WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

MULTIPLE COMPARISON PROCEDURES BASED ON GAPS*

Roy E. Welsch

October 1975 WP 816-75

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

MULTIPLE COMPARISON PROCEDURES BASED ON GAPS*

Roy E. Welsch

October 1975 WP 816-75

Massachusetts Institute of Technology

National Bureau of Economic Research

* This research was supported in part by NSF Grant GJ-1154x3 to the National Bureau of Economic Research.

ABSTRACT

This paper discusses four sequential multiple comparison significance tests and compares them with some existing multiple comparison procedures. Two of the proposed tests begin by examining the gaps between adjacent ordered sample means, then the three-stretches, four-stretches and so on until the range is reached. The remaining two tests reverse this procedure. All four are designed to control the experimentwise type I error rates.

Tables for the gap tests were constructed using improved Monte Carlo techniques. A simulation study showed that one of the gap tests proved to be the best and provided significantly greater power than the commonly used Tukey Honestly Significant Difference procedure.

MULTIPLE COMPARISON PROCEDURES BASED ON GAPS

1. Introduction

The purpose of this paper is to propose several multiple comparison (MC) procedures based on gaps and to compare their performance with some commonly used MC procedures. In order to avoid an extended philosophical discussion, we state now that we are taking a non-Bayesian and significance oriented approach to multiple comparisons. We tend to use MC procedures for data exploration, so we will emphasize experimentwise error rates.

Why gaps first? Well, for one, this author always seems to notice the gaps first in an ordered set of treatment means. Second, a number of MC procedures fail when applied to synthetic data with large spacings between, say, pairs of equal population means. For example, a protected LSD (the FSD2 of Carmer and Swanson [1973], hereafter C and S) has a high experimentwise type I error rate (E1) in such situations. In short, a protected LSD over-emphasizes the very special null hypothesis of all population means equal at the expense of hypotheses where subgroups of population means are equal. Looking at gaps first has the effect of giving these subgroup hypotheses more attention. Preliminary investigations indicated that compared with the Tukey HSD (TSD in C and S), a gap procedure would be more powerful for a given E1.

Finally, we had noticed an additional problem with the Newman-Keuls (SNK in C and S) procedure which, incidentally, suffers from the same defect as the protected LSD when there are subgroups of equal means. The SNK procedure requires that if a group of ordered means is not declared significant, then we are barred from looking at any subgroups of that group. If we were to look at gaps first (and then groups of three, etc.), we would say that if a group is

significant, then all groups containing that group are significant. Since in significance testing we tend to emphasize disproving the null hypothesis, this is a more appealing way to develop a sequential significance procedure.

The paper is organized as follows. The next section discusses error rates, the third discusses the design of sequential multiple comparison procedures, and the fourth and fifth discuss the new MC procedures. Section six contains the comparison of the new tests with existing ones. The appendices discuss the details of the Monte Carlo used to obtain the tables and to compare the various procedures. Finally, there is an analysis of the tables to show how they may be reduced in size by certain approximations.

The author would like to acknowledge many helpful conversations with John Tukey, David Hoaglin, Paul Holland and John Hartigan. David Jones provided invaluable programming assistance.

2. Multiple Comparisons in General

We view multiple comparison procedures as mainly appropriate for the exploration of data rather than for decision-making. In particular, significance oriented procedures should give us hints and clues about the ordering of the underlying populations with respect to some attribute. These we will use to think about what may be going on and which experiments to perform next.

Why then should any one configuration of means be particularly sacred? We are interested in contemplating those configurations which remain after the data has given us an indication about those we should reject. If we are interested in specific configurations, then we should concentrate our statistical power on them. But, if we are exploring, how can we risk considering just a few alternatives?

The literature contains extensive discussions about errors and error rates. Useful references are Kurtz, et. al. (1965), Miller (1966), and O'Neill and Wetherill (1971). For those interested in exploring data, a particular definition of error and error rate provides a way to compute a set of critical values and perhaps make some power calculations. It is certainly conceivable that a data analyst might use more than one set of critical values in analyzing a particular batch of data, weighing the results in light of the definition of error and error rate used to determine each set of critical values.

We distinguish three types of errors:

Type I: Two population means are declared significantly different when they are, in fact, equal.

Type II: Two population means are not declared to be significantly different when they in fact are.

Type III: Two population means are declared significantly different when they are in fact different, but the order is reversed.

These errors are considered in two ways--experimentwise and comparisonwise. An experiment is the determination of a sample mean value for each of the populations under consideration. The experimentwise type I error rate (E1) is defined as the number of experiments with one or more type I errors divided by the number of experiments. The comparisonwise type I error rate (C1) is defined as the number of type I errors divided by the number of comparisons.

We should note that some authors define type I errors to include both type I and type III errors.

3. Sequential Multiple Comparison Significance Tests

We will be considering MC significance tests which have the following structure. Let $m_1 \leq m_2 \leq \ldots \leq m_t$ be the ordered treatment determinations that we wish to compare. Then we seek a set of critical values (gap gages), C_k, and a scale measure, s, such that:

i. The quantity sC_k is used to measure all k-stretches $(m_{i+k} - m_i,\ i=1,\ldots,t-k)$. If a k-stretch exceeds sC_k it is declared significant.

ii. If a k-stretch is declared significant, then all h-stretches containing that k-stretch are declared significant.

iii. If a k-stretch is not declared significant, no h-stretch contained in that k-stretch can be declared significant.

The order relations two and three are automatically satisfied when the C_k are equal for all k. Thus the LSD and HSD satisfy the three conditions, but the LSD fails to control E1. (It is designed to control C1.)

The protected LSD (e.g., FSD2) satisfies the above criteria but a special check is made first with an F-test (or a test on the range) before the MC procedure is applied.

The SNK procedure chooses different C_k such that $C_t \geq C_{t-1} \ldots \geq C_2$ and examines the t-stretch first, then the (t-1)-stretch etc. and enforces rules two and three by not checking h-stretches within a non-significant k-stretch.

At this point, it is possible to see some of the reasons why a test starting with the 2-stretches would be appealing. Looking at the 2-stretch first, then the 3-stretches, etc., makes it natural to declare a k-stretch significant when it contains a significant h-stretch, $h < k$. This is enough to satisfy rules two and three.

4. Gap Tests

We would like to find a procedure starting with the gaps that allows us to control, at predetermined levels, the EI. Why not use the HSD? When t=3 and σ is known, we found critical numbers $C_2 < C_3$ analytically from the bivariate Gaussian distribution. These critical numbers were then used in a small Monte Carlo which showed that a gap test would be generally more powerful than the HSD for a given EI level.

When $t > 3$ we need special methods to get the critical numbers for a gap test. We assume that the data comes from t independent Gaussian populations $M_1, M_2, \ldots, M_t$ with population means $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_t$ and variance σ^2. (Note that m_1, the smallest sample determination, does not necessarily come from M_1.) There are many possible configurations of the true means, but for our purposes we need only consider whether two means are equal or not equal and we shall set $\mu_1 = 0$. This implies that μ_1, μ_2 etc. are partitioned into blocks of equal population means with the blocks possibly ranging in size from 1 to t.

Let $(X_1, \ldots, X_k)$ denote k ordered independent variates from $G(0,\sigma^2)$, S^2 an independent estimator for σ^2, and

$$T_i(k) = \max_{j=1,k-i+1} (X_{j+i-1} - X_j) \quad .$$

We will use H to denote any configuration of true means having at least one block with more than one mean. Denote the blocks by $B_1(d_1), B_2(d_2), \ldots, B_q(d_q)$ where d_i refers to the number of means in the

i^{th} block. Let

$$P_{d_i} = P\{T_{d_i}(d_i) > SC_{d_i} \text{ or } T_{d_{i-1}}(d_i) > SC_{d_{i-1}}$$

$$\text{or } \ldots \text{ or } T_2(d_i) > SC_2\} \quad .$$

Theorem 1. For a gap procedure with $C_2 \leq C_3 \leq \ldots \leq C_t$ we have,

$$P\{\text{one or more type I errors } |H\} \leq P_{d_1} + P_{d_2} + \cdots + P_{d_q}.$$

Proof. First we check the gaps in the ordered sample determinations $m_1 \leq m_2 \leq \ldots \leq m_t$ with scale estimate, s. A type I error can only be made if, when $d_i \geq 2$, $T_2(d_i) > sC_2$. Next look at the three-stretches. If $d_i \geq 3$, an error possibly occurs when $T_3(d_i) > sC_3$ or $T_2(d_i) > sC_3$, since determinations from other blocks could lie between two determinations from the i^{th} block. If $d_i=2$, then an error may occur when $T_2(d_i) > sC_3$. We require that $C_3 \geq C_2$ so that the event $T_2(d_i) > sC_3$ is included in $T_2(d_i) > sC_2$. The four-stretches may be treated in a similar way. Thus $P_{d_1} + P_{d_2} + \ldots P_{d_q}$ is an upper bound on El.

The key quantities are obviously the P_{d_i}. Theorem 1 tells us that to control E1 at level α for all hypotheses, H, we must have

$$\sum_{i=1}^{q} P_{d_i} \leq \alpha$$

for all sequences $\{d_i\}$ such that $\sum_{i=1}^{q} d_i \leq t$, and $C_2 \leq C_3 \leq \ldots \leq C_t$.

We would like to choose the C_i in order to achieve maximum power for a given level α. We would also like relatively simple tables. We considered:

A. $P_j = \frac{j}{t} \cdot \alpha \quad j=2,\ldots,t-2, t \quad$ with $P_{t-1} = \alpha$ and

B. $P_j = \frac{j}{t} \cdot \alpha \quad j=2,\ldots,t$

We suspected that A would be more powerful but that B would lead to smooth and simple tables. We shall call these procedures GAPA and GAPB. In Appendix A we discuss how tables were computed for these tests.

5. Modified Newman-Keuls Procedures

So far we have focused on gap tests that look like a reversal of the Newman-Keuls (NK) procedure. It is natural at this point to think of modifying the SNK procedure in order to control E1. To show how to do this we prove a result analogous to Theorem 1. For simplicity we set $RP_{d_i}=P\{T_{d_i}(d_i) > SC_{d_i}\}$.

Theorem 2. For an NK procedure with $C_t \geq C_{t-1} \geq \ldots \geq C_2$ we have,

$$P\{\text{one or more type I errors } |H\} \leq RP_{d_1} + \cdots + RP_{d_q} .$$

Proof. First we check the range. We can only make an error if $T_{d_i}(d_i) > sC_t$ for some i. Since $C_t \geq C_{d_i}$ for all i, we have included these errors. Next we have two cases, $d_1=t$ or no $d_i=t$. We only look at (t-1)-stretches if the range has exceeded sC_t. If $d_1=t$, we will never look at any k-streches with $k<t$ unless we have already made an error (i.e. the range exceeded sC_t). Hence we do not need to count errors made by checking (t-1)-stretches within a group of t equal means.

If $d_1 \neq t$, then we have not made a type I error at the first stage so we consider the (t-1)-stretches. An error can only be made if $T_{d_i}(d_i) > sC_{t-1}$ for some i. Since $C_{t-1} > C_{d_i}$ when $d_i \leq t-1$, this error is counted in the statement of the theorem. This process continues for (t-2)-stretches, etc. in an analogous manner.

The key quantities for this test are the RP_{d_i}. To control the type I error rate we must have

$$\sum_{i=1}^{q} RP_{d_i} \leq \alpha$$

for all sequences $\{d_i\}$ with $\sum_{i=1}^{q} d_i \leq t$, and $C_t \geq C_{t-1} \geq \ldots \geq C_2$.

We allocate the error as with the gap tests and call these tests WNKA and WNKB.

6. Comparison of the Tests

We used the Monte Carlo study by Carmer and Swanson [1973] as a basis for our comparison of the new procedures. A macro of commands from the NBER TROLL system running on an IBM 360/67 was used for the simulation experiment.

In order to overlap one of the sampling situations of C and S, we chose t=5 with 20 error degrees of freedom (i.e., a two-way ANOVA with six replications). The data was generated by using a Gaussian random generator developed by Marsaglia et al. (1972). Five sets of 4000 replications each from G(0, 1/6) were drawn and then adjusted to represent the various configurations in Table 1. (This implies that the ANOVA model has $\sigma^2=1$.) New seeds were used for each of the eight configurations.

[Table 1 about here.]

The scale was generated from a χ^2 generator with 20 degrees of freedom (a sum of squared Gaussian variates obtained using the method of Box and Muller [1958]) with a different uniform driver from that used for the Marsaglia generator.

The variability of the results was measured by dividing the 4000 samples into 10 batches of 400 and finding the standard error of the results over the 10 batches.

For benchmark purposes we included our own proposed tests plus the LSD, HSD(TSD), and SNK. Since there is no indication of sampling error in C and S, it is difficult to tell if our results are within the C and S sample error. With the possible exception of the SNK procedure, our results are in reasonable agreement with C and S if we use our own measure of standard error.

TABLE 1

CONFIGURATIONS OF TRUE MEANS

Set	True means
1*	0 0 0 0 0
2	0 0 0 .2 .2
3	0 0 0 .5 .5
4	0 0 0 1 1
5	0 0 0 2 2
6	0 0 0 3 3
7*	-.5 -.5 0 .5 .5
8*	-1 -1 0 1 1

*#1 is equivalent to #1 of C and S
#7 is equivalent to #3 of C and S
#8 is equivalent to #5 of C and S

To save space we have not listed the results for WNKB since it was the worst of the four procedures we proposed.

6a. Type I Error Rates

Table 2 shows the experimentwise and comparisonwise type I error rates when all five population means are equal and $\alpha=.05$. Of more interest are the E1 rates for other configurations (all means not equal) listed in Table 3. We see that for the new tests this error rate is less than five percent to within standard error as we have proved it should be. We note that this is not the case for SNK.

[Tables 2 and 3 about here.]

6b. Type III Error Rates

We found that E3 rates were very small relative to the E1 rates. Our results agreed with those described in C and S. Clearly the HSD is designed to control both type I and type III error. We cannot prove such a result for the tests we propose, except in very special cases. We conjecture that such a result is true generally.

6c. Type II Error Rates

Since all of the tests we have proposed control the E1 rate, we are most interested in comparing them on the basis of power. For non-zero true differences we use the C and S definition of power which is

100 - Type II error.

[Table 4 about here.]

TABLE 2

E1 AND C1 ERROR RATES FOR EQUAL MEANS

Error	Procedure					
	HSD	GAPA	GAPB	WNKA	SNK	LSD
E1	4.60	4.55	4.70	4.73	4.73	23.53
Standard Error	.28	.25	.30	.31	.31	2.11
C1	.69	.85	.79	.82	.91	4.51
S.E.	.05	.06	.05	.06	.07	.42

TABLE 3

E1 ERROR RATES FOR UNEQUAL MEANS

Set	Procedure					
	HSD	GAPA	GAPB	WNKA	SNK	LSD
2	2.43	2.90	2.70	2.75	2.93	15.45
3	2.53	3.58	3.23	3.53	4.18	15.55
4	2.73	4.53	4.23	4.35	7.20	17.70
5	2.63	5.08	5.08	5.05	9.48	16.20
6	2.33	4.15	4.15	4.18	9.28	16.23
7	1.13	2.30	2.20	2.13	4.08	9.50
8	1.18	3.00	3.00	2.90	7.65	9.08

Approximate standard error is .4

The results are listed in Table 4 for various values of δ_{ij}/σ where δ_{ij} is the true difference between μ_i and μ_j ($|\mu_i-\mu_j|$) and σ is the population standard deviation, which we took to be 1. These results are based only on configurations 2 through 6 in Table 1.

As we expected GAPA dominates GAPB. All of our proposed tests improved upon the HSD with GAPA the winner. Clearly WNKA is not far behind, and since it requires somewhat easier to compute tables (see Appendix A) we cannot completely set it aside (as we thought we might before running the simulation). Strictly interpreted, these results apply only to the types of configurations listed in Table 1.

We conclude that when we desire to control E1, sequential procedures beat the HSD with the gap procedures having a slight edge over Newman-Keuls type procedures. The tables for sequential procedures are more complicated but the approximations developed in Appendix B can make them compact and easy to use. This author hopes that sequential procedures which control E1 will receive serious consideration from the statistics community.

TABLE 4

100-TYPE II ERROR RATE

δ_{ij}/σ	Procedure					
	HSD	GAPA	GAPB	WNKA	SNK	LSD
.2	.97	1.16	1.11	1.13	1.24	6.52
.5	2.81	3.39	3.13	3.31	3.52	13.09
1	14.05	17.07	16.03	16.80	18.17	38.13
2	67.29	75.70	74.77	74.95	80.22	91.01
3	97.69	99.10	99.08	99.08	99.64	99.88

Approximate standard error is .4

References

1. Arnold, H. J., Bucher, B. D., Trotter, H. F. and Tukey, J. W. (1956). "Monte Carlo techniques in a complex problem about normal samples." Symposium on Monte Carlo Methods, ed. H. A. Meyer, 80-88. New York: Wiley.

2. Box, G. E. P. and Muller, M. E. (1958). "A Note on the Generation of Normal Deviates," Annals of Mathematical Statistics, 29, 610-11.

3. Carmer, S. G. and Swanson, M. R. (1973). "An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods". JASA, 68, 66-74.

4. Harter, H. L., Clemm, D. S., and Guthrie, E. H. (1959). The probability integrals of the range and of the studentized range-probability integral and percentage points of the studentized range; critical values for Duncan's new multiple range test. Wright Air Development Center Technical Report 58-484, Vol. II. (ASTIA Document No. AD231733).

5. Kurtz, T. E., Link, R. F., Tukey, J. W. and Wallace, D. L. (1965). Short-cut multiple comparisons for balanced single and double classifications: Part 1, Results. Technometrics, 7, 95-165.

6. Lewis, P. A., Goodman, A. S., and Miller, J. M. (1969). "A Pseudo-random Number Generator for the System/360. IBM Systems Journal 8, 136-146.

7. Marsaglia, G., Ananthanarayanan, K. and Paul, N. (1972). The McGill Random Number Package "Super-Duper". Unpublished notes and program distributed at the 1972 ASA meeting in Montreal.

8. Miller, R. G. (1966). Simultaneous Statistical Inference. New York: McGraw-Hill.

9. O'Neill, R. and Wetherill, G. B. (1971). The present state of multiple comparison methods. JRSS B, 33, 218-250.

10. Relles, D. A. (1970). Variance reduction techniques for Monte Carlo sampling from Student distributions. Technometrics, 12, 499-515.

11. Thomas, D. A. H. (1974). "Error Rates in Multiple Comparisons among Means-Results of a Simulation Exercise". Applied Statistics, 23, 284-294.

12. Trotter, H. F. and Tukey, J. W. (1956). "Conditional Monte Carlo for normal samples," Symposium on Monte Carlo Methods, ed. H. A. Meyer, 64-79. New York: Wiley.

13. Welsch, R. E. (1965). Conditional Monte Carlo and measures of significance for a new multiple range test. Unpublished senior thesis, Department of Mathematics, Princeton University.

14. Welsch, R. E. (1972). A modification of the Newman-Keuls procedure for multiple comparisons. Working Paper 612-72, Sloan School of Management, Massachusetts Institute of Technology.

15. Welsch, R. E. (1972). A Multiple Comparison Procedure Based on Gaps. Working Paper 628-72, Sloan School of Management, Massachusetts Institute of Technology.

APPENDIX A

A. Computation of the Critical Numbers

We first consider the gap tests. There does not seem to be a feasible computational way to find the C_i, (i=2,...,t) simultaneously so, for a given E1 level α (we used .05), ν (error degrees of freedom) and t, we found C_2, then C_3, etc. We will suppress the dependency of C_i on t, ν, and α in our notation.

Given t, ν, α and the critical numbers $C_2, C_3, \ldots, C_{k-1}$, $k \leq t$, our problem is to find C_k such that

$$P\{T_2(\underset{\sim}{X}) > SC_2 \text{ or } T_3(\underset{\sim}{X}) > SC_3 \text{ or } \ldots \text{ or } T_k(\underset{\sim}{X}) > SC_k\} = P_k \qquad (A.1)$$

where, in order to emphasize the $\underset{\sim}{X}$, we put $T_i(\underset{\sim}{X})$ in place of $T_i(k)$. Since the C_i are independent of σ^2 we set $\sigma^2=1$ and assume that νS^2 is distributed independently of $\underset{\sim}{X}$ as χ^2_ν. If C_k turns out to be less than C_{k-1} we shall set $C_k = C_{k-1}$ in order to preserve the ordering of the C_i's.

We propose to find C_k by evaluating

$$P\{T_2(\underset{\sim}{X}) > SC_2 \text{ or } \ldots \text{ or } T_k(\underset{\sim}{X}) > Sb\} \qquad (A.2)$$

for several values of b and using inverse interpolation for the C_k corresponding to P_k. Now (A.2) is equivalent to

$$P\{T_2(\underset{\sim}{X}) > SC_2 \text{ or } \ldots \text{ or } T_{k-1}(\underset{\sim}{X}) > SC_{k-1} \text{ and } T_k(\underset{\sim}{X}) \leq Sb\} \qquad (A.3)$$

$$+ P\{T_k(\underset{\sim}{X}) > Sb\} .$$

Since $T_k(\underset{\sim}{X})$ is just the range of $\underset{\sim}{X}$ the second term in (A.3) can be obtained directly from tables of the studentized range.

Let $\underset{\sim}{1}$ denote a vector of ones. An interesting property of the statistics, $T_j(\underset{\sim}{X})$, is that for any scalar λ,

$$T_j(\lambda\underset{\sim}{X}) = \lambda T_j(\underset{\sim}{X})$$

and

$$T_j(\underset{\sim}{X}-\lambda\underset{\sim}{1}) = T_j(\underset{\sim}{X}) \quad .$$

In other words, T_j is equivariant with respect to scale and invariant with respect to location. When this situation arises, Relles (1970) noticed that a considerable reduction in Monte Carlo sampling error can be obtained by considering a standardized configuration such as $\underset{\sim}{c}(\underset{\sim}{X}) = (\underset{\sim}{X}-X_1\underset{\sim}{1})/R$ where $R = X_k - X_1$.

If we condition the first part of (A.3) with respect to this configuration, we have

$$\begin{aligned}
&P\{T_2(\underset{\sim}{X}) > SC_2 \text{ or } \ldots \text{ or } T_{k-1}(\underset{\sim}{X}) > SC_{k-1} \text{ and } T_k(\underset{\sim}{X}) \leq Sb \,|\, \underset{\sim}{c}(\underset{\sim}{X})\} \\
&= P\{T_2\left(\frac{\underset{\sim}{X}-X_1\underset{\sim}{1}}{R}\right) > \frac{SC_2}{R} \text{ or } \ldots \text{ and } T_k\left(\frac{\underset{\sim}{X}-X_1\underset{\sim}{1}}{R}\right) \leq \frac{Sb}{R} \,\Big|\, \underset{\sim}{c}(\underset{\sim}{X})\} \\
&= P\{T_2(\underset{\sim}{c}(\underset{\sim}{X})) > \frac{SC_2}{R} \text{ or } \ldots \text{ and } T_k(\underset{\sim}{c}(\underset{\sim}{X})) \leq \frac{Sb}{R} \,\Big|\, \underset{\sim}{c}(\underset{\sim}{X})\} \\
&= P\{\min_{j=2,k-1}\left(\frac{C_j}{T_j(\underset{\sim}{c}(\underset{\sim}{X}))}\right) < \frac{R}{S} \leq b \,|\, \underset{\sim}{c}(\underset{\sim}{X})\}
\end{aligned} \tag{A.4}$$

since $T_k(\underset{\sim}{c}(\underset{\sim}{X})) \equiv 1$. The quantity R/S is just the studentized range and conditional on $\underset{\sim}{c}(\underset{\sim}{X})$ we can compute (A.4) from tables of the studentized range by using an appropriate interpolation procedure. Then the integral over $\underset{\sim}{c}(\underset{\sim}{X})$ can be obtained by simple random sampling. We used 1000 samples.

Since $\underset{\sim}{c}(\underset{\sim}{X}) = (0,Y_2,\ldots,Y_{k-1},1)$ where $0 \le Y_2 \le Y_3 \le \ldots \le Y_{k-1} \le 1$ we may as well take the Y_i to be a random sample of k-2 ordered independent variates from the uniform distribution on [0,1], to be called OUID, and then use weights to convert to the configuration space $\underset{\sim}{c}(\underset{\sim}{X})$. The probability density associated with sampling k variates from OGID (ordered Gaussian) is

$$\frac{k!}{(2\pi)^{k/2}} \exp[-\frac{1}{2}(\sum_{i=1}^{k} x_i^2)]dx_1,\ldots,dx_k. \qquad (A.5)$$

Transforming to $y_i=(x_i-x_1)/r$, i=2,...,k-1 with $r=x_k-x_1$, the probability element (A.5) becomes

$$\frac{k!}{(2\pi)^{k/2}} r^{k-2}\exp\{-\frac{1}{2}[(\sqrt{k}\, x_1 + \frac{r}{\sqrt{k}}(y_2+\cdots+y_{k-1}+1))^2$$

$$+ u(\underset{\sim}{y})r^2]\}dx_1dy_2\ldots dy_{k-1}dr$$

with $\underset{\sim}{y}=y_2,\ldots,y_{k-1}$ and $u(\underset{\sim}{y})=y_2^2+y_3^2+\cdots+1-(y_2+y_3+\cdots+1)^2/k$. To find the probability element for $\underset{\sim}{c}(\underset{\sim}{X})$ in configuration space we integrate over x_1 (a Gaussian integral) and r (a Gamma integral) to obtain

$$\frac{\sqrt{k}\,(k-1)\Gamma((k-1)/2)}{2[\pi u(\underset{\sim}{y})]^{(k-1)/2}} \cdot (k-2)!\, dy_2 \ldots dy_{k-1} \cdot$$

We can now perform sampling in the configuration space by sampling OUID with weights

$$w(\underset{\sim}{y}) = \frac{\sqrt{k}\,(k-1)\Gamma((k-1)/2)}{2(\pi u)(\underset{\sim}{y}))^{(k-1)/2}} .$$

The first step in constructing the tables of C_k was to find C_2 to an accuracy of one unit in the fourth significant digit by using the method of inverse interpolation described in Harter (1959) on the tables of the studentized range contained in the same report. The subroutine ALI from the IBM Scientific Subroutine Package was used for the direct interpotation.

The rest of the computation was carried out sequentially on k starting with k=3. Assume that the computation has been completed for k-1. Then we would have 1000 sets of k-3 ordered uniform pseudo-random numbers available from the k-1st step. The numbers were generated on an IBM 360/165 using a multiplicative congruential generator $Z_{n+1}=aZ_n(\text{mod } p)$, n=0,1,2,... with $p=2^{31}-1$, $a=16807=7^5$, and with starting value $Z_0=524287$ when k=3. (For more details see Lewis et al. (1969).) So for the kth step we generate 1000 more numbers and add one to each of the 1000 sets of k-3 variates, reorder, and call these samples $\underset{\sim}{y}(i)$, i=1,2,...,1000. For k=10 we had generated a total of 8000 pseudo-random numbers.

It is convenient at this point to let $\hat{\underset{\sim}{y}}(i) = (0, \underset{\sim}{y}(i), 1)$. Our next step was to compute $T_j(\hat{\underset{\sim}{y}}(i))$, j = 2,...,k-1 and $w(\underset{\sim}{y}(i))$. These numbers were then used for t=k(1)10 and all ν.

For a given set of k, t, and ν we found which tabled value, $b_{\hat{\ell}}$, of the studentized range Q_k satisified $P\{Q_k > b_{\hat{\ell}}\} \geq P_k$ and $P\{Q_k > b_{\hat{\ell}+1}\} < P_k$. Then we evaluated

$$P\{Q_k > b_\ell\} + \frac{1}{1000} \sum_{i=1}^{1000} w(\underset{\sim}{y}(i)) P\{\min_{2\leq j\leq k-1} \left[\frac{C_j}{T_j(\hat{\underset{\sim}{y}}(i))}\right] < Q_k \leq b_\ell\}$$

with ℓ starting at $\hat{\ell}-4$ until we had enough points (at least 8) to perform reasonably accurate inverse interpolation (Harter (1959), page 673, using direct interpolation tolerance of 5×10^{-5}) for C_k. This computation was actually performed on batches of 200 samples in order to obtain an estimate of the standard error of C_k.

Direct interpolation of the studentized range using the Aitken-Lagrange method with provision for up to 8-point interpolation with a tolerance of 5×10^{-5} was only required when

$$\min_{2\leq j\leq k-1} \frac{C_j}{T_j(\underset{\sim}{y}(i))} < b_\ell$$

and only needed to be performed once and saved as b_ℓ increased.

Every effort was made to ensure that errors associated with direct and inverse interpolation would be small relative to the sampling error. Therefore we feel the standard errors listed in Appendix C are a reasonable measure of the accuracy of the C_k. The standard errors were monotone decreasing with the largest errors occurring for ν=5. It took five minutes of 370/165 CPU time to produce tables for t=2(1)10 and ν=5(1)20, 24, 30, 40, 60, 120, ∞.

Where possible we compared our results with the conditional Monte Carlo approach to this problem developed by Arnold et. al. (1956) and Welsch (1965). The results were in reasonable agreement.

The tables for WNKA required no Monte Carlo. The inverse interpolation method was the same as that described for C_2 above.

APPENDIX B

B. Analysis of Tables

It is clear that GAPA is a more powerful test than GAPB for the situations we have examined. We included GAPB because we thought the tables of critical numbers might be "smoother". Our goal was a table that could be well represented by a

row term + column term + common term.

All of the analyses in this section are done for 20 degrees of freedom.

Figure 1 shows the results of computing the means of differences between rows and between columns of the GAPB table and using $\sqrt{2}$ times the t-statistic for comparing two samples as a common term. We see a rather pleasing fit. For other degrees of freedom the row and column terms are different but the quality of fit is about the same, decaying somewhat when the degrees of freedom are fewer than ten. An analysis like this could be used to compress the tables to 17 numbers instead of 45.

[Figure 1 about here.]

Now we try this analysis on the critical numbers of GAPA, noticing that when $t \geq 5$ the entries for $k=t-1$ are equal to those for $k=t-2$. This occurred because we enforced the monotoninity of the critical numbers as required for Theorem 1. Figure 2 shows this analysis. We see a less desirable fit but since most M C tests probably deal with five or more means, this may not be a serious problem.

[Figure 2 about here.]

Finally we ask how the table for GAPA compares with the one for WNKA. Figure 3 shows GAPA-WNKA and we see a systematic property in the residuals. A deeper analysis (not yet undertaken) may lead us to ways to find the critical numbers for GAP tests with a minimum of computation.

Figure 1

Two-way analysis of GAPB table for 20 degrees of freedom

Column Term

Row Term	1.09	1.02	.94	.85	.75	.63	.48	.28	0
0	-.02	-.01	-.01						
.39	-.01	-.01							
.55									
.67						-.01			
.76					-.01				
.83			-.01	-.01					
.88									
.93									
.97									

Common term = 2.95

Table entry = GAPB - row term - column term - common term

Figure 2

Two-way analysis of GAPA table for 20 degrees of freedom

Column Term

Row Term	1.09	1.02	.94	.85	.75	.63	.48	0	0
0	-.02	-.01	-.01						
.39	-.01	-.01					-.24	.53	
.55						*	.08		
.67					*	.04			
.76				*	.01				
.83			*	.01					
.88		*	-.01						
.93	*								
.97									

Common term = 2.95

*means table entry equal to cell immediately above.

Table entry = GAPA - row term - column term - common term
except for * entries.

Figure 3

GAPA-WNKA for 20 degrees of freedom

K-Gap	Number of Means								
	10	9	8	7	6	5	4	3	2
2									
3	.04	.04	.04	.04	.04	.04		.29	
4	.02	.02	.02	.01	.02	.01	.10		
5	.01	.01	.01	.01	.02	.06			
6	.01	.01	.01	.01	.02				
7	.01	.01	.01	.02					
8		.01	.01						
9									
10									

APPENDIX C

C. Tables of Critical Numbers

We include here a selection of tables for GAPA, GAPB and WNKA. All of these tables are for t=2(1)10 and α=.05. Table 5 lists standard errors for GAPA with ν=5 and ∞. The errors are approximately monotone between these two points. Table 6 lists GAPA critical values for ν=5, 10, 20, 40, 120, ∞ . Table 7 lists GAPB for ν=5, 20, 40, ∞ and Table 8 does the same for WNKA. Linear harmonic ν-wise interpolation is recommended.

To use the tables find the column with the total number of means to be compared, say t=5. For 20 degrees of freedom the GAPA critical numbers would be 3.58, 3.97, 3.97, 4.29 for gaps, 3-stretches, 4-stretches, and the range respectively. For WNKA the numbers are 4.23, 3.96, 3.93, 3.58 for the range, 4-stretch, 3-stretch, and gaps.

Tables for ν=5(1)20, 24, 30, 40, 60, 120, ∞ for all tests are available from the author.

TABLE 5

GAPA Error ($x10^3$)

DF = 5

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2									
3	5	5	5	5	5	5		80	
4	2	2	2	2	2		16		
5	1	1	1	1		2			
6			1		1				
7				1					
8									
9									
10									

DF = ∞

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2									
3	1	2	2	2	2	2		17	
4		1	1	1	1		4		
5				1		2			
6					2				
7				1					
8									
9		1							
10									

TABLE 6

GAPA Critical Numbers

Df = 5

r-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	5.70	5.55	5.39	5.21	5.00	4.76	4.47	3.64	3.64
3	6.41	6.24	6.06	5.86	5.64	5.37	4.60	5.38	
4	6.64	6.47	6.28	6.07	5.84	5.37	5.46		
5	6.79	6.61	6.42	6.21	5.84	5.72			
6	6.88	6.70	6.50	6.21	6.05				
7	6.94	6.75	6.56	6.34					
8	6.97	6.75	6.59						
9	6.97	6.80							
10	7.00								

Df = 11

r-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	4.48	4.39	4.30	4.19	4.06	3.91	3.73	3.15	3.15
3	4.97	4.88	4.77	4.66	4.53	4.37	3.88	4.27	
4	5.16	5.07	4.96	4.84	4.71	4.37	4.47		
5	5.30	5.20	5.09	4.97	4.71	4.71			
6	5.39	5.30	5.18	4.97	4.94				
7	5.47	5.37	5.18	5.14					
8	5.52	5.37	5.32						
9	5.52	5.47							
10	5.60								

TABLE 6 (con't.)

DF = 20

r-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	4.02	3.96	3.93	3.84	3.78	3.68	3.43	2.95	2.95
3	4.42	4.35	4.28	4.19	4.09	3.97	3.68	3.47	
4	4.59	4.52	4.44	4.33	4.25	3.97	4.06		
5	4.71	4.64	4.56	4.47	4.25	4.29			
6	4.80	4.73	4.65	4.47	4.47				
7	4.87	4.80	4.65	4.64					
8	4.92	4.80	4.78						
9	4.92	4.90							
10	5.01								

DF = 40

r-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	3.82	3.77	3.73	3.682	3.63	3.43	3.29	2.86	2.86
3	4.10	4.12	4.05	3.98	3.89	3.79	3.49	3.34	
4	4.34	4.28	4.21	4.14	4.05	3.79	3.83		
5	4.45	4.39	4.32	4.24	4.35	4.09			
6	4.53	4.47	4.40	4.24	4.20				
7	4.60	4.54	4.40	4.43					
8	4.65	4.54	4.53						
9	4.65	4.64							
10	4.74								

TABLE 6 (con't.)

DF = 12

r	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	3.70	3.65	3.59	3.52	3.43	3.33	3.21	2.80	2.80
3	4.03	3.98	3.92	3.85	3.77	3.68	3.35	3.56	
4	4.14	4.13	4.07	4.00	3.92	3.69	3.76		
5	4.28	4.23	4.17	4.10	3.92	3.96			
6	4.36	4.31	4.25	4.11	4.12				
7	4.43	4.37	4.25	4.26					
8	4.48	4.37	4.37						
9	4.48	4.48							
10	4.57								

DF = 13

r	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	3.64	3.59	3.53	3.45	3.39	3.29	3.17	2.77	2.77
3	3.98	3.91	3.85	3.79	3.71	3.62	3.32	3.52	
4	4.10	4.05	4.00	3.93	3.86	3.64	3.71		
5	4.20	4.15	4.10	4.03	3.86	3.90			
6	4.28	4.23	4.18	4.03	4.06				
7	4.34	4.29	4.18	4.19					
8	4.39	4.29	4.30						
9	4.39	4.39							
10	4.48								

TABLE 7

GAPB Critical Numbers

DF= 5

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	5.70	5.55	5.39	5.21	5.00	4.76	4.47	4.12	3.64
3	6.41	6.24	6.06	5.86	5.64	5.37	5.06	4.67	
4	6.64	6.47	6.28	6.07	5.84	5.57	5.24		
5	6.79	6.61	6.42	6.21	5.98	5.68			
6	6.88	6.70	6.50	6.29	6.04				
7	6.94	6.75	6.56	6.33					
8	6.97	6.78	6.59						
9	6.99	6.80							
10	7.00								

DF= 20

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	4.02	3.96	3.88	3.80	3.70	3.58	3.43	3.23	2.95
3	4.42	4.33	4.23	4.10	4.00	3.97	3.82	3.62	
4	4.59	4.52	4.44	4.35	4.25	4.13	3.98		
5	4.71	4.64	4.56	4.47	4.37	4.24			
6	4.80	4.73	4.65	4.56	4.45				
7	4.87	4.80	4.71	4.62					
8	4.92	4.85	4.77						
9	4.97	4.90							
10	5.01								

TABLE 7 (con't.)

DF= 40

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	3.82	3.77	3.70	3.62	3.53	3.43	3.29	3.12	2.86
3	4.18	4.12	4.05	3.98	3.87	3.79	3.66	3.48	
4	4.34	4.28	4.21	4.14	4.05	3.94	3.81		
5	4.43	4.39	4.32	4.24	4.15	4.05			
6	4.53	4.47	4.43	4.33	4.24				
7	4.60	4.54	4.47	4.39					
8	4.65	4.59	4.52						
9	4.70	4.64							
10	4.74								

DF= ∞

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
2	3.64	3.59	3.53	3.46	3.39	3.29	3.17	3.01	2.77
3	3.96	3.91	3.85	3.79	3.71	3.62	3.51	3.35	
4	4.10	4.05	4.00	3.93	3.86	3.76	3.65		
5	4.20	4.15	4.10	4.03	3.96	3.87			
6	4.28	4.23	4.18	4.11	4.03				
7	4.34	4.29	4.24	4.17					
8	4.39	4.35	4.29						
9	4.44	4.39							
10	4.48								

TABLE 8

WNKA Critical Numbers

DF= 5

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
10	6.99								
9	6.97	6.80							
8	6.97	6.75	6.58						
7	6.93	6.75	6.50	6.33					
6	6.87	6.69	6.50	6.19	6.03				
5	6.78	6.60	6.41	6.19	5.81	5.67			
4	6.61	6.44	6.26	6.05	5.81	5.30	5.22		
3	6.32	6.16	5.98	5.78	5.56	5.30	4.60	4.60	
2	5.70	5.55	5.39	5.21	5.00	4.76	4.47	3.64	3.04

DF= 20

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
10	5.01								
9	4.92	4.90							
8	4.92	4.79	4.77						
7	4.86	4.79	4.64	4.62					
6	4.79	4.72	4.64	4.46	4.45				
5	4.70	4.63	4.55	4.46	4.23	4.23			
4	4.57	4.50	4.43	4.34	4.23	3.96	3.96		
3	4.38	4.31	4.24	4.15	4.05	3.93	3.58	3.58	
2	4.02	3.96	3.88	3.80	3.70	3.58	3.43	2.95	2.95

TABLE 8 (con't.)

DF= 40

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
10	4.73								
9	4.65	4.63							
8	4.65	4.53	4.52						
7	4.59	4.53	4.40	4.39					
6	4.53	4.47	4.40	4.24	4.23				
5	4.44	4.38	4.31	4.24	4.04	4.04			
4	4.32	4.26	4.20	4.12	4.03	3.79	3.79		
3	4.15	4.09	4.02	3.95	3.86	3.75	3.44	3.44	
2	3.82	3.77	3.70	3.62	3.53	3.43	3.29	2.86	2.86

DF= ∞

K-GAP	NUMBER OF MEANS								
	10	9	8	7	6	5	4	3	2
10	4.47								
9	4.39	4.39							
8	4.39	4.29	4.29						
7	4.34	4.29	4.17	4.17					
6	4.28	4.23	4.17	4.03	4.03				
5	4.20	4.15	4.09	4.03	3.86	3.86			
4	4.09	4.04	3.98	3.92	3.84	3.63	3.63		
3	3.93	3.88	3.83	3.76	3.68	3.59	3.31	3.31	
2	3.64	3.59	3.53	3.46	3.39	3.29	3.17	2.77	2.77