ON THE THEORY OF RUNS WITH SOME APPLICATIONS TO QUALITY CONTROL[1]

By J. Wolfowitz

Columbia University

1. Recent developments in the theory of runs. The increasing number and importance of recent advances in the theory and statistical applications of runs may make a brief paper on the subject of some interest. The large volume of material and its wide dispersal, together with the limitations of space, will of necessity make these remarks far from exhaustive and complete.

I shall not define a run because new advances and applications of new criteria to new problems would probably soon render most definitions obsolete. Runs as used in statistics are best characterized by a philosophy and a technique rather than by the employment of any one specific device. What is always involved is the ordering of observations according to some characteristic and the resultant effect of this ordering on the ordering according to some other characteristic. For example, if the seats at a meeting of statisticians and engineers are numbered and occupied by m engineers and n statisticians, then if we list the numbers of the occupied seats in ascending order and replace each number by E or S according as the seat is occupied by an engineer or statistician, we shall have a sequence of $m + n$ elements, m E's and n S's. Thus, if $m = 7$ and $n = 6$, such a sequence might be

$$E\ E\ E\ S\ E\ E\ S\ S\ S\ E\ S\ S\ E.$$

If we were interested in knowing how well engineers and statisticians are acquainted with one another, we should find it of interest to study the runs of E's and S's in this sequence. Any subsequence of consecutive E's or S's which cannot be enlarged is called a run. Thus in the example above there is a run of E's of length 3, followed in order by a run of S's of length 1, a run of E's of length 2, a run of S's of length 3, a run of E's of length 1, a run of S's of length 2, and a run of E's of length 1. Runs of this kind are usually called runs of two kinds of elements. Naturally the characteristic according to which we order (in the example above, seat number) and the characteristic whose runs are observed (E or S) may be various. They ought in general to have a meaningful connection.

The order of observations has no value if it is known that the observations are independent and random from the same universe and one seeks to estimate a parameter of the universe. Many of the statistical problems treated in the literature are of this character. In quality control of manufactured articles one

[1] Revised from an expository address delivered at a joint meeting of the Institute of Mathematical Statistics and the American Society of Mechanical Engineers at New York, May 29, 1943, at the invitation of the program committee.

of the fundamental problems is to decide whether the observations are "random," or in the language employed in this field, whether statistical control exists. For this purpose indiscriminate pooling of data which suppresses the order characteristics of the observations represents a loss of valuable information.

The algebra of runs of two kinds of elements is fairly elementary and most of the distribution problems involved have been solved. Suppose an urn contains m white balls and n black balls, thoroughly mixed, and $m + n$ drawings are made without replacement. There are $\frac{(m+n)!}{m!n!}$ different sequences of W's and B's possible, and each sequence has the same probability. Let us find in how many ways the m elements W can be arranged to give k runs. By a trick due to Euler, this is the coefficient of x^m in the purely formal expansion of

$$(x + x^2 + \cdots + x^m)^k$$

which is the same as the coefficient of x^m in the formal expansion of

$$(x + x^2 + x^3 + \cdots)^k = \left(\frac{x}{1-x}\right)^k$$

and is therefore $\binom{m-1}{k-1}$ (which is, of course, the combinatorial symbol for $\frac{(m-1)!}{(m-k)!\,(k-1)!}$).

It is easy to see that the number of sequences of W's and B's which have $2k$ runs of both kinds is

$$2\binom{m-1}{k-1}\binom{n-1}{k-1}$$

and hence that the probability that U, the number of runs of both kinds, be $2k$ is

$$2\binom{m-1}{k-1}\binom{n-1}{k-1}\binom{m+n}{m}^{-1}$$

The details of this and other relevant derivations can be found in Wilks [1], Mood [2], Wald and Wolfowitz [3], and Stevens [12]. The formulae given there are of the type given above; e.g., for the probability that $U = c$. Application to tests of significance usually requires formulae of the type which give the probability that $U \leq c$. This causes some difficulty in application and raises a need for suitable tables. Useful tables have been given by Swed and Eisenhart [4] and by P. S. Olmstead in an article by Mosteller [5]. The latter table really deals with a special case of runs of two kinds of elements.

The devices described above were systematically utilized by Mood [2] to give a valuable collection of formulae. A representative result is that the joint distribution of the numbers of runs of length 1, 2, $\cdots$, p and all those of length greater than p is asymptotically normal, with means and covariance matrix given.

The results given by Mood are limited to a classification of runs into a finite number of classes. The author [6] has given a general result which permits weighting runs of all lengths.

Closely allied to runs of two or more kinds of elements are runs from a binomial or multinomial population. If the observations are classified into k classes, designated by $1, 2, \cdots, k$ say, and each observation has a constant probability p_i of falling into the ith class ($i = 1, 2, \cdots, k$) then a sequence of l observations all of which belong to the same class and which is preceded and followed by observations which belong to another class (except, of course, when the sequence is at the beginning or at the end of the series) is called a run of length l. If a coin, whether unbiassed or not, is tossed repeatedly, the runs of heads and tails are runs from a binomial population (i.e., $k = 2$) and if the coin is unbiassed, $p_1 = p_2 = \frac{1}{2}$.

The algebra of these runs has been studied mainly by von Bortkiewicz [7], von Mises [8], Wishart and Hirshfeld [9], Cochran [10], and Mood [2]. Runs from a binomial population (say) differ from runs of two kinds of elements in that m and n (defined above) are chance variables. If therefore, in general, a distribution formula valid for a fixed m and n be multiplied by the probability of this particular set of m and n $\left(\binom{m+n}{m} p_1^m p_2^n\right)$ and summed over m and n, the result will be the corresponding distribution formula for runs from a binomial population. Von Bortkiewicz [7], Cochran [10] and Mood [2] derived the essential parameters involved. Wishart and Hirshfeld [9] proved the asymptotic normality of the total number of runs from a binomial population, and these results were generalized by Mood [2].

Von Mises [8] proved that if N be the number of observations from a binomial population, the distribution of the number of runs of a length which is of the order of log N approaches the Poisson distribution with increasing N.

Cochran [10], extending the work of Gold [11], made use of runs of this kind in order to study what they called "the persistence of weather", i.e., whether dry months tend to follow dry months and wet months to follow wet months. In a long series of weather observations the months were classified as wet or dry and a four-fold table constructed of the number of months falling into each of the following categories:

(a) wet month following a wet month
(b) wet month following a dry month
(c) dry month following a wet month
(d) dry month following a dry month.

The chi-square test was applied to the four-fold table to test the null hypothesis that the probability of whether a month was wet or dry was independent of what its predecessor had been.

Olmstead [13] has made use of a run which is very similar to that of a run from a binomial population, except that the sequence terminates whenever an observation on a specified one of the two classes (a "failure") is recorded. The author

[6] has used a run defined as a sequence of consecutive integers in a permutation of the first n integers to test whether two variates are independently distributed when nothing is known about their distribution functions except that they are continuous. The rank correlation coefficient is usually employed for this purpose.

Of great importance in quality control of manufactured output are runs up and down. If, in any of the $n!$ equally likely (by hypothesis) permutations of the first n integers, we subtract each element from its successor and replace the result by $+$ or $-$ according as the difference is positive or negative, we get runs of $+$ signs and $-$ signs, called respectively runs up and down. The usage of the term length varies; in this paper we shall say that the length of a run is the number of $+$ or $-$ signs in it. This has the advantage that then the sum of the lengths of all the runs is $n - 1$. (Most quality control literature, which follows Shewhart [14] and Kermack and McKendrick [15], defines the length of a run as one more than the number of $+$ or $-$ signs in it.) Thus, for example, the sequence

$$3\ 4\ 7\ 6\ 5\ 1\ 2$$

will appear as

$$+\ +\ -\ -\ -\ +$$

after the $+$ and $-$ signs have been inserted, and has an ascending run of length 2, followed by a descending run of length 3, followed by an ascending run of length 1.

The distributions associated with runs up and down in general present mathematical difficulties greater than those associated with distributions of runs of two kinds of elements and the results are far from complete. The asymptotic expectation of r_p, the number of runs of length p, was given with great brevity by Fisher [16] and in detail by Kermack and McKendrick [15], and the exact result was supplied by Wallis and Moore [17]. The matrix of covariances among the runs of various lengths is being computed, and, it is hoped, will be available for publication shortly. As far as the author is aware, no explicit formulae giving the probability that $r_p = k$ or that $r_p < k$ are known. Some recursion formulae of limited usefulness are available.

The author has recently obtained the asymptotic distributions of r_p, of $r_{p_1}, r_{p_2}, \cdots, r_{p_k}$ jointly, and of related statistics. These are jointly normal. Hence certain quadratic forms in these variables have approximately the chi-square distribution.

Anticipating somewhat the discussion to be given below, it may be mentioned here that the quadratic forms in certain of the r_p which Kermack and McKendrick [15] use to test for randomness, do not have the chi-square distribution which Kermack and McKendrick imply to them. Wallis and Moore [17] first pointed out that these quadratic forms were not the proper chi-square statistics for goodness of fit because of correlation among the r_p. The author's recent results show that these forms do not have the chi-square distribution.

2. Remarks on applications of runs. Let us now turn to statistical applications of some of the runs described above. Suppose we have a sample of m random independent observations on one variate and a similar sample of n observations on another variate. Suppose further that nothing is known a priori about the distribution of each except that both are continuous, and it is desired to test whether the two distributions are identical. This problem is of great practical importance and occurs frequently. In quality control of manufactured output it may occur, for example, if we wish to test whether the output of two machines, two workers, two different processes, or that from raw material obtained from two different sources, is the same. Naturally the problem not only of two, but in general, of a larger number of samples may arise.

The solution proposed in [3] is as follows: Let the $m + n$ observations be arranged in order of, say, ascending size, and let each observation be replaced by F of S according as it comes from the first or second sample. The total number U of runs in both F and S is the statistic to be used. Small values of U are the critical values for rejecting the hypothesis of identity of distributions. Thus in the example above of the seating of statisticians and engineers in the auditorium, a small value of U, which implies that the S (statisticians) and the E (engineers) each tend to bunch together, would be regarded as evidence that the statisticians and engineers present are not well acquainted with one another.

The statistic U seems a not unreasonable one for the purpose. A discrepancy between the two distribution functions will make alternation of values of the two variates less frequent. This idea was proved for large n in [3], where a generalized concept of statistical consistency is given.

On the other hand, the choice of U as a statistic is arbitrary; other reasonable criteria can certainly be given (see, for example, Dixon [19]. In [3] it is shown that a criterion which had previously been proposed was not acceptable because the statistic was not consistent, but nevertheless consistency is a property enjoyed by many statistics and constitutes only a partial check on the arbitrariness of choice. An "abnormally" long run in one or both variates which would be regarded by "common sense" as an indication that the hypothesis ought to be rejected, might be accompanied by a large number of runs of length one which might make the value of U not critically low. Some writers suggest that the presence of a long run of sufficient length be regarded as indicating rejection of the null hypothesis. In that case, if most of the runs were comparatively long, while none were critically long, the null hypothesis would not be rejected under this criterion, but the value of U would be small. A step has been made in the direction of setting-up a criterion for the choice of statistic ([6]) so as to remove this arbitrariness. This involves an extension of the likelihood ratio principle. It must be remembered, however, that almost any criterion will fail to reject some sequence which, it seems intuitively, ought to be rejected. All statistical inference involves risks of error; one object of the science of statistics is to minimize these risks.

Another possible test for the problem of two samples is to compare the num-

bers of runs of various lengths with their expected numbers by the proper chi-square (Caution: the correlation among the variates must be taken into account). The author [6] has developed another test from an extension of the likelihood ratio.

Whenever a uniformly most powerful test does not exist, and this is the case in most non-parametric problems, it is not usually possible to say that one test is more powerful than another, unless the set of alternatives is sufficiently delimited. The power function is then the ultimate criterion for the choice of statistic.

If a sequence of n unequal numbers be given, a very important question is to decide whether the sequence is a "random" one; if it is and the sequence represents measurements on a characteristic of successive products of some manufacturing process, the latter is said to be in statistical control. A precise mathematical formulation can be given to this statement about randomness. Let $X_1, X_2, \cdots, X_n$ be chance variables, and let $x_1, x_2, \cdots, x_n$ be a set of random observations on the corresponding variables. To test whether $x_1, x_2, \cdots, x_n$ is a "random" sequence means to test the hypothesis that $X_1, X_2, \cdots, X_n$ are independently distributed and have identical distribution functions. This is in general a difficult problem, chiefly because of the large class of alternatives to the null hypothesis.

Since the null hypothesis does not specify the distribution functions but only asserts their identity, the tests most generally sought have been such that their size is independent of the unknown (but identical for all the chance variables) distribution function. Certain reasonable procedures have been based on the numbers and lengths of runs up and down in the sequence.

R. A. Fisher [16] suggested doing this, but gave no indication as to what statistic was to be used. Kermack and McKendrick [15] and Wallis and Moore [17] propose the following procedure, the former writers implicitly and the latter explicitly: Let

$$r'_p = \sum_{i=p}^{n-1} r_i$$

and denote by $\bar{x}$ the expectation of the general chance variable x. The proposed statistic is

$$\sum_{i=1}^{p-1} \frac{(r_i - \bar{r}_i)^2}{\bar{r}_i} + \frac{(r'_p - \bar{r}'_p)^2}{\bar{r}'_p}$$

with the critical region the upper tail. Wallis and Moore recommend $p = 3$ and approximate the distribution by empirical methods. As we have seen above, Kermack and McKendrick err in ascribing to the statistic the chi-square distribution.

The criticism has been made by Olmstead [19] that this statistic is insensitive to pronounced trends in the data. This is correct, and had been pointed out earlier in [17], where the prior removal of a trend is recommended. Since one of

the important problems of quality control is detection of a trend, this would limit the usefulness of the statistic for quality control purposes.

It happens frequently when a new rank statistic has been proposed for testing a non-parametric hypothesis such as that of "randomness" above, that critics of the proposed criterion construct sequences which, they say, appealing to "ordinary common sense", any reasonable statistic ought to place in the region of rejection for almost any size of test. They then cheerfully point to the fact that the proposed statistic does not act in this reasonable fashion. A few remarks about this may not be amiss.

A test for, say, "randomness", which is to be made on the sequence of ranks, is really a numbering of the $n!$ permutations of, say, the first n integers, according to the order in which they ought to be taken into the critical region in order to make the latter of any prescribed size. This numbering could even be done by tabulating, for different n, the various sequences in their proper order. Aside from the obvious practical obstacles to such a tabulation, there would soon arise the difficulty that, after the "obvious" sequences are assigned their places the investigator would have difficulty in assigning to most of the remainder an ordering according to the degree in which they may be held to "contradict" the null hypothesis. Resort is therefore made to a statistic which can be given as an analytic expression in the ranks. Because of the inadequacies of the theory the formula is often chosen by analogy with a similar formula in classical statistics. Difficulties may arise because of this.

Let us examine for a moment this intuitive notion of reasonableness. Most people, and even most statisticians, would agree that the sequence of the first n integers in ascending order is an indication of non-randomness. The basis for this notion is an intuitive conception of an alternative to the null hypothesis for which this sequence is very probable. The fact is, however, that if we admit all alternatives to the hypothesis of randomness, for any sequence of ranks whatever there exist infinitely many alternatives which assign to this sequence a probability of one.

It seems to us that the difficulty can be met to a large extent by delimiting the class of distributions which constitute the alternatives to the null hypothesis, and by assigning to the admissible alternatives a weight function which measures the importance of the various alternatives (e.g., the financial loss caused by each). A profound treatment of this subject for the parametric case has been given by Wald [20]. This method has also the great merit that it removes the need for a choice of size of the region of rejection.

In the control of the quality of mass production output one of the outstanding problems is to decide on the basis of a sequence of observations on the product whether the production process is in statistical control. Shewhart and his school of industrial statisticians base many of their tests on the sequence of ranks. On the basis of their experience they find that the causes which most often lead to a breakdown of statistical control are such as to cause shifts up and down in the level of the observations or trends in the observations. To detect

the former they have devised the technique of runs above and below the median and to detect the latter they use runs up and down. Runs above and below the median may be described briefly as follows: The $2m + 1$ (odd number) of observations furnish a sequence of rankings from 1 to $2m + 1$. The elements 1 to m are considered to be elements of one kind and the elements $m + 2$ to $2m + 1$ elements of another kind. We then have a special case of runs of two kinds of elements. Limitations of space prevent the presentation of more detail or a description of the ingenious scheme by which both kinds of runs are graphically exhibited. The reader is referred to [14], [5], and [21], among others. The tests used in the industrial applications are not always explicitly stated, nor do they always seem to be the same. The most common involve comparison of runs of various lengths with their expected number or else are based on the presence of abnormally long runs.

A pretty application of the theory may be found in Campbell [21]. The corrosion of a copper plate was determined by a delicate mechanism which measured the electrical resistance in various places on the plate. The rectangular plate was divided by rows and columns into forty small rectangles in each of which a measurement was made. The readings were made in each column in successive order from one end to the other, and the columns were also measured in successive order from one edge to the other. The observations, when examined for runs above and below the median and runs up and down, indicated something amiss ("absence of statistical control"). Two causes were considered possible:

(a) variations, over the plate, in the corrosion of the copper;

(b) malfunctioning of the delicate measuring apparatus.

The runs obtained by arranging the observations in successive order according to positions on the plate might be expected to be associated with (a), while the runs obtained by arranging the observations in temporal order might be expected to be associated with (b). The object was therefore to separate the two orderings and this was done as follows: The rectangles were numbered 1 to 40 in the order in which the first observations had been made and a random permutation of this sequence was used to indicate the order in which the next set of observations was to be made. The second set was then ordered in two different ways, first according to the temporal order of the observations, and second according to the original ordering by positions. The runs above and below the median and the runs up and down, in the first ordering of the second set of observations gave evidence of a lack of statistical control, while those in the second ordering of the same set did not. An investigation located the trouble in the measuring apparatus.

3. Conclusion. The manifold achievements of quality control as it is practiced at present point to the desirability of still further development of theory and practice. We conclude this paper by suggesting a few directions in which the theory of runs could develop and be of greater assistance in quality control.

(1) The kinds of runs and the statistics used for making decisions in a production process should be chosen on the basis of the kind of deviations from the

state of statistical control which the engineers consider most likely to occur. It is very likely that different production processes may require different statistical procedures.

(2) General distribution theorems should be developed, power functions obtained, and the correlations between different tests investigated.

(3) The application of the weight function idea of minimizing financial losses should be considered.

In these developments both engineers and mathematical statisticians would have important and complementary roles. The tempo of progress will depend in large part on the cooperation between them.

REFERENCES

[1] S. S. WILKS, *Mathematical Statistics*, Princeton, 1943.
[2] A. M. MOOD, *Ann. Math. Stat.*, 11 (1940), p. 367.
[3] A. WALD and J. WOLFOWITZ, *Ann. Math. Stat.*, 11 (1940) p. 147.
[4] FRIEDA S. SWED and C. EISENHART, *Ann. Math. Stat.*, 14 (1943) p. 66.
[5] FREDERICK MOSTELLER, *Ann. Math. Stat.*, 12 (1941) p. 228.
[6] J. WOLFOWITZ, *Ann. Math. Stat.*, 13 (1942) p. 247.
[7] L. VON BORTKIEWICZ, *Die Iterationen*, Berlin, 1917.
[8] RICHARD VON MISES, *Zeitschrift fur angewandte Mathematik und Mechanik*, 1 (1921), p. 298.
[9] J. WISHART and H. O. HIRSHFELD, *Journal of the London Math. Soc.*, 11 (1936), p. 227.
[10] W. G. COCHRAN, *Quarterly Journal of the Roy. Meteorological Society*, 64 (1938) p. 631.
[11] E. GOLD, *Quarterly Journal of the Roy. Meteorological Society*, 55 (1929), p. 307.
[12] W. L. STEVENS, *Annals of Eugenics*, 9 (1939), p. 10.
[13] P. S. OLMSTEAD, *Ann. Math. Stat.*, 11 (1940), p. 363.
[14] W. A. SHEWHART, "Contribution of statistics to the science of engineering", *Proceedings of the Bicentennial Celebration of the Univ. of Penna.*, Philadelphia, 1941.
[15] W. O. KERMACK and A. G. MCKENDRICK, *Proc. Roy. Soc.*, Edinburgh, 57 (1937), pp. 228–240, 332–376.
[16] R. A. FISHER, *Quarterly Journal of the Roy. Meteorological Soc.*, 52 (1926), p. 250.
[17] W. A. WALLIS AND G. H. MOORE, *A Significance Test for Time Series Analysis*, New York, 1941.
[18] P. S. OLMSTEAD, *Journal of the American Stat. Assn.*, 1942, p. 152.
[19] W. J. DIXON, *Ann. Math. Stat.*, 11 (1940), p. 199.
[20] A. WALD, *Ann. Math. Stat.*, 10 (1939), p. 299.
[21] W. E. CAMPBELL, "Use of statistical control in corrosion and contact resistance studies", *Bell Tel. System Tech. Publications*, 1942.