Tables for a Treatments Versus Control Multiple Comparisons Sign Test*

A. L. RHYNE, JR.
The Bowman Gray School of Medicine, Winston-Salem, N. C.
R. G. D. STEEL
The University of North Carolina at Raleigh, N. C.

Tables are presented for a multiple comparisons sign test comparing each treatment with a control. An illustration of the test procedure is provided, and the tables are also used to construct non-parametric joint confidence intervals. It is also shown how to find the per-comparison error rate for a given experimentwise error rate and vice versa.

1. THE SIGN TEST

Since a multiple comparisons sign test is to be proposed, let us first look carefully at the sign test for paired observations.

The sign test is a simple, distribution-free, test criterion used to test hypotheses about the medians. It is used when experimental units are meaningfully paired, i.e. paired to be more alike in expected response, apart from treatment differences, than units in different pairs. Thus pairing might result from the application of criteria such as lot number, source of material, etc.

The most common hypothesis tested is that the medians of the distributions of signed differences are zero, but modifications provide tests concerning medians other than zero. For example, if (X, Y) represents a pair of observations, then quantities such as $X - (Y + \delta)$ or $X - \alpha Y$ may be examined.

The assumptions are that differences be independent and arise from continuous distributions having a common median. Each difference may arise from a different distribution, except so far as location of the median is concerned. The underlying assumptions are seen to be minimal.

The obvious alternative, when appropriate, is the t-test. However if differences are observed under quite variable conditions, they will likely have different variances and bring the validity of the t-test under question.

Situations for which the sign test is to be recommended are (1) when there are doubts as to the validity of the assumptions underlying the t-statistic, (2) for simplicity and economy where measurement is difficult, unreliable, or expensive but ranking is feasible, (3) when the number of observations is small,

* This paper is a condensation of part of the doctoral dissertation of the senior author at the University of North Carolina at Raleigh [1964].

Received April 1964; final revision Mar. 1965.

(4) when many observations are easily obtained and there is no problem of inability to detect a reasonable difference, and (5) for computational convenience.

Situations (1) and (2) include the summary of data where heterogeneity occurs as the result of having different years, places, investigators, techniques and experimental unit sizes. Here, the assumptions for a valid t-test are patently false and, in any case, the primary concern is to establish the presence or absence of a difference, reasonably consistent in direction. Situation (2) also includes rankings, often feasible when scaling is impossible.

For small samples, the sign test is quite efficient relative to parametric techniques, if such are valid. However, the power of the sign test decreases with increasing sample size. With large samples, often easily obtained if it is sufficient to rank the individuals of a pair, the power of the test may not be an issue.

Finally, the test is performed by counting the number of times the response to one treatment exceeds the other and comparing this with a tabulated critical value. Clearly, gross recording errors are not likely to have an appreciable effect on any conclusions.

The sign test is so simple, and valid under such general conditions, that it is appropriate to consider extensions to multiple comparisons.

2. Extending the Sign Test

The proposed extension of the sign test calls for a single observation on each of a control and two or more treatments in each block of a block experiment, an extension of pairing also called the two-way classification. The procedure is to conduct sign tests comparing each treatment with the control, using a critical region to be discussed. We first consider the problem in a general way.

For an analysis of variance of a block design, the validity of the assumptions of normality, homogeneous variance, and additivity of block and treatment effects is generally considered necessary. In practice, strict adherence is not usually demanded.

Normality appears unnecessary when additivity holds, when randomization is used and when only tests of the significance of treatment effects are wanted; for estimation, within-blocks homogeneity of variance is desirable. If the normality assumption is valid but variances are treatment related, tests of hypotheses concerning linear contrasts require a partitioning of the residual sum of squares to provide a valid error term.

When it is known that additivity of treatment and block effects is not a valid assumption, it is customary to look for a transformation to a scale on which additivity holds or to eliminate removable non-additivity. An appropriate transformation may be difficult to find and the resulting data difficult to interpret. Techniques to define and handle removable non-additivity are still not numerous.

Non-additivity of treatment and block effects is also a failure of treatment differences to be homogeneous over blocks, that is, a treatment x block inter-

action. If in the presence of non-removable interaction, it is still desired to test treatment effects, then only in the case of the fixed model is there no valid error term. This of course, says nothing about the analysis of data clearly not adhering to any standard model.

This extension of the sign test is valid when the analysis of variance is, and when non-additivity of a reasonable nature exists. Treatment-related heterogeneity of variance may be a problem since all permutations of the observations within a block are considered equally likely under the null hypothesis. With this possible exception, the proposed extension seems appropriate when the two-sample sign test is.

The experimenter dealing with more than two treatments must decide whether the comparison or the experiment is the conceptual unit in defining error rate.

If the comparison is the unit, then the error rate is the long run percentage of false significant differences. Comparisons are determined prior to the experiment and are not to be dependent on the outcome; they need not be uncorrelated. The error rate is a comparisonwise or per-comparison one. The sign test with customarily tabulated critical values is a per-comparison test criterion with this error rate.

If the experiment is the unit, then the error rate is the long run proportion of experiments with at least one false significant difference. Tests of this sort are of two types: (1) suitcase tests that give evidence for or against the presence of real differences but no indication as to which are real, and (2) multiple comparisons tests that look at individual comparisons, from a specific set on through those suggested by the data. This error rate is an experimentwise one.

If comparisons are to be made because the data suggest them, then an experimentwise error rate is appropriate. Otherwise, defining error rate is the responsibility of the investigator.

The widespread use of confidence coefficients of .05 and .01, regardless of any considerations of the error rate definition and associated consequences, has possibly led to the notion that the two error rates are irreconcilable. This is not the case. In fact, if we choose to make k orthogonal comparisons when, as in this case, it is not necessary to estimate a single variance as the basis for all comparisons, then a per-comparison error rate, α_C , and an experimentwise error rate, α_E , are related by the equation $(1 - \alpha_E) = (1 - \alpha_C)^k$; if we set one α, the other is determined. In general, when multiple comparisons are not orthogonal or it is necessary to estimate and use a single variance*, we can only say that $\alpha_E \leq 1 - (1 - \alpha_C)^k$. However with adequate tables, we can find the comparisonwise error rate for a fixed per-comparison error rate, and vice versa. The two error rates are uniquely related and if the experimenter can specify the kind and amount of protection he wants, then both are automatically established in value.

Our test has an experimentwise error rate.

* A referee has called our attention to the fact that when the variances of comparisons are unknown and their estimates depend upon a common σ^2, then probability statements concerning orthogonal comparisons are correlated and the correct relationship is $(1 - \alpha_E) \geq (1 - \alpha_C)k$.

3. Treatments versus Control: An Example

Let X_{0j} and X_{ij}, $i = 1, \cdots, k$ and $j = 1, \cdots, n$ be measured responses on the control and i-th treatment in the j-th block of a randomized block experiment. Consider the k sets of n signed differences $d_{ij} = X_{ij} - X_{0j}$. Let r_i be the number of the n d_{ij}'s with minus signs (or plus signs).

We plan k sign tests with a critical value depending on the distribution of $(r_1, \cdots, r_k)$. In particular, critical values are from the distribution of the minimum r_i, Table 1, if the test is against one-sided alternatives, or of $\min_i (\min(r_i, n - r_i))$, Table 2, if alternatives are two sided.

The data, Table 3, on which this multiple comparisons sign test is used are a small part of the results of an experiment designed to compare three methods of measuring serum cholesterol with the method standard at the time of the experiment. Such data have not been studied extensively and may well include at least two populations as determined by sick, and well individuals. In general, technicians will change from time to time and other sources of variation are likely. In consequence, we can say relatively little, with assurance, about the underlying distribution of the observations so that a non-parametric technique is important.

In our case, it is informative to know how each analytical method compares with the standard. This suggests the advisibility of an experimentwise error rate. No comparisons are suggested by the nature of the treatments; if it seems reasonable to compare the Control with the most different treatment, then there is necessity for an experimentwise error rate. Information on how non-standards compare seems unimportant at this time.

Data on all methods for an individual are based on a common blood sample with all methods being used nearly simultaneously. Hence blocking is obvious and large differences from block to block are probable. The conditions required for the sign test are met whereas those for a valid analysis of variance are doubtful.

Discussion up to this point has probably implied a main interest in tests of significance, but confidence intervals on the sizes of differences are of interest. They are also constructed below.

For the 5% two-sided test of the null hypothesis of zero medians against the alternative that the three-tuples (d_{1i}, d_{2i}, d_{3i}) have probability distributions with common medians in which at least one component is different from zero, Table 2 gives a critical value of $\min_i (\min(r_i, n - r_i))$ of 0. Here $(r_1, r_2, r_3) = (4, 8, 0)$ and Method 3 is declared significantly different from the Control.

To obtain joint confidence limits of the differences, note that for a 5% two-tailed test, the critical value is zero. Consider the smallest difference between the Control and Method 1; it is -20. If less than -20 is subtracted from each difference, e.g. add 21 to each, the test would deny the null hypothesis that the median of the d_{1i}'s is -21. Also, if we subtract more than 40 we obtain significance in the opposite direction, i.e. a median of over 40 for the d_{1i}'s is an untenable hypothesis. Hypotheses that the median of the d_{1i}'s is -20, $+40$, or anything in between are not denied. Continuing, we find that the

TABLE 1

Critical values of minimum r_i for comparison of k treatments against one control in n sets of observations: a one-tailed critical region with an experiment-wise error rate

n	Level of significance for min r_i	k = number of treatments (excluding control)							
		2	3	4	5	6	7	8	9
4	.15	0 (.113)[a]	–	–	–	–	–	–	–
	.10	–	–	–	–	–	–	–	–
	.05	–	–	–	–	–	–	–	–
5	.15	0 (.058)	0 (.082)	0 (.104)	0	–	–	–	–
	.10	0 (.058)	0 (.082)	–	–	–	–	–	–
	.05	–	–	–	–	–	–	–	–
6	.15	0 (.030)	0 (.043)	0 (.055)	0	0	0	0	0
	.10	0 (.030)	0 (.043)	0 (.055)	0	0	–	–	–
	.05	0 (.030)	0 (.043)	–	–	–	–	–	–
7	.15	1 (.113)	0 (.022)	0 (.029)	0	0	0	0	0
	.10	0 (.015)	0 (.022)	0 (.029)	0	0	0	0	0
	.05	0 (.015)	0 (.022)	0 (.029)	0	–	–	–	–
8	.15	1 (.066)	1 (.092)	1	0	0	0	0	0
	.10	1 (.066)	1 (.092)	0	0	0	0	0	0
	.05	0 (.008)	0 (.011)	0	0	0	0	0	0
9	.15	1 (.037)	1 (.053)	1	1	1	1	1	1
	.10	1 (.037)	1 (.053)	1	1	0	0	0	0
	.05	1 (.037)	0 (.006)	0	0	0	0	0	0
10	.15	2 (.100)	2 (.139)	1	1	1	1	1	1
	.10	1 (.021)	1 (.030)	1	1	1	1	1	1
	.05	1 (.021)	1 (.030)	1	0	0	0	0	0
11	.15	2 (.061)	2 (.087)	2	2	1	1	1	1
	.10	2 (.061)	2 (.087)	1	1	1	1	1	1
	.05	1 (.011)	1 (.017)	1	1	1	1	0	0
12	.15	3 (.131)	2 (.053)	2	2	2	2	2	2
	.10	2 (.037)	2 (.053)	2	2	1	1	1	1
	.05	2 (.037)	1 (.009)	1	1	1	1	1	1
13	.15	3 (.085)	3 (.119)	2	2	2	2	2	2
	.10	3 (.085)	2 (.031)	2	2	2	2	2	2
	.05	2 (.022)	2 (.031)	2	1	1	1	1	1
14	.15	3 (.054)	3 (.077)	3	3	3	2	2	2
	.10	3 (.054)	3 (.077)	2	2	2	2	2	2
	.05	2 (.013)	2 (.018)	2	2	2	2	1	1
15	.15	4 (.108)	4 (.149)	3	3	3	3	3	3
	.10	3 (.034)	3 (.048)	3	3	3	2	2	2
	.05	3 (.034)	3 (.034)	2	2	2	2	2	2
16	.15	4 (.072)	4	4	3	3	3	3	3
	.10	4 (.072)	3	3	3	3	3	3	3
	.05	3 (.021)	3	3	3	2	2	2	2

[a] ( ) Exact cumulative probability.

TABLE 1 (continued)

Critical values of minimum r_i for comparison of k treatments against one control in n sets of observations: a one-tailed critical region with an experiment-wise error rate

n	Level of significance for min r_i	k = number of treatments (excluding control)							
		2	3	4	5	6	7	8	9
17	.15	5 (.129)	4	4	4	4	4	3	3
	.10	4 (.046)	4	4	3	3	3	3	3
	.05	4 (.046)	3	3	3	3	3	2	2
18	.15	5 (.089)	5	4	4	4	4	4	4
	.10	5 (.089)	4	4	4	4	4	3	3
	.05	4 (.029)	4	3	3	3	3	3	3
19	.15	6 (.149)	5	5	5	4	4	4	4
	.10	5 (.060)	5	4	4	4	4	4	4
	.05	4 (.019)	4	4	4	3	3	3	3
20	.15	6 (.105)	6	5	5	5	5	5	5
	.10	5 (.039)	5	5	5	4	4	4	4
	.05	5 (.039)	4	4	4	4	4	3	3
21	.15	6 (.073)	6	6	5	5	5	5	5
	.10	6 (.073)	5	5	5	5	5	5	5
	.05	5 (.026)	5	5	4	4	4	4	4
22	.15	7 (.121)	6	6	6	6	6	5	5
	.10	6 (.050)	6	6	5	5	5	5	5
	.05	6 (.050)	5	5	5	4	4	4	4
23	.15	7 (.086)	7	6	6	6	6	6	6
	.10	7 (.086)	6	6	6	6	5	5	5
	.05	6 (.033)	6	5	5	5	5	5	5
24	.15	8 (.136)	7	7	7	6	6	6	6
	.10	7 (.060)	7	6	6	6	6	6	6
	.05	6 (.022)	6	6	5	5	5	5	5
25	.15	8	8	7	7	7	7	7	7
	.10	7	7	7	7	6	6	6	6
	.05	7	6	6	6	6	6	5	5
30	.15	10	10	9	9	9	9	9	9
	.10	10	9	9	9	8	8	8	8
	.05	9	8	8	8	8	8	7	7
35	.15	12	12	12	11	11	11	11	11
	.10	12	11	11	11	10	10	10	10
	.05	11	10	10	10	10	9	9	9
40	.15	15	14	14	13	13	13	13	13
	.10	14	13	13	13	13	12	12	12
	.05	13	12	12	12	12	11	11	11
45	.15	17	16	16	16	15	15	15	15
	.10	16	16	15	15	15	14	14	14
	.05	15	14	14	14	14	13	13	13
50	.15	19	18	18	18	17	17	17	17
	.10	18	18	17	17	17	17	16	16
	.05	17	17	16	16	16	16	15	15

TABLE 2

Critical values of minimum$_i$ ($\min(r_i, n - r_i)$) for comparison of k treatments against one control in n sets of observations: a two-tailed critical region with an experiment-wise error rate

n	Level of significance for $\min_i(\min(r_i, n - r_i))$	k = number of treatments (excluding control) 2	3	4	5	6	7	8	9
6	.10	0 (.060)[a]	0 (.085)	–	–	–	–	–	–
	.05	–	–	–	–	–	–	–	–
	.01	–	–	–	–	–	–	–	–
7	.10	0 (.030)	0 (.044)	0 (.057)	–	–	–	–	–
	.05	0 (.030)	–	–	–	–	–	–	–
	.01	–	–	–	–	–	–	–	–
8	.10	0 (.015)	0 (.023)	0	0	0	0	0	0
	.05	0 (.015)	0 (.023)	0	–	–	–	–	–
	.01	–	–	–	–	–	–	–	–
9	.10	1 (.074)	0 (.011)	0	0	0	0	0	0
	.05	0 (.008)	0 (.011)	0	0	0	0	–	–
	.01	–	–	–	–	–	–	–	–
10	.10	1 (.041)	1 (.060)	1	0	0	0	0	0
	.05	1 (.041)	0 (.006)	0	0	0	0	0	0
	.01	0 (.004)	0 (.006)	–	–	–	–	–	–
11	.10	1 (.023)	1 (.034)	1	1	1	1	0	0
	.05	1 (.023)	1 (.034)	0	0	0	0	0	0
	.01	0 (.002)	0 (.003)	0	–	–	–	–	–
12	.10	2 (.073)	1 (.018)	1	1	1	1	1	1
	.05	1 (.012)	1 (.018)	1	1	0	0	0	0
	.01	0 (.001)	0 (.001)	0	0	0	0	–	–
13	.10	2 (.043)	2 (.063)	2	1	1	1	1	1
	.05	2 (.043)	1 (.011)	1	1	1	1	1	1
	.01	1 (.007)	0 (.001)	0	0	0	0	0	0
14	.10	2 (.025)	2 (.037)	2	2	2	2	1	1
	.05	2 (.025)	2 (.037)	1	1	1	1	1	1
	.01	1 (.004)	1 (.005)	0	0	0	0	0	0
15	.10	3 (.067)	3 (.096)	2	2	2	2	2	2
	.05	2 (.014)	2 (.021)	2	2	1	1	1	1
	.01	1 (.002)	1 (.003)	1	1	0	0	0	0
16	.10	3 (.041)	3	3	2	2	2	2	2
	.05	3 (.041)	2	2	2	2	2	2	2
	.01	2 (.008)	1	1	1	1	1	1	0
17	.10	4 (.093)	3	3	3	3	3	2	2
	.05	3 (.024)	3	2	2	2	2	2	2
	.01	2 (.005)	1	1	1	1	1	1	1

[a] ( ) Exact cumulative probability

TABLE 2 (continued)

Critical values of minimum$_i$(min(r_i, $n - r_i$)) for comparison of k treatments against one control in n sets of observations: a two-tailed critical region with an experiment-wise error rate

n	Level of significance for min$_i$(min(r_i, $n - r_i$))	k = number of treatments (excluding control)							
		2	3	4	5	6	7	8	9
18	.10	4 (.059)	4	3	3	3	3	3	3
	.05	3 (.015)	3	3	3	2	2	2	2
	.01	2 (.003)	2	2	1	1	1	1	1
19	.10	4 (.037)	4	4	4	3	3	3	3
	.05	4 (.037)	3	3	3	3	3	3	3
	.01	3 (.009)	2	2	2	2	2	2	1
20	.10	5 (.079)	4	4	4	4	4	4	3
	.05	4 (.023)	4	3	3	3	3	3	3
	.01	3 (.005)	2	2	2	2	2	2	2
21	.10	5 (.051)	5	4	4	4	4	4	4
	.05	4 (.014)	4	4	4	4	3	3	3
	.01	3 (.003)	3	3	2	2	2	2	2
22	.10	6 (.099)	5	5	5	5	4	4	4
	.05	5 (.033)	4	4	4	4	4	4	4
	.01	4 (.009)	3	3	3	3	3	2	2
23	.10	6 (.066)	5	5	5	5	5	5	5
	.05	5 (.021)	5	5	4	4	4	4	4
	.01	4 (.005)	3	3	3	3	3	3	3
24	.10	6 (.043)	6	6	5	5	5	5	5
	.05	6 (.043)	5	5	5	5	5	4	4
	.01	4 (.003)	4	4	3	3	3	3	3
25	.10	7	6	6	6	6	6	5	5
	.05	6	6	5	5	5	5	5	5
	.01	4	4	4	4	4	4	3	3
30	.10	9	8	8	8	8	7	7	7
	.05	8	8	7	7	7	7	7	7
	.01	6	6	6	6	5	5	5	5
35	.10	11	10	10	10	10	9	9	9
	.05	10	9	9	9	9	9	9	8
	.01	8	8	8	7	7	7	7	7
40	.10	13	12	12	12	12	11	11	11
	.05	12	12	11	11	11	11	11	10
	.01	10	10	9	9	9	9	9	9
45	.10	15	14	14	14	14	13	13	13
	.05	14	14	13	13	13	13	12	12
	.01	12	12	11	11	11	11	11	11
50	.10	17	17	16	16	16	16	15	15
	.05	16	16	15	15	15	15	14	14
	.01	14	14	13	13	13	13	13	12

TABLE 3
Total cholesterol data

Patient No.	Control	Method 1	Method 2	Method 3
1	260	240 (−20)[a]	270 (+10)	200 (−60)
2	300	290 (−10)	290 (−10)	240 (−60)
3	290	320 (+30)	320 (+30)	240 (−50)
4	250	240 (−10)	270 (+20)	210 (−40)
5	270	250 (−20)	260 (−10)	190 (−80)
6	180	220 (+40)	230 (+50)	160 (−20)
7	200	190 (−10)	210 (+10)	140 (−60)
8	220	230 (+10)	250 (+30)	180 (−40)
9	410	420 (+10)	430 (+20)	270 (−140)
10	310	300 (−10)	320 (+10)	200 (−110)
Number of Minuses, r_i		6	2	10
Number of Pluses, $(n - r_i)$		4	8	0

[a] Differences $d_{ij} = X_{ij} - X_{0j}$ are given in parentheses.

95% joint confidence limits on the medians are as in the equation below.

$$\begin{bmatrix} -20 \\ -10 \\ -140 \end{bmatrix} \leq \begin{bmatrix} m_1 - m_0 \\ m_2 - m_0 \\ m_3 - m_0 \end{bmatrix} \leq \begin{bmatrix} 40 \\ 50 \\ -20 \end{bmatrix}.$$

If the critical value is unity, as for a joint confidence coefficient of $\alpha = .10$, then begin with the second smallest difference. Here the second smallest d_{1j} is -10. If we subtract -11 from each d_{1j}, we obtain two negative differences so cannot reject the hypothesis that the median is -11. This will still be the case for -19. This situation has occurred because two values were extreme, namely -20.

The same problem does not occur at the other end of the interval. Any difference greater than $+30$ is rejected because it leads to one or zero positives, cause for rejection of the corresponding hypothesis concerning the median. Hence, the 90% joint confidence limits on the median are as in the following equation:

$$\begin{bmatrix} -20 \\ -10 \\ -110 \end{bmatrix} \leq \begin{bmatrix} m_1 - m_0 \\ m_2 - m_0 \\ m_3 - m_0 \end{bmatrix} \leq \begin{bmatrix} 30 \\ 30 \\ -40 \end{bmatrix}.$$

For $k = 3$, $n = 10$, and an experimentwise error rate of .05, the critical value is zero. In fact, the exact error rate is .006 as is seen from Table 2. If we now find the probability of zero pluses and zero minuses in n binomial

trials, each with $p = .5 = q$, we have the corresponding per-comparison error rate for a two-tailed test. Here, it is .002.

For an error rate of .10, the critical value is one. Here the exact error rate is .060. The probability of zero or one pluses and minuses is .021. This is the corresponding per-comparison error rate for a two-tailed test.

Finding experimentwise error rates for given per-comparison error rates is the inverse of the procedure above but may require more adequate tables than those given here.

4. An Alternative Test

Tables 1 and 2 may also be used as follows, a procedure calling for neither a per-comparison nor an experimentwise error rate.

Compare Control with the most different alternative. If this difference is not significant, no further tests are to be made since the evidence supports an hypothesis of homogeneous responses. If the difference is significant, assume the difference to be real and test the next most different alternative using a critical value calling for $k - 1$ non-control treatments. Proceed in this manner till a difference is declared non-significant.

For our data and a two-sided test with $\alpha = .05$, the critical value is zero and Method 3 is significantly different from Control. Now we use $k = 2$ and find a critical value of one. The remaining treatments and Control are not significantly different.

Friedman and Brown have both presented non-parametric tests for the two-way classification, but these are not specifically treatments versus control tests. Rather, they are analogues of the F-test used in the analysis of variance. Such tests can indicate the presence but not the location of real differences.

5. Some Problems

Ties between control and treated units in the same block are theoretically not acceptable for measured observations; in practice, they occur. Various recommendations for the treatment of ties have been made, including omitting them as well as assigning them equally to positive and negative differences. In terms of false significant differences, the latter is the conservative procedure.

Identical differences should not occur but do in practice. This leads to no real difficulty in using the tables but it is certainly not satisfying. For example, because of identical extreme differences in d_{1i} and d_{2i}, the 95% and 99% joint confidence intervals on the medians of these differences have the same lower value.

In the parametric case, equal numbers of observations on control and treated units are not optimum in terms of detecting differences. The optimum occurs when the ratio of number of observations on control to number on any one treatment is approximately equal to the square root of the number of treatments. Certainly, something is to be gained by investigating other than equal allocation.

What are possible alternatives which are also multiple comparisons procedures? Some extension of Wilcoxon's method of signed ranks could well be investigated.

The efficiency and power of the test against various alternatives need investigation.

6. Computation of Tables

The test criterion is based on the distribution of $(r_1 , \cdots , r_k)$ as defined in Section 3. To obtain $(r_1 , \cdots , r_k)$, record each of the k differences $(X_{1j} - X_{0j} , \cdots , X_{kj} - X_{0j})$ as one if the difference is negative and zero if positive. Each of the n blocks yields a vector whose k elements are zeros and ones. Add the n vectors to obtain $(r_1 , \cdots , r_k)$.

Each vector represents a single observation on a distribution for which we define median $(\cdots , d_{ij} , \cdots) = (\cdots , \text{median } d_{ij} , \cdots)$. There are 2^k possible vectors. The null hypothesis calls for the $(k + 1)!$ possible arrangements to be equally likely and the n distributions to have zero medians. The ratio of the product of the number of possible arrangements of the observations on each side of the control to the total number of arrangements is the probability the vector appears in a single trail.

Steel [4] denotes the set of vectors in a single trial by $v_1 , \cdots , v_s$ where $s=2^k$ and their corresponding probabilities by $p_1 , \cdots , p_s$. Then the probability of obtaining a given $(r_1 , \cdots , r_k)$ as the sum of the vectors in n trials is the sum of the coefficients of the products of powers of the x_i's in one or more terms of the expansion of (1).

$$(p_1x_1 + \cdots + p_sx_s)^n . \tag{1}$$

To find an appropriate term, (2) must be solved for the n_i's subject to the condition $\sum_{i=1}^{s} n_i = n$.

$$\sum_{i=1}^{s} n_iv_i = (r_1 , \cdots , r_k). \tag{2}$$

Each solution of (2) is a set of exponents of the x_i's in (1) and therefore determines a term.

For $k = 2$, the $s = 4$ vectors with their probabilities are given by (3).

$$\begin{aligned} v_1 &= (1, 1) \quad \text{with} \quad p_1 = 1/3 \\ v_2 &= (1, 0) \quad \text{with} \quad p_2 = 1/6 \\ v_3 &= (0, 1) \quad \text{with} \quad p_3 = 1/6 \\ v_4 &= (0, 0) \quad \text{with} \quad p_4 = 1/3. \end{aligned} \tag{3}$$

The probability that the joint event of v_1 occurring n_1 times, $\cdots$, v_4 occurring n_4 times in n trials is given by (4).

$$\frac{n!}{n_1!n_2!n_3!n_4!}(1/3)^{n_1+n_4}(1/6)^{n_2+n_3} \quad \text{with} \quad \sum n_i = n. \tag{4}$$

It follows that the probability that $(r_1 , r_2) = (a, b)$ is the sum of terms like (5)

$$\frac{n!}{\alpha!(a - \alpha)!(b - \alpha)!(n - a - b + \alpha)!}(1/3)^{n-a-b+2\alpha}(1/6)^{a+b-2\alpha} \tag{5}$$

where summation is over all values of α which make the terms in the denominator non-negative.

The distributions of min r_i and $\min_i$ $(\min(r_i, n - r_i))$ are obtained by solving (5) for every possible value of a and b and classifying accordingly. For values of $n = 1(1)24$, the distributions of min r_i and $\min_i$ $(\min(r_i, n - r_i))$ have been tabulated by means of an IBM 704 computer; the cumulative distributions were also obtained.

When k is greater than two, expressions similar to (5) can be obtained where summation is over more than two parameters. Simplification does not appear feasible and it seems better to expand the multinomial (1), evaluating every term and recording its value as the partial sum of the $(r_1, \cdots, r_k)$ to which it belongs. For example, when $k = 3$, the possible vectors are given by (6).

$$\begin{aligned} v_1 &= (1, 1, 1) & v_5 &= (0, 1, 1) \\ v_2 &= (1, 1, 0) & v_6 &= (0, 1, 0) \\ v_3 &= (1, 0, 1) & v_7 &= (0, 0, 1) \\ v_4 &= (1, 0, 0) & v_8 &= (0, 0, 0) \end{aligned} \tag{6}$$

$$\text{with} \quad p_1 = p_8 = \tfrac{1}{4} \quad \text{and} \quad p_2 = \cdots = p_7 = \tfrac{1}{12}.$$

The required coefficients of individual terms in the expansion of (1) are of the form (7).

$$\frac{n!}{n_1! \cdots n_8!} (1/4)^{n_1+n_8} (1/12)^{n-n_1-n_8} . \tag{7}$$

If $(n_1, \cdots, n_8) = (1, 1, 0, 0, 2, 2, 0, 0)$, the value of (7) is given by (8). Note that $\sum_{i=1}^{8} n_i = n = 6$.

$$\frac{6!}{1!1!2!2!} (1/4)(1/12)^5 = .0001808. \tag{8}$$

This is one of the terms giving rise to $(r_1, r_2, r_3) = (2, 6, 3)$ as may be seen by evaluating (9) using the above n_i .

$$\begin{aligned} r_1 &= n_1 + n_2 + n_3 + n_4 = 1 + 1 + 0 + 0 = 2 \\ r_2 &= n_1 + n_2 + n_5 + n_6 = 1 + 1 + 2 + 2 = 6 \\ r_3 &= n_1 + n_3 + n_5 + n_7 = 1 + 0 + 2 + 0 = 3. \end{aligned} \tag{9}$$

By this means, namely generating all probabilities, the distribution of $(r_1, \cdots, r_k)$ and, in turn, of cumulative min r_i and cumulative $\min_i$ (min $(r_i, n - r_i)$) have been obtained for $k = 3$, $n = 1(1)15$ and $k = 4$, $n = 1(1)7$. An IBM 1620 computer was used. As n and/or k increases, computer time soon becomes the major factor in restricting the tabulation of more extensive tables.

The equal correlations, namely $\rho = \frac{1}{3}$, between the various r_i's, and the fact that the individual r_i's are distributed binomially with $p = \frac{1}{2}$, suggest the multivariate normal distribution with all correlations equal to $\frac{1}{3}$ as an appropriate approximation.

Gupta [3] has tabled, among other distributions, the probability that $k=1(1)12$ standard normal variables with common correlation equal to $\frac{1}{3}$ are simultaneously less than or equal to H. To do this, he evaluates equation (10).

$$\Pr(\text{all } x\text{'s} < H \mid \rho) = \int_{-\infty}^{\infty} F^k\left(\frac{H + \rho^{\frac{1}{2}}y}{(1-\rho)^{\frac{1}{2}}}\right) f(y)\, dy \tag{10}$$

where $f(y)$ is the unit normal density function and

$$F(z) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z} \exp(-t^2/2)\, dt.$$

From relation (11) for $k = 2$,

$$\Pr(\min r_i < H \mid k = 2) = \Pr(\text{at least one } r_i < H \mid k = 2)$$

$$= \Pr(r_i < H \mid k = 1) + \Pr(r_2 < H \mid k = 1) - \Pr(r_1, r_2 < H \mid k = 2) \tag{11}$$

and similar relations for larger k, along with Gupta's tables, the probability that the minimum r_i is equal to or less than some given value may be derived. Clearly, these results will be useful in tabulating critical values for one-tailed tests. The minimum r_i are standardized, as in equation (12), in order to tabulate the approximation probabilities.

$$H = \frac{\text{minimum } r_i + 0.5 - np}{\sqrt{npq}} \quad \text{with} \quad p = q = 1/2. \tag{12}$$

Comparison of the exact cumulative distribution with its approximation reveals only small differences which quickly diminish as n increases. For probabilities of .05(.05).25, the critical values given by the exact distribution differ from those given by the approximation in only one case among those for $k = 2$, $n = 4(1)24$, for $k = 3$, $n = 4(1)15$, and for $k = 4$, $n = 4(1)7$; in particular, for $k = 3$, $n = 6$, the true critical value of zero is not given by the approximation. The exact probability for the value zero is .043.

Table 1 gives critical values of minumum r_i for all comparisons of k treatments against a single control in an experiment with n replications. Tabulation is for $\alpha = .05(.05).15$ and for $n = 4(1)25$ and $k = 2(1)9$. This table of critical values is constructed from the exact distributions, supplemented with values based on the multivariate normal approximation.

As for the two-tailed situation, a multivariate normal distribution with equal correlations, namely $\rho = \frac{1}{3}$, would probably produce a reasonable approximation to the exact distribution of $\min_i (\min(r_i, n - r_i))$. Unfortunately, the distribution of $\Pr(\max |x_i| \geq H)$ for equally correlated variables, $\rho = \frac{1}{3}$, in a multivariate normal distribution seems not to be tabulated. Stuart (5), David and Thigpen (1), among others, have considered this problem. David and Thigpen (1) examined the distribution of the extremes in a normal sample when the variables are equally correlated with common, known mean and common variance. Their report gives the upper 100α, $\alpha = .10, .05, .01$, percentage points of the distribution of the maximum of the modulus of $k = 2(1)9$ standard normal variables with equicorrelations, $\rho = 0, .2, .4, .8, 1$.

Steel [4] computed a table of critical values by taking the integral part of

the number computed by means of equation (13) with t from Dunnett's [2] tables. It was assumed that no value should be declared significant when the computation gave negative values.

$$r = (n - 1 - t\sqrt{n})/2 \tag{13}$$

where t is Dunnett's t with infinite degrees of freedom, and based on $\rho = .5$. Replacing the values of t in equation (13) by the corresponding values obtained by David (1) for $\rho = .2$ and $\rho = .4$, the two resulting tables of critical values, $k = 2(1)9$, differ only slightly from those obtained by Steel [4].

The critical value for each n, k, and α obtained using $\rho = .2$ was compared with that for $\rho = .4$. When these values differed, interpolation to $\rho = \frac{1}{3}$ was used to obtain the critical value. By supplementing the critical values obtained from the exact distributions with those from the approximation, Table 2 was formed. This table gives critical values for $\alpha = .10$, .05, and .01 for the test criterion $\min_i (\min(r_i, n - r_i))$. Tabulation is for $n = 6(1)25$ and $k = 2(1)9$.

A comparison shows that critical regions for $\alpha = .10$, .05 and .01 corresponding to available tabulated exact distributions differ only slightly from those obtained by the approximation. Discrepancies occurred for $k = 2$, $\alpha = .10$ at $n = 17$ and 22, for $k = 2$, $\alpha = .05$ at $n = 10$, 13, for $k = 2$, $\alpha = .01$ at $n = 9$, 13, 16, 19, for $k = 3$, $\alpha = .10$ and $n = 15$, for $k = 3$, $\alpha = .05$ and $n = 7$, and for $k = 3$, $\alpha = .01$ and $n = 10$.

In each discrepancy, the critical value of the approximation was smaller than that of the exact distribution by one integer. Although this approximation underestimates the critical region in some cases, it is felt that for practical purposes it is usable for all values of k.

References

1. David, H. A. and C. C. Thigpen, 1961. Distribution of extremes in normal samples when the variables are equally correlated with common, known mean and common variances. Technical Report No. 15, Dept. of Statistics. Virginia Polytechnic Institute, Blacksburg.
2. Dunnett, C. W., 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc. 51*, 1096–1121.
3. Gupta, S. S., 1963. Probability integrals of multivariate normal and multivariate t. *Annals of Math. Stat. 34*, 792–828.
4. Steel, R. G. D., 1959. A multiple comparison sign test: treatments versus control. *J. Amer. Stat. Assoc. 54*, 767–775.
5. Stuart, A., 1958. Equally correlated variates and the multinormal integral. *J. Roy. Stat. Soc. B, 20*, 373.