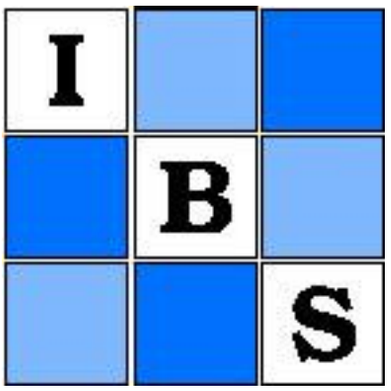

Comparing Individual Means in the Analysis of Variance

Author(s): John W. Tukey

Source: *Biometrics*, Jun., 1949, Vol. 5, No. 2 (Jun., 1949), pp. 99-114

Published by: International Biometric Society

Stable URL: https://www.jstor.org/stable/3001913

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/3001913?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

International Biometric Society is collaborating with JSTOR to digitize, preserve and extend access to *Biometrics*

JSTOR

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

COMPARING INDIVIDUAL MEANS IN THE ANALYSIS OF VARIANCE*

JOHN W. TUKEY

Princeton University

The practitioner of the analysis of variance often wants to draw as many conclusions as are reasonable about the relation of the true means for individual "treatments," and a statement by the F-test (or the z-test) that they are not all alike leaves him thoroughly unsatisfied. The problem of breaking up the treatment means into distinguishable groups has not been discussed at much length, the solutions given in the various textbooks differ and, what is more important, seem solely based on intuition.

After discussing the problem on a basis combining intuition with some hard, cold facts about the distributions of certain test quantities (or "statistics") a simple and definite procedure is proposed for dividing treatments into distinguishable groups, and for determining that the treatments within some of these groups are different, although there is not enough evidence to say "which is which." The procedure is illustrated on examples.

2. DISCUSSION OF THE PROBLEM

LET US BEGIN by considering how the latest and most advanced statistical theory would approach this problem and then explain why such a solution seems impractical. To make things more precise, let us suppose as a fictitious example that seven varieties of buckwheat; A, B, C, D, E, F, and G have been tested for yield in each of 12 locations, and that our interest is in the average yield of the buckwheat varieties in a region of which the 12 locations are a respectable sample, and in years exactly like the one in which the experiment was made. We will then have a simple and straightforward analysis of variance into varieties, locations, and interaction. We shall be concerned with the seven observed variety means and with an unbiased estimate of their variance, which will be given by 1/12th of the interaction mean square, which is itself on 66 degrees of freedom. What can we say about the varieties under these conditions?

We will wish to say, for example, that B and F yield better than A, C, and G, which yield better than D and E. Perhaps we might wish to add that A, C and G are not alike, although we do not know which one

*Prepared in connection with research sponsored by the Office of Naval Research.

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

yields better. The most modern approach would require us to proceed as follows: Write down all the possible conclusions to which we might come—the one illustrated above is one of the 120,904 similar possibilities for seven "treatments." Then for each combination of seven true mean yields we should decide how much it would "cost" us to make each of these 120,904 decisions. Making the usual assumptions about the distribution of fluctuations in yield, we would have begun to state a mathematically well-posed problem. We are unlikely to get this far in a practical problem in my lifetime! Then we find, to our horror, that there are many competing methods of decision, and that which one risks the least will depend on the true variety yields, which we will never know. The problem is not as hopeless as it sounds, for Wald has taken a large step forward, and shown that any decision method can be replaced by one derived from *a priori* probability considerations without increasing the risk under any set of true variety yields. This is a great simplification—but the mathematical complications of dealing with 120,904 functions of seven variables are still awe-inspiring. If we were able to carry through this program—to set the risks intelligently, to carry out the mathematics, and to choose wisely among the admissible decision functions—we would surely do much better than we can hope to do now, but for the present we need to adopt a simpler procedure. (Note. The case of 3 or 4 means has been attacked within the scope of Wald's theory by Duncan [7] using a different philosophy which emphasizes conclusions about pairs of means.)

At a low and practical level, what do we wish to do? We wish to separate the varieties into distinguishable groups, as often as we can without too frequently separating varieties which should stay together. Our criterion of "not too frequently" is a rough one, and may frequently be expressed by saying "at the 5% level" or "at the 1% level." The meaning of these words deserves a little discussion. To the writer they do not mean, "so that an entirely nonexistent effect will be called real once in twenty times, or once in a hundred times", but rather that "with the same sort of protection against false positives that I usually have when I make tests of significance on hypotheses suggested by the results tested, successive tests of hypotheses, tests of regression on selected variables, etc." For these reasons, working "at the 5% level" may involve the successive use of tests, each of which yields false positives five times in a hundred, but, when used together, will yield seven, eight or nine false positives in a hundred. It is such a primitive and rough standard that we wish to combine with a primitively and roughly outlined desire to detect effects which are really there. From these primitive desires we are to seek a method.

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

3. THE STIGMATA OF DIFFERENCE

When the real differences between variety means are large, how do we realize this fact? Three vague criteria come naturally to mind:

(1) There is an unduly wide gap between adjacent variety means when arranged in order of size,
(2) One variety mean struggles too much from the grand mean,
(3) The variety means taken together are too variable.

It is these three criteria we are going to apply in order to break up an observed set of means. We need, then quantitative tests for detecting (1) excessive gaps, (2) stragglers, (3) excess variability. These must be used when the variance of an individual observed mean is not known exactly, but rather when it is estimated from some other line of an analysis of variance table. The tests which we use must therefore be Studentized tests. Exact tests for (2) and (3) are available, but for the present we shall confine ourselves to an approximate and conservative test for (1).

If there are only two variety means, the largest gap between adjacent means is the same as the absolute value of the difference of the means. If $m_1 > m_2$, and s_m^2 is the estimated variance of a single mean, then

$$\frac{m_1 - m_2}{s_m 2^{1/2}}$$

has one-half of a t-distribution and assuming normality, exceeds 2.447 only 5% of the time when the two true means are equal and s_m is based on 6 degrees of freedom. There are good reasons based on experimental sampling (Section 9) and numerical integration (Section 8) to believe that the one-sided 5%, 2%, 1% points of

$$\frac{\text{largest gap between adjacent means}}{s_m 2^{1/2}}$$

are smaller than the corresponding two-sided percentage points of t. If this is true we will be conservative to use this ratio and the two-sided percentage points of t as a test of excessive gapping. The reasons are discussed in a later section.

The exact test of

$$\frac{m_1 - \overline{m}}{s_m}$$

where m_1 is the largest mean and $\overline{m}$ is the grand mean has been discussed for the case of normality by K. R. Nair [4] in a very recent number of *Biometrika*. Simple and satisfactory *empirical* approximation to the

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

upper percentage points (between 10% and 0.1%) can be obtained by treating

$$\frac{\left(\frac{m_1 - \overline{m}}{s_m}\right) - \frac{6}{5}\log_{10} k}{3\left(\frac{1}{4} + \frac{1}{n}\right)} \qquad (k > 3 \text{ means})$$

or

$$\frac{\left(\frac{m_1 - \overline{m}}{s_m}\right) - \frac{1}{2}}{3\left(\frac{1}{4} + \frac{1}{n}\right)} \qquad (3 \text{ means})$$

as unit normal deviates, where s_m is based on n degrees of freedom. The adequacy of this approximation—which avoids the use of multiple entry tables—is also discussed in Section 6.

The exact test of excessive spread in general will of course be the familiar F-test (or z-test).

We propose to use these tests successively, and in the following order and manner. First, apply the gap test to break up the means into one or more broad groups. Second, apply the straggler test *within* these groups to further break off stragglers within groups. Third, apply the F-test to these new subgroups to detect excess variability. It is hard to see how to find the frequency of false positives with the whole system analytically, but the writer conjectures that, if the same level, such as 5%, is used in all three tests, the frequency of false positives will be between 1.2 and 1.6 times the level used (i.e., between 6% and 8% when a 5% level is used). This is about where the frequency of false positives stands for many repeated and result-guided tests of significance now in actual practice.

4. DETAILED PROCEDURE ILLUSTRATED BY EXAMPLES

The two examples we are going to use are those discussed by Newman [5] in connection with the use of the Studentized range. The advantages of continuing with the same examples may compensate for disadvantages of lack of simplicity, and in the case of the first example, lack of appropriateness. This first example is a 6 × 6 Latin square with potatoes, cited by Fisher [1] in Article 36 of *The Design of Experiments*. As first presented this example is stated to be six fertilizer treatments in a Latin Square, and Newman seems to have based his example on this discussion. Later on in the book (Article 64), Fisher points

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

out that these treatments were a 2 × 3 factorial design in nitrogen and phosphorus, so that there were specific individual degrees of freedom whose analysis was planned when the experiment was designed. These were *not* 6 treatments all on an equal footing, and overall analysis is not appropriate, *but* we shall proceed to analyze them as if they were six treatments about which there is no advance information. The six means were (*A*) 345.0, (*B*) 426.5, (*C*) 477.8, (*D*) 405.2, (*E*) 520.2, (*F*) 601.8, and the estimated standard deviation of a mean was $s_m = 15.95$.

Step 1. Choose a level of significance. For this example we shall choose 5%.

Step 2. Calculate the difference which would have been significant if there were but two varieties.

The two-sided 5% point of t on 20 degrees of freedom is 2.086. For this example, then, this least significant difference is $2.086\ (2^{1/2})15.95 = 47.0$.

Step 3. Arrange the means in order and consider any gap longer than the value found in Step 2 as a group boundary.

Arranged in order, the means are 345.0, 405.2, 426.5, 477.8, 520.2, 601.8 and the differences 405.2 − 345.0 = 60.2, 477.8 − 426.5 = 51.3, and 601.8 − 520.2 = 81.6 exceed 45.7, so that we have divided the varieties into four groups: 345.0 (*A*) by itself, 405.2 (*D*) and 426.5 (*B*) together, 477.8 (*C*) and 520.2 (*E*) together, and 601.8 (*F*) by itself.

If no group contains more than two means, the process terminates. The first example having terminated, we must pass to another to illustrate the continuance of the process. Snedecor [6] gives as Example 11.28 on p. 274 (of the 4th edition) the results of a 7 × 7 Latin Square with potatoes. The means were (*A*) 341.9, (*B*) 363.1, (*C*) 360.5, (*D*) 360.4, (*E*) 379.9, (*F*) 386.3, (*G*) 387.1 and s_m on 30 degrees of freedom was 9.52. Choosing the 5% level, for which t on 30 degrees of freedom is 2.042, we find $t(2^{1/2})s_m = 27.5$. In order, the means are 341.9, 360.4, 360.6, 363.1, 379.9, 386.3, and 387.1 No difference between adjacent means exceed 27.5, so that there is only one group at the end of Step 3.

Step 4. In each group of 3 or more means find the grand mean, the most straggling mean and the difference of these two divided by s_m. Convert these ratios into approximate unit normal deviates by finding

$$\frac{\dfrac{m - \overline{m}}{s_m} - \dfrac{6}{5}\log_{10} k}{3\left(\dfrac{1}{4} + \dfrac{1}{n}\right)}, \qquad (k > 3 \text{ means in the group}),$$

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

$$\frac{\dfrac{m - \overline{m}}{s_m} - \dfrac{1}{2}}{3\left(\dfrac{1}{4} + \dfrac{1}{n}\right)}, \qquad (3 \textit{ means in the group}).$$

Separate off any straggling mean for which this is significant at the chosen two-sided significance level for the normal.

For the Snedecor example we find $\overline{m} = 368.5$, and the most straggling mean is $m = 341.9$. The ratio is $26.6/9.51 = 2.80$. Further $\log_{10} 7 = .845$ and we are to consider

$$\frac{2.80 - \dfrac{6}{5}\,.845}{3\left(\dfrac{1}{4} + \dfrac{1}{30}\right)} = \frac{60}{51}\,(2.80 - 1.01) = 2.10.$$

Since the two-sided 5% level for the unit normal is well known to be 1.96, we must separate 341.9 (A).

Step 5. If Step 4 changed any group, repeat the process until no further means are separated in the old groups. The means separated off from one side of a group form a subgroup. If there are any subgroups of three or more when no more means are being separated from groups, apply the same process (Steps 4 and 5) *to the subgroups.*

The old group in the Snedecor example now contains 6 means, and its grand mean has increased to $\overline{m} = 372.9$. The most straggling mean is 387.1 for which $(387.1 - 372.9)/9.51 = 1.49$. The approximate unit normal deviate is $60/51\ (1.49 - 0.93) = 0.66$, which is far from significance. Step 5 has produced no further effect.

Step 6. Calculate the sum of squares of deviations from the group mean and the corresponding mean square for each group of or subgroup 3 or more resulting from Step 5. Using s_m^2 as the denominator, calculate the variance ratios and apply the F-test.

In the Snedecor example, we have one group of six, for which the sum of squares of deviations is 829 and the mean square 166. The denominator is $(9.51)^2 = 90.4$ and the F-ratio 1.83 on 4 and 30 degrees of freedom, which is near the 12% point. Thus there is no overall evidence of difference in yield for these six varieties.

If varieties (B) 363.1, (C) 360.6, and (D) 360.4 had been *known in advance* to be different *as a class* from varieties (E) 379.9, (F) 386.3, and (G) 387.1, it would be fair to introduce a single degree of freedom for this

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

comparison, giving an analysis of variance (in terms of means) like this.

	Degrees of Freedom	Mean Square
BCD vs *EFG*	1	794
Varieties within classes	4	35
Error	30	90.4

From this we could conclude that *BCD* and *EFG* were different, even at the 1% level. There is no valid basis for this particular conclusion *unless* the classes are *uniquely* known in advance of the experiment. (There are 20 ways to split six varieties into two classes of three varieties each, so that the apparent significance of the most significant split would be expected to be at a percentage level near 1/20th of the percentage level of the whole group. The actual figures are, approximately, 0.6% and 12% and their agreement with the 1-to-20 ratio is unusually close.)

In the Fisher example, the proposed procedure gave the following result: Variety *A* (345.0) *is significantly lower than varieties D* (405.2) *and B* (426.5), *these in turn are significantly lower than C* (477.8) *and E* (520.2), *and in turn these are significantly lower than F* (601.8). *All significance statements are statistical, and are at the 5% level or better.*

In the Snedecor example, the proposed procedure gave the following result: *Variety A* (341.9) *was significantly lower than some of the varieties C* (360.1), *D* (360.4), *B* (363.1), *E* (379.9), *F* (386.3), *and G* (387.1) *at the 5% level or better, the group of 6 varieties showed no overall evidence of internal differences at the 5% level.*

These conclusions should be compared with those of Newman, who used the Studentized range to conclude in the first case that even taking *ADB* and *CEF* as two groups, neither was homogeneous. This is consistent with the result of the present analysis, but far less detailed. For the Snedecor example, Newman found that if either *A* or *F* and *G* together were made a separate group, the remainder seemed homogeneous. This is again consistent, but less detailed, since the present process finds definite reason to suppose that it is *A* which is inhomogeneous. (How *much* stronger is the evidence we have against *A* than against *F* and *G* is another matter.)

The writer feels that the proposed procedure is direct, reasonably simple, involves no new tables, and is ready to be used in practice and thereby put to the ultimate test.

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

5. THE DISTRIBUTION OF THE MAXIMUM GAP

We are interested in the following problem:

"Let a sample of k values (in our case means) be drawn from a normal distribution, of which we know only an independent estimate s of its standard deviation, based on n degrees of freedom. What is the distribution of

$$\frac{\text{largest gap between ordered observed values}}{s}\ ?"$$

The methods of Hartley, reviewed in detail by Nair [4], would allow us to solve this problem for finite n if we knew the answer for infinite n, that is for the case where we know σ, the standard deviation of the normal population.

The problem of the distribution of the largest gap in a sample of k values from a unit normal distribution can easily be attacked by experimental sampling (see Section 9). The fact that the random normal deviates of Mahalanobis [3] are printed in blocks of five leads one to study $k = 5$ and $k = 10$ first. The first 1000 blocks of five in that table were used (skipping block 768, which was marked as an error in the copy available to the author).

The results are shown below:

TABLE 1

UPPER PERCENTAGE POINTS OF THE LARGEST GAP IN AN ORDERED SAMPLE OF k FROM A UNIT NORMAL

%	$k = 2$ theory	$k = 5$ sample of 1000 cases	$k = 10$ sample of 500 cases
10	2.33	1.86	<1.50
5	2.77	2.13	1.68
2	3.29	2.49	1.95
1	3.64	2.77	2.42

The theoretical values for $k = 2^{1/2}$ are values of $t(2^{1/2})$ and are accurate, the others are as found by experimental sampling and may deviate from accuracy by perhaps 1 or 2 in the first decimal. They are sufficiently accurate, however, to indicate that the upper percentage point decreases as k increases. Thus if we use the values for $k = 2$ we will make a conservative test. This is true for $n = \infty$, and by the nature of Hartley's expansion it will continue to hold for all reasonable values of n.

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

TABLE 2

QUALITY OF APPROXIMATION OF PERCENTAGE POINTS FOR THE STRAGGLER TEST

Normal percentage point minus accurate percentage point	Occurs for	Cases
0.15 to 0.20	3 means, $n \leq 15$	6
0.10 to 0.15	5%, 3 or 4 means, $n \leq 24$ 1%, 4 means, $n \leq 11$ 1%, 3 means, $n \leq 30$	33
0.05 to 0.10	5%, 5 means, $n \leq 24$ 5%, 3 or 4 means, $n \leq 60$ 1%, 4 means, $n \leq 11$ 1%, 3 means, $n \leq 120$	21
−0.05 to +0.05	otherwise	154
−0.10 to −0.05	10%, all cases 5%, 9 means, $n = 10, 11$ 1%, 8 or 9 means, $n = 20$	20

The discussion in Section 2 suggests, of course, that it would be correct and wise to find accurately the percentage points of the largest gap for various values of k and then use the appropriate values of k. This is not being suggested for the present, because:

(1) the necessary table does not exist,
(2) it would complicate the procedure,
(3) there are problems in choosing the appropriate value of k,
(4) the simpler proposed procedure has not yet been used enough to show its characteristics.

6. THE STUDENTIZED EXTREME DEVIATE

In his recent paper, Nair [3] has given the following upper percentage points for 3 to 9 samples: (A) the 10%, 5%, 2.5%, 1%, 0.5% points for $n = \infty$, (B) the 5% points for n from 10 to 20 and 24, 30, 40, 60, 120, ∞, (C) the 1% points for the same values of n. The accuracy of our rough approximation is most easily considered by transforming them into percentage points for the approximate unit normal deviates—these are what should be used for accuracy,—and comparing these with the percentage points of the normal—these are what we propose to use. Such a comparison has the following results, (Table 2).

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

Thus for about two-thirds of the cases tabulated by Nair, the error is less than 0.05, and is surely negligible in practice.

In doubtful cases, a more precise approximate test may be made as follows. Let

$$w = \frac{|\, m - \overline{m} \,|}{s_m} \qquad (m \text{ an extreme mean})$$

Then treat

$$\left(\frac{k}{k-1}\right)^{1/2}\left(w - \frac{10(w-1.2)}{3n}\right)$$

as a unit normal deviate and multiply the tail area by k if only one kind of straggler (high or low) could be considered, and by $2k$ otherwise. Thus if $\overline{m} = 52$, $m = 43$, $s = 4$, $k = 13$, $n = 28$

$$w = \frac{143 - 521}{2} = \frac{9}{4} = 2.25,$$

$$\left(\frac{13}{12}\right)^{1/2}\left(2.25 - \frac{10(1.05)}{3.28}\right) = 1.041(2.25 - 0.13) = 2.20$$

Now the probability of a unit normal deviate = 2.14 is 0.01390 (from any normal table, e.g. Fisher and Yates [2] Table IX where 98.610% corresponds to a probit of 7.1200). Multiplying by 11 gives 15.3% as the approximate significance, if only low means are of interest, while the level is 30.6% when either high or low means are involved.

This approximation is discussed by Nair [4] for the case $n = \infty$, where it is due to McKay. Nair shows that it is very good indeed. The effectiveness of the term in n^{-1} may be tested by calculating the true percentage points for $w - 3n^{-1}(w - 1.2)$ from Nair's tables.

TABLE 3

UPPER PERCENTAGE POINTS FOR $w - 10/3n\ (w - 1.2)$

	5% points				1% points			
n	$k = 3$	5	7	9	$k = 3$	5	7	9
10	1.75	2.06	2.24	2.35	2.24	2.57	2.73	2.85
15	1.76	2.08	2.26	2.39	2.27	2.62	2.81	2.93
20	1.76	2.08	2.27	2.39	2.25	2.62	2.82	2.92
30	1.75	2.09	2.27	2.40	2.25	2.61	2.82	2.93
∞	1.74	2.08	2.27	2.39	2.22	2.57	2.76	2.88

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

The errors involved in the use of the values at the bottom of the columns of Table 3 instead of those above them can hardly ever be of practical importance.

The previous approximation is recommended for routine work since it involves less computation and no changing of significance levels. Both approximations are only good for upper percentage points in the significance test range. The latter approximation should meet all practical needs.

The writer would rarely bother with the more precise approximation except possibly for the cases where the error of the rough test is between -0.10 and -0.05. The original experimental values are likely to be somewhat non-normal with large tails. An accurate allowance for this would be hard to compute, but it would increase the accurate percentage point slightly, more for smaller n. The rough approximation tends to compensate for this fact in most cases.

7. THE DISTRIBUTION OF LONG GAPS IN A SAMPLE OF k FORM ANY POPULATION

While we could concern ourselves with the distribution of the longest gap, the next longest gap, and so on, it seems theoretically better and practically simpler to do something somewhat different. We are going to calculate the expected number of gaps longer than a length G, which we denote by p_1. For the sort of test considered above, there is much reason to use p_1. For p_1 is the fraction of gaps per sample which will be falsely judged significant. If it is as bad to find two false gaps in a sample as to find one false gap in each of two samples, then we should consider p_1.

Now we shall take the definition of a gap starting at y to be that y is the left hand of the gap. If y is the left-hand end of a gap of length at least G, we have the following table of elementary probabilities:

Event	Probability
One observation must fall between y and $y + dy$	$k\,dF(y)$
$k - 1$ observations must fall between $-\infty$ and y or between $y + G$ and $+\infty$	$\{F(y) + 1 - F(y + G)\}^{k-1}$
Not all $k - 1$ observations can fall between $-\infty$ and y	$-(F(y))^{k-1}$

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

hence

$$p_1 = k \int_{-\infty}^{+\infty} \{(F(y) + 1 - F(y + G))^{k-1} - (F(y))^{k-1}\}\, dF(y).$$

8. THE SYMMETRICAL CASE

If the distribution of x is symmetrical about zero, we may count only the gaps with centers to the left of the origin and then double. The expression for p_1 follows from:

Event	Probability
One observation must fall between y and $y + dy$	$k\, dF(y)$
$k - 1$ observations must fall between $-\infty$ and y or $y + G$ and $+\infty$	$(F(y) + 1 - F(y + G))^{k-1}$
Not all $k - 1$ observations can fall between $-\infty$ and y or $-y$ and $+\infty$	$-(2F(y))^{k-1}$

Since $y \leq -\frac{1}{2}G$, and since the result is to be doubled, we have

$$p_1 = 2k \int_{-\infty}^{-\frac{1}{2}G} \{(F(y) + 1 - F(y + G)^{k-1} - (2F(y))^{k-1}\}\, dF(y)$$

Making the substitutions $u = F(y)$, $h(u) = F(y) + 1 - F(y + G)$, this becomes

$$p_1 = 2k \int_{-\infty}^{F(-\frac{1}{2}G)} h^{k-1}\, du - \left\{2F\left(-\frac{G}{2}\right)\right\}^k$$

For reasonably large G, the second term is fairly small and we can get an accurate value of p_1 with a reasonable amount of labor.

As an example, let us take the unit normal distribution and $G = 2$. Since $h(u)$ is non-analytic near 1 and has a minimum at $F(-1) = .1587$, it is natural to break the integral up into parts as follows:

$$p_1 = 2k \int_0^{.0004} h^{k-1}\, du + 2k \int_{.0004}^{.004} h^{k-1}\, du + 2k \int_{.004}^{.04} h^{k-1}\, du$$

$$+ 2k \int_{.04}^{.16} h^{k-1}\, du - 0.0013(2k)(h(.1587))^{k-1} - (.3174)^k$$

Calculating h to four decimals, applying Simpson's rule to the range from 0 to .004, and the corresponding six-panel rule to the other three

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

ranges yields the following results, where the terms are given in the order of the formula above:

k	+	+	+	+	−	−	p_1
2	.00151	.01168	.07875	.16781	.00165	.10074	.15736
3	.00214	.01426	.06603	.08885	.00079	.03206	.13843
4	.00270	.01553	.04227	.04224	.00032	.01014	.09228
5	.00320	.01590	.03035	.01908	.00013	.00322	.06518
6	.00363	.01567	.02635	.00835	.00005	.00102	.05293
7	.00402	.01508	.01880	.00353	.00002	.00032	.04109
8	.00436	.01426	.01342	.00154	.00001	.00010	.03347
9	.00468	.01330	.00959	.00065	.00000	.00003	.02819
10	.00493	.01230	.00691	.00029	.00000	.00000	.02443

The value for $k = 2$ can of course be calculated directly as

$$2(1 - F(2^{1/2})) = 2(.0787) = .1574$$

The results are probably accurate to 1 or 2 in the fourth place. They can be conveniently stated as in the following table:

TABLE 4

NUMBER OF GAPS LONGER THAN 2.00 EXPECTED PER 100 SAMPLES OF k FROM THE UNIT NORMAL

k	2	3	4	5	6	7	8	9	10
gaps / 100 samples	15.74	13.84	9.23	6.52	5.29	4.11	3.35	2.82	2.44

9. RESULTS OF EXPERIMENTAL SAMPLING

The results of the experimental sampling of 1000 sets of 5 from Mahalanobis' approximation to the unit normal are given in the following table, (Table 5).

The approximate normality of $(\text{largest gap})^{1/2}$ in this sample, as indicated by the correspondence of the last two columns between the 2% points is striking. For comparison it seemed worthwhile to examine the normality of $(\text{largest gap})^{1/2}$ for $k = 2$, where the probability of a gap $\geq G$ is $2N(G/2)$, where $N(u)$ is the unit normal cumulative. This gives the following results, (Table 6).

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

TABLE 5

RESULTS OF EXPERIMENTAL SAMPLING. DISTRIBUTION OF LARGEST GAPS IN 1000 SAMPLES OF 5

Cell	Number	Cumu.	Equiv. Norm. Dev.	$\frac{(\text{gap})^{1/2} - 1.07}{.23}$
.185– .199	2	2	−2.88	(−2.70)
.200– .299	9	11	−2.29	(−2.26)
.300– .399	20	31	−1.87	−1.90
.400– .499	28	59	−1.56	−1.57
.500– .699	97	156	−1.01	−1.00
.700– .899	141	297	−0.53	−0.52
.900–1.099	172	469	− .08	− .09
1.100–1.299	149	618	0.30	0.30
1.300–1.499	126	744	0.66	0.68
1.500–1.699	110	854	1.05	1.00
1.700–1.899	56	910	1.34	1.34
1.900–2.099	36	946	1.61	1.64
2.100–2.299	24	970	1.88	1.90
2.300–2.499	11	981	2.07	2.12
2.500–2.699	8	989	2.29	(2.51)
2.700–2.899	4	993	2.46	(2.77)
2.900–3.099	4	997	2.75	(2.99)
3.100–3.299	2	999	3.09	(3.20)
...				
4.000–4.099	1	1000	∞	

Here the fit is good between the 10% points. This suggests that the $(\text{largest gap})^{1/2}$ may be a convenient interpolation variable.

The number of cases ≥ 2.00 actually found was 68, while the number to be expected according to the last section was 65.2 less an allowance for large double gaps which might amount to one unit. Finding 68 instead of 64 is a deviation of 0.5σ, and is highly reasonable.

For $k = 10$, the count was only made for gaps ≥ 1.5, with the following results, (Table 7).

The fit here is reasonably good out to the 5% point. Since theory predicts about 12.2 beyond 2.00 instead of 9 observed, there is no serious disagreement here.

If we want to make real use of this $(\text{gap})^{1/2}$ variable, we may use the known percentages beyond 1.414, found for k between 2 and 10 in the last section to fix lines in the plane of the mean and standard deviation

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

TABLE 6

CUMULATIVE FOR (LARGEST GAP)$^{1/2}$ IN SAMPLES OF 2 FROM THE UNIT NORMAL

%	gap	(gap)$^{1/2}$	Equiv. Norm. Deviate	$\frac{(\text{gap})^{1/2} - .98}{.44}$
1	.0177	.134	−2.33	(−1.95)
2	.0357	.189	−2.05	(−1.80)
5	.0891	.299	−1.64	(−1.55)
10	.1781	.423	−1.28	−1.26
20	.360	.600	−0.84	−0.86
50	.960	.980	0.00	0.00
80	1.825	1.353	0.84	0.85
90	2.350	1.536	1.28	1.26
95	2.794	1.672	1.64	(1.57)
98	3.308	1.821	2.05	(1.91)
99	3.650	1.914	2.33	(2.12)

TABLE 7

RESULTS OF EXPERIMENTAL SAMPLING
DISTRIBUTION OF LARGEST GAPS IN 500 SAMPLES OF 10

Cell	Number	Cumul.	Equiv. Norm. Deviate	$\frac{(\text{gap})^{1/2} - 0.85}{.24}$
−1.499	454	454	1.33	1.38
1.500–1.599	15	469	1.54	1.53
1.600–1.699	9	478	1.71	1.68
1.700–1.799	9	487	1.94	1.82
1.800–1.899	2	489	2.01	1.98
1.900–1.999	2	491	2.10	2.08
2.000–2.199	1	492	2.14	(2.39)
2.200–2.399	2	494	2.26	(2.48)
2.400–2.599	3	497	2.51	(2.93)
2.600–2.799	2	499	2.88	(3.13)
...				
3.100–3.199	1	500	∞	

of the approximation. A little bold, dashing, freehand, two-dimensional interpolation produces the following results:

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms

TABLE 8

TENTATIVE BEHAVIOR OF (LARGEST GAP)$^{1/2}$ FOR SAMPLES OF k FROM THE UNIT NORMAL

Parameters			Levels for (gap)$^{1/2}$			Levels for gap		
k	m	s	5%	2.5%	1%	5%	2.5%	1%
2	0.98	0.43	1.69	1.82	1.98	2.8	3.3	3.9
3	1.03	0.36	1.62	1.74	1.87	2.6	3.0	3.5
4	1.06	0.27	1.50	1.59	1.69	2.3	2.5	2.8
5	1.06	0.23	1.43	1.51	1.60	2.0	2.3	2.6
6	1.06	0.22	1.42	1.49	1.57	2.0	2.2	2.5
7	1.04	0.21	1.39	1.45	1.53	1.9	2.1	2.3
8	1.02	0.21	1.37	1.43	1.51	1.9	2.0	2.3
9	1.00	0.21	1.35	1.41	1.49	1.8	2.0	2.2
10	0.99	0.22	1.33	1.40	1.48	1.8	2.0	2.2

By a stroke of luck, the levels for the gap itself might be accurate to one or two tenths. These are, of course, unstudentized levels.

REFERENCES

[1] R. A. Fisher 1935–1947, *The Design of Experiments*, 4th edition 1947, Oliver and Boyd, Edinburgh.

[2] R. A. Fisher and F. Yates, 1938–1948, *Statistical Tables*, 4th edition 1948, Oliver and Boyd, Edinburgh.

[3] P. C. Mahalanobis et al 1934, "Tables of Random Samples from a Normal Population", *Sankhyā 1* (1933–34), pp. 289–328.

[4] K. R. Nair 1948, "The distribution of the extreme deviate from the sample mean and its Studentized form". *Biometrika 35* (1948) pp. 118–144.

[5] D. Newman 1939, "The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation". *Biometrika 31* (1939–40) pp. 20–30.

[6] G. W. Snedecor 1937–1946, *Statistical methods*, 4th edition 1946, Collegiate Press, Ames, Iowa.

[7] David B. Duncan 1947, Iowa State College Thesis, iii + 117 pp.

This content downloaded from
200.144.62.89 on Sun, 07 Apr 2024 22:51:33 +00:00
All use subject to https://about.jstor.org/terms