JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1967, 10, 319-330 NUMBER 3 (MAY)

SOME STATISTICAL TREATMENTS COMPATIBLE WITH INDIVIDUAL ORGANISM METHODOLOGY[1]

SAMUEL H. REVUSKY

U.S. ARMY MEDICAL RESEARCH LABORATORY

Consider experimental treatments with consequences so irreversible that baseline performance cannot be recovered. The conventional method of assessing the effects of such treatments by statistical means involves separate experimental and control groups. An alternative proposed here is to administer the experimental treatment to each subject, one subject at a time and in a random order; whenever any subject receives the experimental treatment, those subjects which have not yet received it receive a control treatment. This procedure permits results significant at the one-tailed 0.05 level to be obtained with four subjects; if a two-group procedure evaluated by means of the U test is used, a minimum of six subjects is needed for the same significance level. More generally, the procedure permits equal sensitivity to any experimental effect with over 30% fewer subjects than a two-group procedure. Extensions of the basic method are made to a variety of levels of the experimental treatment and to treatments without irreversible effects, and limitations of the method are discussed.

CONTENTS

Probability theory, the source of inferential statistics, began as a mathematical analysis of games of chance. It is such an accurate description of these games that if the outcome of a game diverges radically from its predictions, there is reason to suspect that it is not a game of chance; for instance, a poker game in which somebody receives five successive royal flushes is probably dishonest. The role of inferential statistics in the analysis of experiments is similar. An experiment is designed so that it may be interpreted as a game of chance if there is no experimental effect. Inferential statistics are then used to assess the likelihood that the experimental result would have been obtained if it were a game of chance (*i.e.*, if the null hypothesis were true). If the result is quite unlikely for a game of chance, it is assumed that the experiment has detected a real effect. Thus gambling, an activity in which the role of chance is ideally maximized, has contributed to the analysis of experimentation, an activity in which the role of chance is ideally minimized.

The requirement that statistically designed experiments be interpretable as games of chance in the absence of an experimental effect often conflicts with the classical principle underlying experimentation, the minimization of the role of chance. In the realm of operant conditioning, this conflict is so great that the use of inferential statistics is avoided. Analysis of the statistical requirement of random sampling, a major source of this conflict, led to the statistical procedures proposed here and therefore will be discussed in detail.

[1]Reprints may be obtained from the Publications Section, Army Medical Research Laboratory, Fort Knox, Kentucky 40121.

RANDOM SAMPLING

Random sampling, the central assumption of nearly all inferential statistics, requires that all scores used in a statistical test be drawn at random from a population. Two examples illustrate how the statistical requirement of random sampling interacts with operant methodology. In the first, involving reversible treatments, a number of scores are randomly sampled from a single subject. In the second, involving irreversible treatments, only one score is randomly sampled from each subject, so that several animals are required.

Reversible Treatments

Of 20 test days with a well-trained animal, five are randomly assigned to placebo injection and five days to each of three drug levels. Under these conditions, the outcomes of each day are statistically independent, so that if the drug has no effects, use of separate scores as inputs into a statistical test (such as the analysis of variance) should yield chance results. As long as the test days are assigned to treatments at random, such random sampling is valid even if the test days themselves are preselected; it may be desirable, for instance, to test the animals every third session and allow recovery during the sessions in between. Thus, random assignment of different time intervals to different experimental treatments makes statistical analysis possible. This does not mean that it is necessary, since many results are entirely credible without statistics. Nor does this statistical validity insure scientific meaningfulness. To be scientifically meaningful, the randomly assigned treatments must have no important permanent effects on performance.

Irreversible Treatments

Figure 1 contains some hypothetical experimental data. Three pigeons were trained until they reached some criterion of stable performance. Then, after 20 additional daily sessions, they were poisoned at the point shown by the arrow. The poison is lethal between 10 and 20 days after administration. It is perfectly obvious from Fig. 1 that the poison interferes with performance beginning with the first session after injection. But rigorous use of statistics alone to evaluate the result does not indicate that the poison has any effect on performance at all; that is, the information used to gauge the reliability of this result cannot be analyzed statistically. Thus, any purely statistical decision is likely to be an absurdity.

At a maximum level of statistical rigor, there is no evidence that the decrement in performance, even if it is real, is due to the

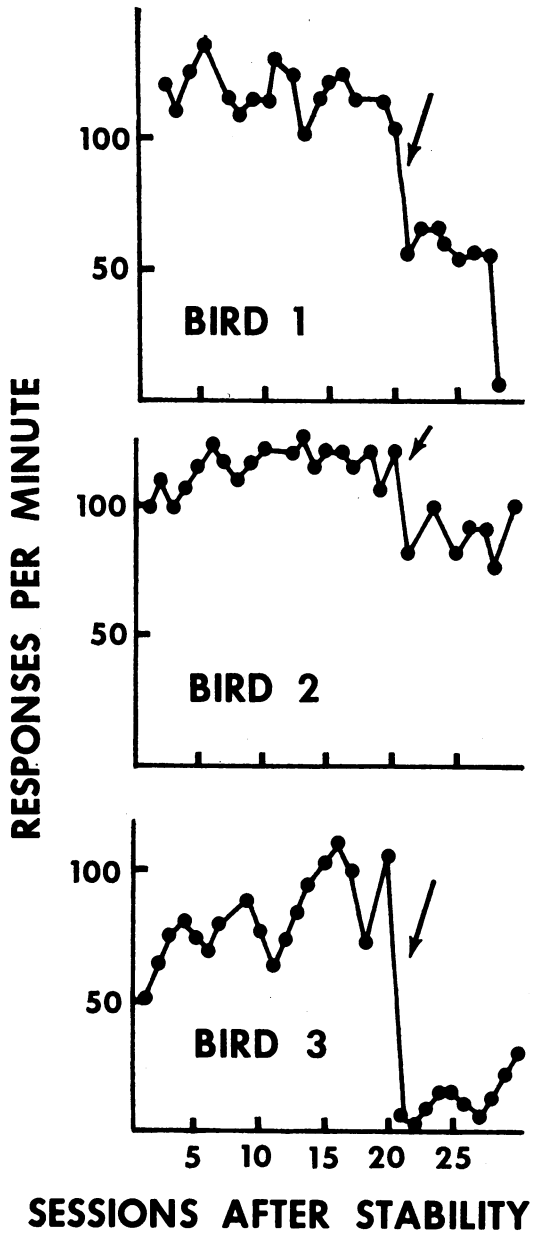

Fig. 1. Effects of poison, injected at the point indicated by the arrow, on performance (imaginary data).

poison because there are no control animals. Strictly speaking, the prepoison sessions are not controls because, with no other information, there is the possibility that 20 days after the animals reach the criterion of stability, performance decreases regardless of the experimental treatment.

Even if any statistically reliable decrement in performance after poison is administered is assumed to be due to the poison, the data in Fig. 1 are not adequate for a firm conclusion on statistical grounds alone. Figure 1 appears to show a real effect because there is no single post-poisoning score for any animal greater than any single pre-poisoning score. But this information cannot be used by inferential statistics. To allow statistical use of a number of scores from a single subject, the scores must be obtained by a random sampling process; that is, blocks of sessions, each yielding a single score, must be randomly assigned, some to the poison and some to a non-poison treatment. This is patently impossible; once a bird has been poisoned, it is impossible to recover baseline performance: the baseline data must precede the post-poison data. The only permissible random sampling is one data item per subject. For instance, the mean of the scores from the days following poisoning may be subtracted from the mean of the scores in the 10 days preceding poisoning; if these differences are significantly larger than zero, the effect is statistically reliable. It so happens that in the clear-cut case shown in Fig. 1, the differences would not be significantly positive if tested by the *t* distribution.

The inability to verify statistically the effect in Fig. 1 due to the requirement of random sampling is illustrated by what could be done if this requirement were eliminated. The five sessions before poisoning could be compared with the five sessions after it by means of the F test. The columns would be pre-poisoning and post-poisoning, and the rows would be subjects; there would be five scores in each cell. Without a doubt, such a procedure would show a highly significant effect. But if no effect existed, the same procedure would tend to detect effects with a greater probability than would be expected from the F distribution.

For practical purposes, non-random sampling may yield spurious statistical results because the magnitude of any performance score may be directly related to the magnitude of the score obtained on the immediately preceding occasion; this is a positive serial correlation (with lag one). A method of using serial correlation statistics to estimate the seriousness of the lack of random sampling is described elsewhere (Revusky, 1961); regrettably, it is not rigorous because it depends on commonsensical assumptions.

THE R_n STATISTIC

For these reasons, analysis of the effects of irreversible treatments on relatively few subjects is probably best done without statistics if the effect is clear-cut; it is better to rely on judgment and experience in interpreting data like those in Fig. 1 than to use a statistical method of evaluation which cannot take most of the available information into account. In some cases, however, the situation may be so ambiguous or novel as to demand statistical criteria of reliability. The remainder of this paper describes new statistical procedures for analyzing irreversible treatments without violating any statistical principles. These procedures are more sensitive than the conventional method of separate experimental and control groups for rigorously dealing with irreversible treatments. Therefore the experimenter who, for any reason whatsoever, feels that he must use separate groups in a particular experiment, may find that the following methods offer an alternative somewhat more compatible with individual organism methodology.

Consider a concrete example (a more formal description is available in Cronholm and Revusky, 1965). Suppose six animals are trained on an avoidance performance established over 100 daily 4-hr sessions. A hypothetical poison kills the animals in 8 to 12 hr. The problem is to determine if the poison has any effect on performance during the 4 hr immediately following its administration.

As indicated in the discussion of Fig. 1, if inferential statistics are to be used as the sole basis for deciding if an irreversible treatment has any effects on performance, there must be control animals as well as experimental animals. In a conventional, statistically orientated experimental design, the six animals would be split into two groups of three animals each: one group to receive the experimental treat-

ment (poison) just before its experimental session, and the other group to receive a control treatment (saline). The innovation which is the basis of the proposed statistic is that these six animals are subjected to the experimental treatment one at a time and in random order. Each time an animal is subjected to the experimental treatment, those animals not yet subjected to this treatment are subjected to the control treatment. The performance of each experimental subject is compared with those of the controls. Thus, instead of the conventional single experiment with three experimental subjects and three controls, there are six subexperiments each containing one experimental subject and some number of controls. In the conventional method, the experimental treatments are administered in one day; in the present method, they require six days, one for each subexperiment. The net result of this new procedure is that more statistically usable information is obtained from fewer animals than can be obtained from the conventional two-group design. Note that this procedure includes random sampling. Consider each subexperiment: the experimental treatment is administered to the subjects in random order, so that each of the animals that has not yet received the experimental treatment is equally likely to be selected as the experimental animal of a given subexperiment.

The statistical method involves ranking the scores in each subexperiment; the rank of the experimental subject is the rank outcome of the subexperiment. The statistic, R_n, is the sum of the rank outcomes of the experimental subjects in each subexperiment. Table 1 is a precis of the experimental procedure for six animals.

Table 1

Precis of the experimental design used with R_n. Each line contains the possible outcomes of one subexperiment. The subexperiments are numbered in chronological order.

Subexperiment	*Possible, equiprobable ranks*
1	1, 2, 3, 4, 5, 6
2	1, 2, 3, 4, 5
3	1, 2, 3, 4
4	1, 2, 3
5	1, 2
6	1

A Priori Probabilities

The *a priori* probabilities are calculated under the assumption that the poison has no effect, so that the random selection of the experimental subject in each subexperiment alone determines the probability of any rank outcome. This implies that in each subexperiment each of the possible rank outcomes is equally probable. Thus, in subexperiment 1, each rank shown in Table 1 has a probability of 1/6. In subexperiment 2, each possible rank has a probability of 1/5, and so on. In terms of dice, subexperiment 1 is similar to the toss of a true die with rank outcomes equivalent to the number of pips on each side. Subexperiment 2 is the toss of a five-sided die, and so on. This analogy concretely shows that the rank outcome of a subexperiment is independent of the rank outcome of any other subexperiment as long as the experimental operation has no effect. When first familiarized with the R_n procedure, people frequently decide that any interdependence of the scores obtained by a given subject in different subexperiments must affect the probability a given rank will be obtained. Because of the random selection of the experimental subject in each subexperiment, this impression is incorrect as long as no nonchance (experimental) effect is present.

The Probability Generating Function

Given the above assumptions, each subexperiment has a probability generating function (Feller, 1957) of its own, which is of no intrinsic interest, but is necessary for the understanding of the probability generating functions of the statistics to be described in this paper. When k is the number of possible ranks, this function[2] is

$$\frac{1}{k}\sum_{i=1}^{k} s^i.$$

The coefficient of the ith power of s in this function is equal to the probability that the rank outcome of the subexperiment will be equal to i; s has no numerical meaning and its

[2]Typically, a mathematical label is used for a function so that it is shown as an equation with a term designating the function to the left of the symbol "=" and the function itself to its right. In this paper, however, the left term is described verbally, so that only the generating function itself is shown.

only purpose is to supply a place for the superscript i, which indicates the outcome for which the coefficient of s^i is the probability. For instance, if k = 5, as in the second subexperiment of Table 1, the function is

$$\frac{s^1 + s^2 + s^3 + s^4 + s^5}{5}.$$

This function means that each rank from 1 to 5 has a probability of 1/5. As previously mentioned, the statistic used to evaluate the probability of the entire series of subexperiments is simply the sum of the rank outcomes in each subexperiment, called R_n. To obtain its probability generating function, multiply together the generating functions for each subexperiment (Feller, 1957). In this multiplication, s^i is treated as a number so that i acts as an exponent; when the product is interpreted, values of s with the same exponent are segregated and their coefficients are added together. The exponent of s then corresponds to a sum of ranks and its coefficient to the probability that the sum will be obtained by chance. For instance, with six subjects:

$$\left(\frac{\sum_{i=1}^{6} s^i}{6}\right)\left(\frac{\sum_{i=1}^{5} s^i}{5}\right)\left(\frac{\sum_{i=1}^{4} s^i}{4}\right)\left(\frac{\sum_{i=1}^{3} s^i}{3}\right)\left(\frac{\sum_{i=1}^{2} s^i}{2}\right)\left(\frac{\sum_{i=1}^{1} s^i}{1}\right)$$

$$= \frac{s^6}{720} + \frac{5s^7}{720} + \frac{14s^8}{720} + \frac{29s^9}{720} + \frac{49s^{10}}{720} + \frac{71s^{11}}{720}$$
$$+ \frac{90s^{12}}{720} + \frac{101s^{13}}{720} + \frac{101s^{14}}{720} + \frac{90s^{15}}{720} + \frac{71s^{16}}{720}$$
$$+ \frac{49s^{17}}{720} + \frac{29s^{18}}{720} + \frac{14s^{19}}{720} + \frac{5s^{20}}{720} + \frac{s^{21}}{720}.$$

The logic of this generating function depends on the fact that each distinguishable result of a series of six subexperiments is equally probable if the experimental treatment has no effect, where a result is defined as distinguishable from some other result if the rank outcome of at least one subexperiment is different. The numerator of each term is the number of distinguishable results which yield a rank sum equal to the power of s, while its denominator is the total number of distinguishable results. Therefore, the coefficient of each power of s is the probability that a sum of ranks equal to that power will be obtained. The probabilities derived by means of this generating function are not cumulative; to obtain the cumulative probability, the probabilities of all more extreme events must be added to the probability of the event itself. For instance, the probability that $R_n = 8$ is 14/720; the probability that $R_n \leq 8$ is 1/720 + 5/720 + 14/720 = 20/720.

Transformations of the Raw Scores Prior to Ranking

In order to permit random sampling, the comparisons used to assess the probability of the outcomes are between subjects, although it is usually true that within-subject comparisons are more sensitive experimentally. Thus, statistical rigor has yielded an absurdity in experimental terms. Fortunately, the raw scores can be transformed to minimize individual differences and these transformed scores can then be ranked. For instance, absolute or percentage change from baseline can be compared, where the baseline may be mean or median performance during some number of days immediately preceding the subexperiment. Any transformation of the raw scores is permissible provided that the experimental subject is equally likely to take on any one of the possible ranks under chance conditions. Thus, knowledge of the subject matter and experimental control can make the scores used as inputs into the statistical test more sensitive to the experimental treatment.

Significance Levels

Table 2 shows the maximum values of R_n significant at one-tailed probability levels under the assumption that the ranks of the experimental scores tend to be smaller than

Table 2

Maximum values of R_n significant at the indicated one-tail probability levels when the experimental scores tend to be smaller than the control scores.

No. of Subjects	*Significance Level*				
	0.05	*0.025*	*0.02*	*0.01*	*0.005*
4	4				
5	6	5	5	5	
6	8	7	7	7	6
7	11	10	10	9	8
8	14	13	13	12	11
9	18	17	16	15	14
10	22	21	20	19	18
11	27	25	24	23	22
12	32	30	29	27	26

the ranks of the control scores. Table 3 shows the minimum values of R_n significant under the assumption that the ranks of the experimental subjects tend to be higher than the ranks of the controls. If a two-tailed test is used, the null hypothesis may be rejected on the basis of either Table 2 or Table 3, provided the indicated probability level is

Table 3

Minimum values of R_n significant at the indicated one-tail probability levels when the experimental scores tend to be larger than the control scores.

No. of Subjects	*Significance Level*				
	0.05	*0.025*	*0.02*	*0.01*	*0.005*
4	10				
5	14	15	15	15	
6	19	20	20	20	21
7	24	25	25	26	27
8	30	31	31	32	33
9	36	37	38	39	40
10	43	44	45	46	47
11	50	52	53	54	55
12	58	60	61	63	64

doubled. Note that detection of a nonchance result at the one-tail 0.05 level is possible with four subjects; the Mann-Whitney U test, a rank test based on the two group procedure, requires six subjects. A more general comparison of the sensitivity of R_n and U is contained in the Appendix; roughly, R_n permits the use of 35% fewer subjects than a two-group procedure without any reduction in the likelihood that an experimental effect will be detected.

A normal approximation to the generating function is applicable when more than 12 subjects are used (Cronholm and Revusky, 1965). The following quantity is distributed with zero mean and unit variance so that its probability can be obtained from standard tables of the normal distribution:

$$Z_n = \frac{12\,R_n - 3n\,(n+3)}{\sqrt{2n\,(n-1)\,(2n+5)}}$$

where n is the number of subjects. When z_n yields a probability above 0.02, it may be desirable to reduce the absolute value of the numerator of the above equation by 6 to correct for continuity (Cronholm and Revusky, 1965).

Hypothetical Example

The effect of the hypothetical poison on the avoidance performance of five rats was investigated by means of the R_n procedure. The experimental treatment was injection with the poison just before the session, while the control treatment was saline injection. Table 4 shows the experimental results during the series of subexperiments and for five preceding baseline days; the label E after a score indicates poison was injected; C indicates saline was injected.

The differences between subjects suggested transformation of the raw scores into proportion change from baseline, where the baseline for any rat during any subexperiment is the mean of its responses during the immediately preceding five sessions. Symbolically,

$$\frac{R\text{-}B}{B}$$

where R is the number of responses during the subexperiment and B is the baseline.

Table 5 shows the input into the test for each subexperiment. The input score for the experimental subject of each subexperiment is labelled E; the other inputs are labelled C. For Rat 1 in subexperiment 1, the baseline, B, is the mean of the scores shown in Table 4 for days 46 through 50 and equals 916; the number of responses per hour during the subexperiment itself, R, is 890. Therefore, the input as shown in Table 5 is

$$\frac{R\text{-}B}{B} = \frac{890\text{-}916}{916} = -0.028.$$

The bottom row of Table 5 shows the rank outcome of each subexperiment. The sum of these, R_n, is 6; Table 2 indicates this result is significant at the one-tailed 0.05 level. Note that if the raw scores had been used instead of transformed scores or if the score of the preceding day had been used as the baseline, R_n would have equalled 5 and the result would have been significant at the 0.01 level; unfortunately, *ex post facto* choice of the type of score to be used in the test is invalid. Note also that a change in the performance of Rat 3 in Table 4 probably reduced the sensitivity of the test, which cannot magically turn poor data into good data.

Other Properties of R_n

Implicit in the above discussion are certain matters, which will now be made more explicit.

Table 4

Effects of Hypothetical Poison on Response Output in Avoidance

Rats	Days 46	47	48	49	50	51	52	53	54	55
1	890	780	970	1010	930	890C	770C	980C	710E	
2	1570	1470	1490	970	1130	740E				
3	640	680	590	580	530	890C	990C	770E		
4	930	900	840	960	860	750C	1130C	980C	860C	390E
5	1180	1040	1210	1120	970	990C	490E			

C—Control session.
E—Experimental session.
Days 46-50: Baseline for subexperiment 1.

(a) Note in Table 5 that the fifth subexperiment contains only Rat 4 as the experimental rat and no controls. Thus, the rank outcome is bound to be 1 regardless of the data and the only contribution of Rat 4 to the statistical outcome is its role as a control animal in earlier subexperiments. For this reason, statistical evaluation is possible without the final subexperiment. In fact, it may be possible to reach a valid statistical decision even earlier: calculate the value of R_n, which will be obtained if each remaining subexperiment yields the highest possible rank, and then calculate its value if each remaining subexperiment yields the lowest possible rank. If both yield the same statistical decision, the experiment is over for statistical purposes. If, for instance, Table 5 had shown a rank of 3 for the first subexperiment, a significant departure from chance would be impossible no matter what the results of the remaining subexperiments, and it might be desirable to save the well-trained animals for some other experiment.

(b) The above procedure requires that animals receive the control procedure repeatedly. In the case of some control procedures, such as surgical procedures, this may preclude a scientifically meaningful result.

(c) When the effects of the experimental treatment require a long time to emerge, the duration of a series of subexperiments may be prohibitively long. R_n is most useful when pretraining is lengthy and the test is short.

(d) Often, it is not feasible to have a number of animals all at the same performance level in order to initiate a series of subexperiments in exactly the manner described. Typically, the behavior of different animals becomes stable at different times; to wait until all have reached a criterion is frequently wasteful. Furthermore, facilities may not permit simultaneous training of all the animals. Under such circumstances, it is valid to assign each subject randomly to a subexperiment, run an appropriate number of control sessions after the criterion is reached, and then apply the experimental treatment. Interpret the data as though the animals had been run simultaneously. The reader is cautioned, however, that any procedure which confounds something else with the experimental treatment cannot

Table 5

Inputs into the R_n Test (Transformed from Table 4)

Rats	Subexperiments 1	2	3	4	5
1	−0.028C	−0.159C	+0.072C	−0.225E	
2	−0.442E				
3	+0.474C	+0.514C	+0.075E		
4	−0.165C	+0.311C	+0.079C	−0.081C	−0.574E
5	−0.103C	−0.540E			
Rank Outcome	1	1	2	1	1

C—Control session.
E—Experimental session.

yield results which can be assessed by statistical means; for instance, if it somehow happens that a number of subjects are administered the experimental treatment on the same day, characteristics of that particular day may produce a result which may be spuriously attributed to the experimental treatment.

EXTENSION OF THE R_n METHOD

As previously noted, a physical model of the R_n procedure under chance conditions is a dice throwing situation in which the number of faces of the die is reduced by one in each subexperiment. Other related experimental designs can also be interpreted as dice throwing situations; whenever this is true, the preceding methods probably can be adapted to permit a statistical analysis. The number of such possible techniques is so large that the experimenter will have to supply his own probability generating function if he departs from a straightforward use of R_n. In the hope that they will help experimenters devise techniques for their unique problems, a number of possible examples are considered below.

Several Levels of the Experimental Treatment; One Level in Each Subexperiment

Consider the effects of a poison on stabilized performance. Suppose there are three dose levels and all dose levels are assumed (*a priori*) to act in the same direction on performance. Begin with 10 subjects under a modification of the R_n procedure described by Table 6. The change from the R_n procedure is that for each subexperiment, one of the three dose levels is used for the experimental subject. Thus, three subexperiments are selected to test each of the three dose levels; the last subexperiment is not used for purposes of statistical inference because its outcome is predetermined. To assess the probability of the overall effect, simply use R_n, ignoring the individual dose levels. (If it is reasonable to suppose one dose level improves performance and a second dose level interferes with it, as may indeed be the more usual case, methods for assessing the overall effect used in the following section are more appropriate; in fact, these may be preferable if all dose levels act in the same direction, but some doses are considerably more potent than others.)

Table 6

The procedure by which the R_n technique is used to study the effects of three levels of a treatment.

Subexp.	*Level*	*Possible, equiprobable ranks*
1	A	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
2	B	1, 2, 3, 4, 5, 6, 7, 8, 9
3	C	1, 2, 3, 4, 5, 6, 7, 8
4	A	1, 2, 3, 4, 5, 6, 7
5	B	1, 2, 3, 4, 5, 6
6	C	1, 2, 3, 4, 5
7	C	1, 2, 3, 4
8	B	1, 2, 3
9	A	1, 2

To obtain a separate statistic for each dose level, add the rank outcomes obtained in the subexperiments in which that dose level was used. For dose levels A, B, and C, these sums are called r_A, r_B and r_C, respectively. To obtain a generating function for r_A, note that dose level A was used in subexperiments 1, 4, and 9 (Table 6). The dice analogy for each of these subexperiments, respectively, is a 10-sided die, a seven-sided die, and a two-sided die. Thus, the probability generating function or r_A is constructed much like the probability generating function for R_n

$$\frac{\left(\sum_{i=1}^{10} s^i\right)\left(\sum_{i=1}^{7} s^i\right)\left(\sum_{i=1}^{2} s^i\right)}{10 \cdot 7 \cdot 2}$$

where, as before, the coefficient of any power of s corresponds to the probability that a sum of ranks equal to that power may be obtained.

For similar reasons, the generating function for r_B *is*

$$\frac{\left(\sum_{i=1}^{9} s^i\right)\left(\sum_{i=1}^{6} s^i\right)\left(\sum_{i=1}^{3} s^i\right)}{9 \cdot 6 \cdot 3}$$

and the generating function of r_C *is*

$$\frac{\left(\sum_{i=1}^{8} s^i\right)\left(\sum_{i=1}^{5} s^i\right)\left(\sum_{i=1}^{4} s^i\right)}{8 \cdot 5 \cdot 4}.$$

Inspection of the denominators of these three generating functions, shows 140 possible outcomes for dose level A, 162 outcomes for B, and 160 outcomes for C. The sequence of administration of the dose levels was contrived to make the number of outcomes for each dose level nearly equal, so that the statistical power at each dose level would then be similar. Of course, this may not be desirable in some cases.

In short, both the significance of an overall effect can be determined and the significance of the effect at each dose level, given an overall effect, can be determined. Unfortunately, however, there is no reasonably powerful technique to assess differences between the effects of the different dose levels. At best, the Kruskal-Wallis one-way analysis of variance (Siegel, 1956) may be used to compare differences in the effects of the different levels; the control scores cannot be used without violating random sampling.

Still more statistical sensitivity may be obtained if some results are discounted before the data are collected, just as in a one-tailed hypothesis, it is assumed, *a priori*, that any result not in the predicted direction is a sampling error. Thus, it might be reasonable to suppose that if any effect exists, dose level A (the lowest level) will have the smallest effect and dose level C (the highest level) will have the greatest effect. If any other result may confidently be attributed to chance, the obtained probability levels may be divided by 6 because there are $3! = 6$ possible and equiprobable permutations of the results obtained for the three dose levels, and only one of these can presumably be nonchance. Alternatively, a significant result may also be accepted if A has the largest effect and C had the smallest effect, in which case two permutations may be considered nonchance and the obtained probability level may be divided by 3. In making such an *a priori* decision, the experimenter is taking a considerable risk; if the data seem clearly to contradict his preconceptions, he is in the unenviable position of discarding data only because of the foolishness of his *a priori* notions. On the other hand, if he does accept the unexpected result the true probability of rejection of the null hypothesis at the chance 0.05 level will be 0.30 if only one permutation had been expected and 0.15 if one of two permutations had been expected. The best solution in event of an unexpected outcome may be to repeat the experiment unless the unexpected result is convincing without any formal statistical evidence in its favor.

Several Levels of the Experimental Treatment in Each Subexperiment

The preceding application included nine subexperiments. A variant on this procedure, also utilizing 10 subjects, permits a reduction to three subexperiments as follows: (a) Subexperiment 1. Beginning with 10 subjects, randomly assign one subject to each dose level and utilize seven controls. (b) Subexperiment 2. Of the seven controls of subexperiment 1, randomly assign one subject to each of the dose levels and use the four remaining subjects as controls. (*c*) Subexperiment 3. Repeat the procedure with three experimental subjects and one control. In this design, the probability generating function for each dose level is straightforward, but the assessment of whether an overall effect occurred is difficult. Therefore, the separate dose levels will first be considered.

Consider dose level A. A rank is obtained for each subexperiment by ranking the subject receiving dose level A with respect to the controls and ignoring the results obtained with levels B and C. These ranks are then summed over the three subexperiments. The following probability generating function is applicable:

$$\frac{\left(\sum_{i=1}^{8} s^i\right)\left(\sum_{i=1}^{5} s^i\right)\left(\sum_{i=1}^{2} s^i\right)}{8 \cdot 5 \cdot 2}.$$

Identical probability generating functions apply to levels B and C. Note that the denominator of the generating function shows 80 possible outcomes; when only one experimental subject was run at a time in the otherwise similar design of the preceding section, the smallest number of outcomes was 140, so that this method reduces the number of subexperiments needed in the preceding section at the price of some power.

The next problem is whether the overall pattern is due to chance; obviously the probability that at least one of these statistics will be significant at the 0.05 level has a higher chance level than 0.05, which will be taken, in this discussion, to be the rejection level for the null hypothesis. There are three ways of doing this and the experimenter must select the most reasonable way for his particular experiment before he has seen the data. The first two of these ways are also applicable to the method of the preceding section.

(*a*) If the result is significant at the 0.05 level at the highest dose level, assume any other apparently significant results are real. If it is not, assume any other significant results are spurious.

(*b*) If each of three statistical probabilities were independent, one or more of the three results would be significant at the 0.017 level with a probability of 0.05. Since the results are not entirely independent, because they all depend on the same control scores, a conservative guess at the chance level is 0.02. If one of the three results has a chance probability below 0.02, regard any other results significant at the 0.05 level as not due to chance.

(*c*) Combine all three dose levels for each subexperiment and regard it as the comparison of an experimental with a control group. Then, for each subexperiment, obtain a probability level by some conventional test; the Mann-Whitney *U* test (Siegel, 1956) would be very consistent with our other tests because it is a rank test. Then, combine the three obtained probabilities by means of the z-transformation (Mosteller and Bush, 1954). If, and only if, the combined probability level is below 0.05, there is a significant overall effect. If this method is to be sensitive, it must be reasonable to suppose that all dose levels act in the same direction on the performance. Because U is a discrete distribution, the combined probability will be conservative.

Reversible Effects

So far, cases in which the subjects are irreversibly affected by the experimental treatment have been discussed because it is to these that the new statistical method makes a unique contribution. Nevertheless, an extension in which a subject is used for control data after it has been subjected to the experimental treatment may be of interest to some experimenters, particularly psychopharmacologists.

Suppose there are n subjects. On each of k occasions, one subject is randomly selected for the experimental treatment and the remaining subjects are used as controls. For the foregoing material to be rigorous, it is necessary that the selection be entirely at random, even if it results in the same subject being administered the experimental treatment on each of the k occasions. The probability generating functions for the sum of the ranks obtained by the experimental subjects is

$$\frac{\left(\sum_{i=1}^{n} s^i\right)^k}{n^k}.$$

Irreversible effects will not affect the statistical validity of any rejection of the null hypothesis, although the sensitivity of the test may be reduced, so that it is only necessary that the experimental treatment be reversible enough so that a significant result is conceivable and will make scientific sense.

Now consider a concrete example. There are four subjects, each trained to a high performance criterion. On each of the eight occasions, one of these subjects is randomly selected for drug administration and the remaining three subjects act as controls. The probability generating function looks like this:

$$\frac{(s^1 + s^2 + s^3 + s^4)^8}{4^8}.$$

The denominator of the above function, $4^8 = 65{,}536$, is the number of possible outcomes. This huge number is probably indicative of remarkable sensitivity to small effects.

Because of this large number of outcomes, the probability generating function discussed in the preceding two paragraphs cannot usually be computed except by an electronic computer. Fortunately, both editions of Feller's (1950, 1957) textbook on probability theory include equivalent equations for the chance probability of any sum of ranks under this procedure. For the 1950 edition: examples 11 and 12 on page 236 with necessary background on pages 40-41. For the 1957 edition: examples 18 and 19 on page 266 with necessary background on pages 48-49.

As already mentioned, if the use of the statistics is to be mathematically rigorous, the experimental subject to be used in each subexperiment must be selected entirely at random so that some subjects may receive the experimental treatment more often than others. From an experimental viewpoint, however, it usually seems more desirable to administer the experimental treatment in a restricted random sequence in which no subject receives the treatment a second time until all subjects have received it once. Most likely, doing this reduces the probability of a significant result due to chance.

CAUTION

Because the techniques suggested here are new, there probably are even more pitfalls for the unwary than when conventional statistics

are used; it is unlikely that the author has detected all of them for every potential application. Therefore, if these statistics are used, the experimenter must satisfy himself that there are no such pitfalls in his own particular application and that they are preferable to older statistics. The author's fondest hope is that they may be useful in some small proportion of operant experiments. Certainly no claim is made that operant conditioning has been crippled by a lack of appropriate statistical techniques or that use of the proposed statistics should become standard.

REFERENCES

Cronholm, J. N. and Revusky, S. H. A sensitive rank test for comparing the effects of two treatments on a single group. *Psychometrika*, 1965, 30, 459-467.

Feller, W. *An Introduction to Probability Theory and Its Application. Vol. 1.* New York: John Wiley, First edition, 1950; second edition, 1957.

Mosteller, F. and Bush, R. R. Selected quantitative techniques in G. Lindzey (Ed.) *Handbook of Social Psychology*, Reading, Massachusetts: Addison-Wesley, 1954, pp. 289-334.

Revusky, S. H. Some effects of hunger and frequency of reinforcement on timing. Unpublished doctoral dissertation, Indiana University, 1961.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956.

Received October 11, 1965

APPENDIX: SENSITIVITY OF R_n RELATIVE TO MANN-WHITNEY U

This Appendix supplies a measure of the sensitivity of R_n to nonchance effects and, for comparison, a measure of the sensitivity of the Mann-Whitney U Test (Siegel, 1956). Both U and R_n are rank tests; the difference is that U is a test of a result obtained under a conventional statistical design, while R_n is a test of a result obtained under the present experimental design. Because U is nearly as sensitive to nonchance results as the *t*-test (Siegel, 1956), if R_n is far more sensitive than U, it is also more sensitive than the *t*-test.

Consider the number of control scores when the R_n procedure is used with *n* subjects. In the first subexperiment there are n-1 control scores, and this number is reduced by one in each subsequent subexperiment until in the final subexperiment, there are none. Thus, the total number of control scores in the entire procedure is

$$\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2}.$$

Of these control scores, $R_n - n$ are smaller than their respective experimental scores (from the same subexperiment) because the number of control scores smaller than the experimental score in each subexperiment is one less than its rank. Thus, the proportion of control scores smaller than their respective experimental scores is

$$\frac{2(R_n - n)}{n(n-1)}$$

A proportion of 0.50 is chance expectation. The largest value, smaller than 0.50, which this proportion can have if significance at the one-tailed 0.05 level is obtained, is defined as the sensitivity of R_n for *n* subjects. It is calculated by substituting the value of R_n just significant at the 0.05 level into the above term. Due to the symmetry of the R_n distribution, we need not bother with the case in which the experimental scores usually are larger than the control scores. This measure of sensitivity is also the best estimate of the probability that any randomly selected control score will be smaller than any randomly selected experimental score in the populations from which the data were drawn when significance is just obtained.

For the Mann-Whitney U test (Siegel, 1956), a corresponding measure of sensitivity is

$$\frac{U}{n_1 n_2},$$

where the value of U chosen is just significant at the one-tailed 0.05 level and n_1 and n_2 are the number of subjects in each group. The number of different ways in which one experimental score can be paired with one control score is $n_1 n_2$; U is the number of times the control score is smaller than the experimental score with which it is paired.

Table 7 shows sensitivity as a function of the total number of subjects for R_n and U; when U is used, each group is equally large. It is evident that R_n permits equal sensitivity with perhaps 30% to 50% fewer subjects than

Table 7

A Comparison of the Sensitivity of R_n and U (Mann-Whitney)

	Number of Subjects									
	4	*6*	*8*	*10*	*12*	*16*	*20*	*24*	*32*	*40*
R_n	0	0.133	0.214	0.267	0.303	0.342	0.363	0.377	0.397	0.409
U	–	0	0.062	0.160	0.194	0.234	0.270	0.292	0.324	0.345

U. To use Table 7 as a guide for deciding between the R_n procedure and a two-group procedure for some particular application, the following approximation to the truth is used: if the sensitivity of R_n for some number of subjects is equal to the sensitivity of U for some other number of subjects, both tests are equally likely to detect a nonchance effect (assuming that a larger number of subjects can be tested as carefully as a smaller number). Thus, the experimenter can select the procedure easier for him at some level of sensitivity. Of course, there may be considerations in addition to sensitivity; if there is reason to expect that the effect of concern is a function of the amount of training, the U test may have an advantage, because each experimental subject is tested after the same amount of training.

As previously noted, both the present use of R_n and of U includes controls. Controls may not be necessary in many experimental procedures; if a change in performance occurs after the experimental treatment, it often can be safely attributed to the treatment. In such a case, all the subjects can be administered the experimental treatment and the difference between pre-treatment and post-treatment results can be evaluated by means of the Wilcoxon T test for paired scores (or *t*-test for paired scores). However, R_n is probably more sensitive than T, because for an equal number of subjects, it permits rejection of the null hypothesis at a higher level of significance. R_n's presumed advantage in sensitivity over T is not as great as its advantage over U.