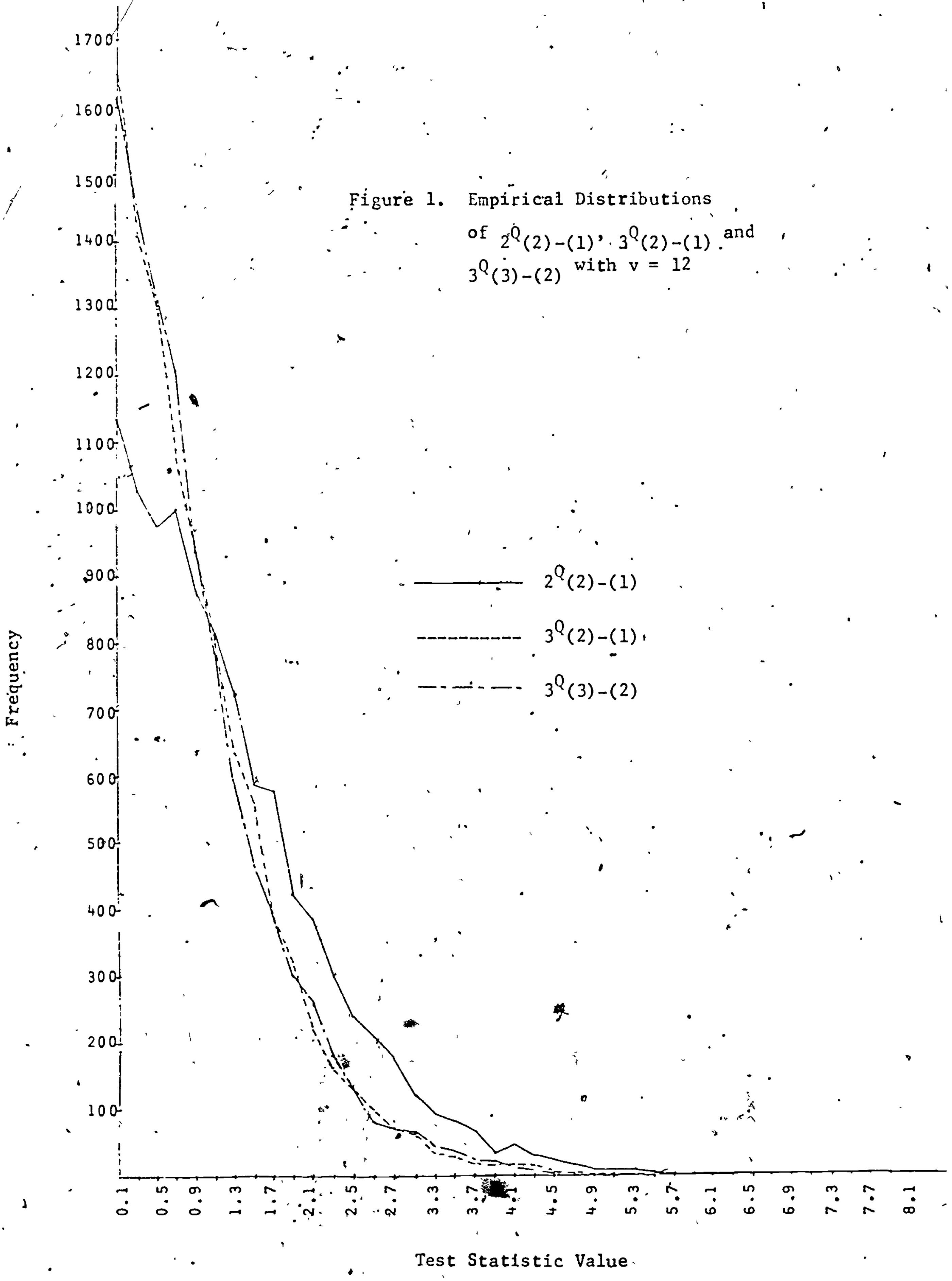


Figure 1. Empirical Distributions of $_2Q_{(2)-(1)}$, $_3Q_{(2)-(1)}$ and $_3Q_{(3)-(2)}$ with v = 12

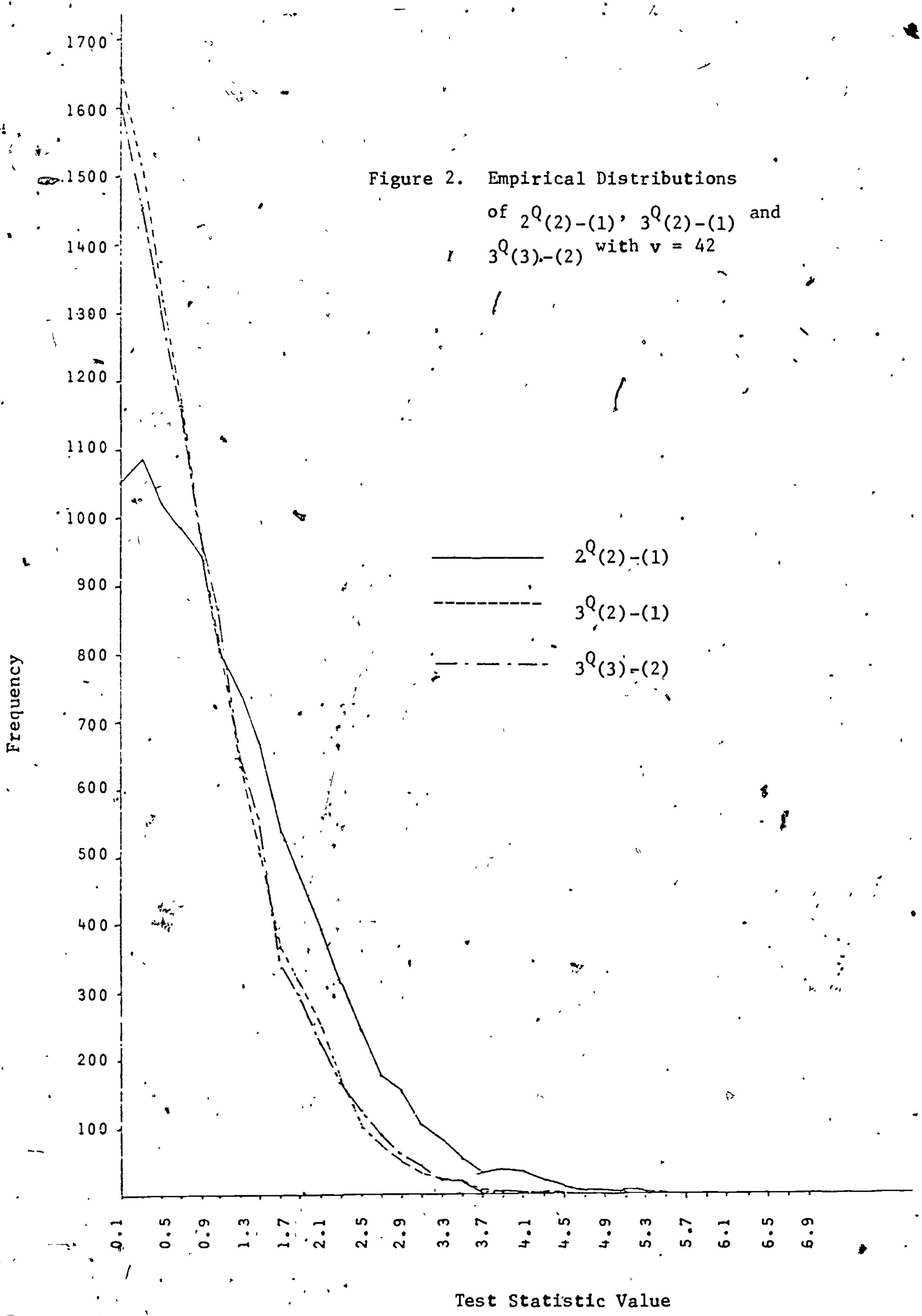
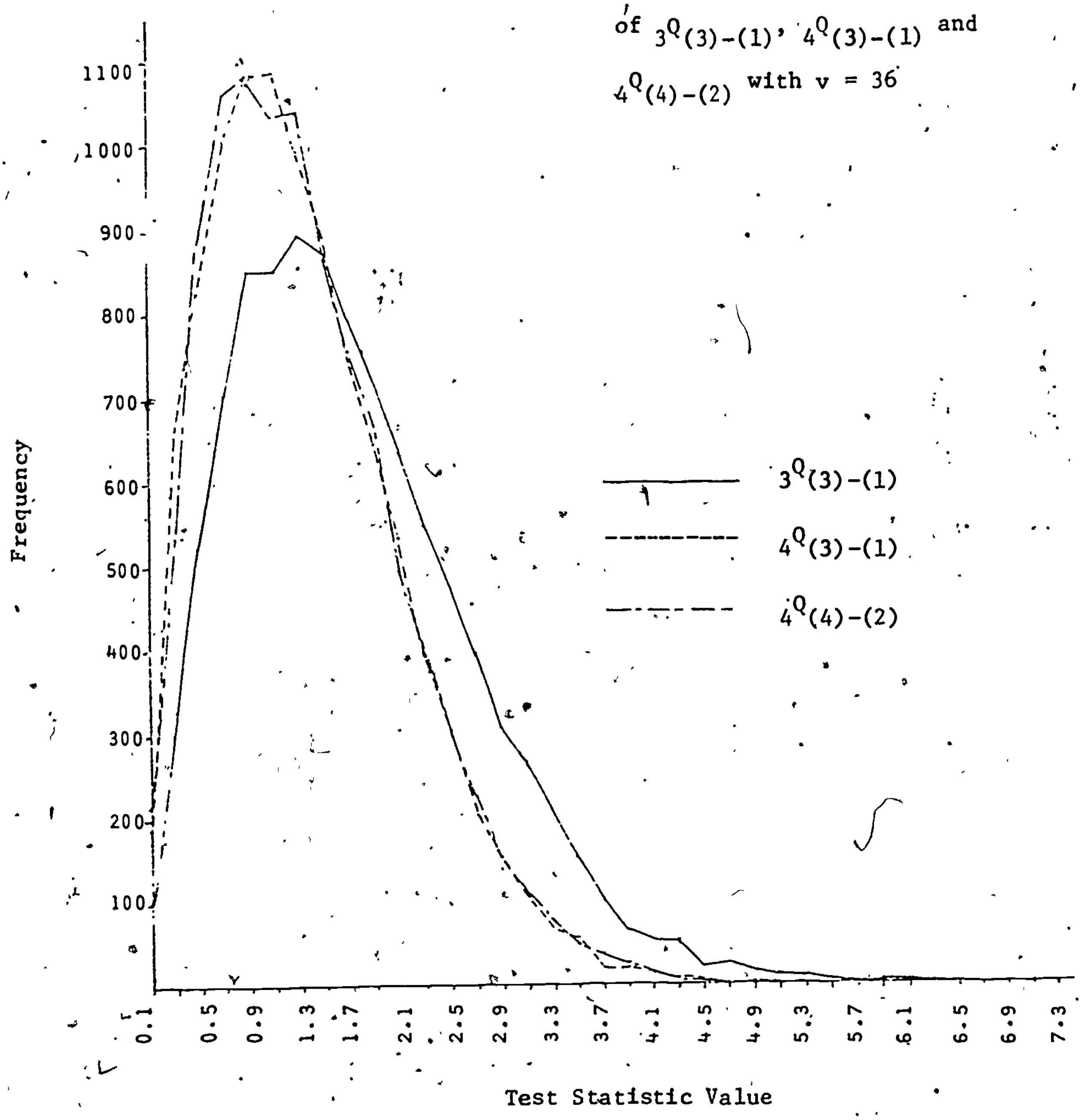


Figure 2. Empirical Distributions of $_2Q_{(2)-(1)}$, $_3Q_{(2)-(1)}$ and $_3Q_{(3)-(2)}$ with v = 42

Figure 3. Empirical Distributions of $_3Q_{(3)-(1)}$, $_4Q_{(3)-(1)}$ and $_4Q_{(4)-(2)}$ with v = 36



Comparisons of the actual means of the generated distributions with the expected values of the Studentized range distributions used in the Newman-Keuls technique are presented in Table 5. This Table has four sections. The first section contains the comparisons for the distributions that are theoretically correct (no Type I errors previously made). The other sections contain the comparisons for cases in which one, two and three Type I errors have been previously made and the expectations in these sections are those for the specific Studentized range distributions used in the Newman-Keuls technique.

Table 5

Comparison of Empirical Means and Their Expectations

Statistic	v	No. of Previous Type I Errors	Distribution Used	Mean	Expectation	Difference Mean-E(Mean)
$2^Q(2)-(1)$	12	0	$q_{2,12}$	1.2116	1.2056	.0060
$2^Q(2)-(1)$	42	0	$q_{2,12}$	1.1665	1.1490	.0175
$3^Q(3)-(1)$	12	0	$q_{3,42}$	1.8099	1.8084	.0015
$3^Q(3)-(1)$	27	0	$q_{3,27}$	1.7416	1.7415	.0001
$3^Q(3)-(1)$	36	0	$q_{3,36}$	1.7154	1.7289	-.0135
$3^Q(3)-(1)$	42	0	$q_{3,42}$	1.7191	1.7236	-.0045
$4^Q(4)-(1)$	16	0	$q_{4,16}$	2.1720	2.1620	.0100
$4^Q(4)-(1)$	36	0	$q_{4,36}$	2.0897	2.1029	-.0132
$4^Q(4)-(1)$	56	0	$q_{4,56}$	2.0767	2.0868	-.0101
$3^Q(2)-(1)$	12	1	$q_{2,12}$	.9079	1.2056	-.2977
$3^Q(3)-(2)$	12	1	$q_{2,12}$	.9019	1.2056	-.3037
$3^Q(2)-(1)$	27	1	$q_{2,27}$	.8722	1.1609	-.2887
$3^Q(3)-(2)$	27	1	$q_{2,27}$	.8694	1.1609	-.2915
$3^Q(2)-(1)$	36	1	$q_{2,36}$	.8642	1.1526	-.2894
$3^Q(3)-(2)$	36	1	$q_{2,36}$	.8512	1.1526	-.3014
$3^Q(2)-(1)$	42	1	$q_{2,42}$	.8501	1.1490	-.2989
$3^Q(3)-(2)$	42	1	$q_{2,42}$	.8690	1.1490	-.2800
$4^Q(3)-(1)$	16	1	$q_{3,16}$	1.4082	1.7774	-.3692
$4^Q(4)-(2)$	16	1	$q_{3,16}$	1.3932	1.7774	-.3842
$4^Q(3)-(1)$	36	1	$q_{3,36}$	1.3440	1.7289	-.3849
$4^Q(4)-(2)$	36	1	$q_{3,36}$	1.3529	1.7289	-.3760
$4^Q(3)-(1)$	56	1	$q_{3,56}$	1.3443	1.7157	-.3714
$4^Q(4)-(2)$	56	1	$q_{3,56}$	1.3396	1.7157	-.3761

Table 5 (cont.)

Statistic	v	No. of Previous Type I Errors	Distribution Used	Mean	Expectation	Difference Mean-E(Mean)
$4^Q(2)-(1)$	16	2	$q_{2,16}$	.7788	1.1850	-.4062
$4^Q(4)-(3)$	16	2	$q_{2,16}$	.7638	1.1850	-.4212
$4^Q(2)-(1)$	36	2	$q_{2,36}$	.7368	1.1526	-.4158
$4^Q(4)-(3)$	36	2	$q_{2,36}$	.7457	1.1526	-.4069
$4^Q(2)-(1)$	56	2	$q_{2,56}$	.7371	1.1438	-.4067
$4^Q(4)-(3)$	56	2	$q_{2,56}$	.7324	1.1438	-.4114
$4^Q(3)-(2)$	16	3	$q_{2,16}$	.6295	1.1850	-.5555
$4^Q(3)-(2)$	36	3	$q_{2,36}$	.6073	1.1526	-.5453
$4^Q(3)-(2)$	56	3	$q_{2,56}$	.6072	1.1438	-.5366

Values in the first section of Table 5 show that the means of the generated distributions are very close to their expectations, the largest deviation from expectation for the 9 distributions being only .0175. After one Type I error has been made, however, the mean of the empirical distribution differs from the expectation of the distribution used in the Newman-Keuls procedure by amounts ranging from .2800 to .3849 for the 14 different distributions. Similarly, the mean of the empirical distributions for the case of two previous Type I errors deviate from the expectations of the distributions actually used by amounts varying from .4062 to .4212 for six distributions, and the similar deviations for three distributions following three Type I errors range from .5366 to .5555.

These data indicate, as we had predicted, that the means of the distributions of the Newman-Keuls test statistic definitely do decrease following a Type I error on a previous test.

As a final basis for comparison of the distributions generated in this study we present, in Table 6, estimates of the means, variances, and indices of skewness (gamma-one) and kurtosis (gamma-two). For convenience of reading this Table presents the particular contrasts in the same order as used in Tables 3 and 4.

Table 6

Estimates of Parameters of the Distributions

Statistic	v	No. of Previous Type I Errors	Estimates of Parameters			
			Mean	Variance	Skewness	Kurtosis
$3^{Q}(3)-(1)$	12	0	1.8099	1.1353	1.2098	2.8572
$3^{Q}(3)-(2)$	12	1	.9019	.6025	1.5351	3.3147
$3^{Q}(2)-(1)$	12	1	.9079	.6077	1.5934	4.0802
$2^{Q}(2)-(1)$	12	0	1.2116	.9314	1.2794	2.2504
$3^{Q}(3)-(1)$	27	0	1.7416	.9274	.9130	1.2369
$3^{Q}(3)-(2)$	27	1	.8694	.5137	1.3637	2.4987
$3^{Q}(2)-(1)$	27	1	.8722	.5123	1.2731	1.9543
$3^{Q}(3)-(1)$	42	0	1.7191	.8723	.7398	.4990
$3^{Q}(3)-(2)$	42	1	.8690	.4971	1.2010	1.5185
$3^{Q}(2)-(1)$	42	1	.8501	.4831	1.2271	1.6903
$2^{Q}(2)-(1)$	42	0	1.1665	.7825	1.0681	1.1629
$3^{Q}(3)-(2)$	36	1	.8512	.4963	1.2901	2.0159
$3^{Q}(2)-(1)$	36	1	.8642	.4983	1.2745	2.0331
$4^{Q}(4)-(1)$	16	0	2.1720	1.0523	.8727	1.2383
$4^{Q}(4)-(2)$	16	1	1.3932	.6794	1.0277	1.6104
$4^{Q}(3)-(1)$	16	1	1.4082	.6784	1.0156	1.5548
$4^{Q}(4)-(3)$	16	2	.7638	.4262	1.4196	2.5623
$4^{Q}(2)-(1)$	16	2	.7788	.4306	1.3750	2.4693
$4^{Q}(3)-(2)$	16	3	.6295	.2991	1.5281	3.4331

Table 6 (Cont.)

Statistic	v	No. of Previous Type I Errors	Estimates of Parameters			
			Mean	Variance	Skewness	Kurtosis
$4^Q(4)-(1)$	36	0	2.0897	.8750	.6720	.5417
$4^Q(4)-(2)$	36	1	1.3529	.5896	.8581	.8951
$4^Q(3)-(1)$	36	1	1.3440	.5820	.8139	.7356
$3^Q(3)-(1)$	36	0	1.7154	.8963	.8356	.8442
$4^Q(4)-(3)$	36	2	.7457	.3968	1.3975	2.6537
$4^Q(2)-(1)$	36	2	.7368	.3877	1.2802	1.8960
$4^Q(3)-(2)$	36	3	.6073	.2590	1.2802	1.8353
$4^Q(4)-(1)$	56	0	2.0767	.8397	.6263	.4946
$4^Q(4)-(2)$	56	1	1.3396	.5633	.8328	.8877
$4^Q(3)-(1)$	56	1	1.3443	.5804	.7862	.5126
$4^Q(4)-(3)$	56	2	.7324	.3677	1.2768	1.9909
$4^Q(2)-(1)$	56	2	.7371	.3881	1.3349	2.0971
$4^Q(3)-(2)$	56	3	.6072	.2649	1.3264	2.1171

Examining first the estimates in Table 6 for the cases of theoretically correct distributions we can see several tendencies. For these cases the distributions are positively skewed and leptokurtic. As is to be expected the means and variances for the same number of groups (k) decrease as the degrees of freedom (v) increase, and increase as k increases when v remains constant. There is a definite tendency for the degrees of skewness and kurtosis to decrease as v increases for the same k and also a tendency for these indices to decrease as k increases with v held constant.

Next we examine those cases in which one Type I error has been committed previous to the test of a mean difference, and for which we were able to compute estimates of the parameters of the distribution that would be used in actually applying the Newman-Keuls technique. For example, we compare the estimates of parameters of $_3Q_{(2)-(1)}$ and $_3Q_{(3)-(2)}$ with those for $_2Q_{(2)-(1)}$ for the same number of degrees of freedom. On making these comparisons we see that, as we had originally hypothesized, the means and variances are smaller than tne values from the distribution that would be used in applying the Newman-Keuls procedure. There is also a tendency for the actual distributions generated under the condition that one Type I error has been made to be more skewed and more leptokurtic than the theoretical distributions. In all cases the tendency is for all estimates to decrease in value as v increases for the same k, for the mean and variance of these distributions to decrease as k increases for the same v; and for the indices of skewness and kurtosis to decrease as k increases for the same v.

Further examination of the data reveals that when more than one Type I error has been committed the mean and variance decrease even further but the indices of skewness and kurtosis appear not to be greatly affected.

8. Discussion

As had been expected by the authors prior to generating the data reported in this paper, the distribution of the Newman-Keuls test statistic changes shape after a Type I error is committed. Also, as expected this change results in an actual Type I error rate that is smaller than the significance level chosen by the researcher. This means, of course, that the Newman-Keuls procedure becomes rather conservative when one Type I error has been made, and even more conservative following additional Type I errors. Thus the chance of making two or more Type I errors is lower than would be expected if the distributions were not affected by the results of the previous significance tests.

It is difficult to attempt to relate our findings to those of Petrinovich and Hardyck (1969) and Carmer and Swanson (1973) because there was no reason for them to estimate separately the error rates of second and third comparisons after making errors on the initial comparison. Our results do, of course, relate to the fact that Petrinovich and Hardyck found the experimentwise Type I error rate for the case of all populations having identical means to be better than that for mixtures of populations.

Spjøtvoll (1974) in his section on the justification of the Newman-Keuls technique makes the statement, "At step 2 we *assume* (italics ours) that the conclusion reached at step 1 is correct [p. 100]." Our findings point out the importance of this assumption.

References

Bennett, C. A., & Franklin, N. L. Statistical analysis in chemistry and the chemical industry. New York: Wiley, 1954.

***

Carmer, S. G., & Swanson, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 1973, 68, 66-74.

Games, P. A. Inverse relation between the risks of type I and type II errors and suggestions for the unequal n case in multiple comparisons. Psychological Bulletin, 1969, 71, 43-54.

Hammersley, J. M., & Handscomb, D. C. Monte Carlo Methods. London: Methuen, 1964.

Harter, L. H.. Tables of range and studentized range. Annals of Mathematical Statistics, 1960, 31, 1122-1147.

Keuls, M. The use of the "studentized range" in connection with an analysis of variance. Euphytica, 1952, 1, 112-122.

Kirk, R. E. Experimental design: procedures for the behavioral sciences. Belmont, Cal.: Brooks-Cole, 1968.

MacLaren, M. D., & Marsaglia, G. Uniform random number generators. Journal of the Association for Computing Machinery, 1965, 12, 83-89.

Mendenhall, W. Introduction to linear models and the design and analysis of experiments. Belmont, Cal.: Wadsworth, 1968.

Miller, R. G. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

Miller, R. G. Personal communication, 1971.

Newman, D. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. Biometrika, 1939, 31, 20-30

Newman, T. G., & Odell, P. L. The generation of random variates. New York: Hafner, 1971.

O'Neill, R., & Wetherill, G. B. The present state of multiple comparison methods. Journal of the Royal Statistical Society, Series B, 1971, 33, 218-241.

Petrinovich, L. F., & Hardyck, C. D. Error rates for multiple comparison methods: some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

***Insertion—(see page 26)

Pingel, K. R. The generation of random numbers from chi-square distributions with small degrees of freedom. Unpublished M. A. Thesis, University of Pittsburgh, 1975.

Spjøtvoll, E. Multiple testing in the analysis of variance. Scandinavian Journal of Statistics, 1974, 1, 97-114.

Winer, B. J. Statistical principles in experimental design (2nd ed.) New York: McGraw-Hill, 1971.

***

Box, G. E. P. & Muller, M. E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.