ASYMPTOTIC DISTRIBUTION OF RUNS UP AND DOWN[1]

By J. Wolfowitz

Columbia University

1. Introduction. Let $a_1, a_2, \cdots, a_n$ be any n unequal numbers and let $S = (h_1, h_2, \cdots, h_n)$ be a random permutation of them, with each permutation having the same probability, which is therefore $\frac{1}{n!}$. Let R be the sequence of signs (+ or −) of the differences $h_{i+1} - h_i$ $(i = 1, 2, \cdots, n-1)$. Then R is also a chance variable. A sequence of p successive + (−) signs not immediately preceded or followed by a + (−) sign is called a run up (down) of length p. The term "run" applies to both runs up and runs down. As an example, if $S = (4\ 6\ 2\ 3\ 5)$, then in $R = (+\ -\ +\ +)$ there are three runs, one up of length one, one down of length one, and one up of length two.

The purpose of this paper is to establish several theorems about the limiting distributions of a class of functions of runs up and down. These results are applicable to certain techniques which have been employed in quality control and the analysis of economic time series. They are also shown to apply to a large class of "runs."

2. Joint distribution of runs of several lengths. Let r_p be the number of runs of length p in R and r'_p the number of runs of length p or more in R. Then r_p and r'_p are chance variables. The expectations $E(r_p)$ and $E(r'_p)$, the variances $\sigma^2(r_p)$ and $\sigma^2(r'_p)$, and the covariances $\sigma(r_{p_1} r_{p_2})$ are given by Levene and Wolfowitz [1]. They are all of the order n. Let

$$y_p = \frac{r_p - E(r_p)}{\sqrt{n}},$$

$$y'_p = \frac{r'_p - E(r'_p)}{\sqrt{n}}.$$

Our first results are embodied in the following theorem:

Theorem 1. *Let l be any non-negative integer. The joint distribution of $y_1, \cdots, y_l, y'_{(l+1)}$, approaches the normal distribution as $n \to \infty$.*

We shall give the proof for the case $l = 1$, but it will easily be seen to be perfectly general.

Let $x_{pi} = 1$ if the sign (+ or −) of $h_{i+1} - h_i$ is the initial sign of a run of length p, and let $x_{pi} = 0$ otherwise. Let $w_{pi} = 1$ if the sign of $h_{i+1} - h_i$ is the

[1] Part of the results of this paper was presented to the Institute of Mathematical Statistics and the American Mathematical Society at their joint meeting in New Brunswick, N. J., on September 13, 1943.

initial sign of a run of length p or more, and let $w_{pi} = 0$ otherwise. Let $x_{pn} = w_{pn} = 0$. Then

$$r_1 = \sum_{i=1}^{n} x_{1i},$$

$$r_2' = \sum_{i=1}^{n} w_{2i}.$$

Now write $\alpha = n^{\frac{2}{3}}$, $\beta = n^{\frac{1}{3}}$, and consider the β sequences

$$h_{(j-1)\alpha+1}, \ h_{(j-1)\alpha+2}, \ \cdots, \ h_{j\alpha} \qquad (j = 1, 2, \cdots, \beta).$$

(Strictly speaking, we should employ the largest integer in α. Since what is meant is clear and since we are dealing with an asymptotic property, we shall omit this useless nicety.) Let x'_{pi} and w'_{pi} have the same definitions relative to each of these sequences that x_{pi} and w_{pi} have relative to the sequence S. The accented and unaccented x's and w's are not always the same, because the partitioning of the sequence S sometimes breaks up runs and creates others. Thus we might have $x_{p\alpha} = 1$, but $x'_{p\alpha}$ always $= 0$.

It is easy to see that there exists a positive number d such that

$$\sum_{i=1}^{n} | x_{1i} - x'_{1i} | < d\beta,$$

$$\sum_{i=1}^{n} | w_{2i} - w'_{2i} | < d\beta.$$

If, therefore, we define

$$z_1 = \frac{\sum_{i=1}^{n} [x'_{1i} - E(x'_{1i})]}{\sqrt{n}},$$

$$z_2' = \frac{\sum_{i=1}^{n} [w'_{2i} - E(w'_{2i})]}{\sqrt{n}},$$

we have

$$| z_1 - y_1 | < \frac{2\,d\beta}{\sqrt{n}}$$

$$| z_2' - y_2' | < \frac{2\,d\beta}{\sqrt{n}}$$

and

$$\frac{d\beta}{\sqrt{n}} \to 0.$$

Hence, if the joint limiting distribution of z_1 and z_2' is normal, so is that of y_1 and y_2'.

The chance variables

$$r_{1j} = \sum_{i=(j-1)\alpha+1}^{j\alpha} x_{1i}'$$

$$r_{2j}' = \sum_{i=(j-1)\alpha+1}^{j\alpha} w_{2i}' \qquad (j = 1, 2, \cdots, \beta)$$

have the same joint distribution for all values of j. For x_{1i}' and w_{2i}', $((j-1)\alpha < i < j\alpha)$, depend only on the relative magnitude of the elements of the sequence

$$h_{(j-1)\alpha+1}, \cdots, h_{j\alpha},$$

not upon the particular values which the elements take, and all permutations of the sequence have equal probability. Clearly r_{1j} and r_{2j}' are independent, in the probability sense, of $r_{1j'}$ and $r_{2j'}'$ $(j \neq j')$, because of the definitions of x_{1i}' and w_{2i}'. (However, r_{1j} and r_{2j}' are not independent, because x_{1i}' and w_{2i}' cannot both be 1.) From the results of [1] it follows that for sufficiently large n the absolute value of the correlation coefficient between r_{1j} and r_{2j}' is less than a number smaller than 1. By the methods of [1] it can easily be shown that the ratio of the fourth order moments of r_{1j} and r_{2j}' about their means to the square of the variance of either, is bounded for sufficiently large n. Hence by Liapounoff's theorem (see, for example, Cramer [2], Uspensky [5]), z_1 and z_2' are jointly normally distributed in the limit. Hence so are y_1 and y_2' and the theorem is proved.

3. Generalization of Theorem 1. Examination of the proof of Theorem 1 shows that it rests on the following two properties of runs up and down:

a) Partition of the sequence S into subsequences affects at most d runs in each sub-sequence, where d is a fixed positive number independent of n.

b) After partition the totals of runs of each length in any sub-sequence (the definition now relates to the subsequence) are independent in the probability sense of the totals of runs in any other subsequence, and satisfy some condition (such as the Liapounoff) sufficient to make the components of the sum of the vectors jointly normally distributed in the limit.

Hence if we adopt other definitions of runs which meet conditions (a) and (b) above, the total numbers of each of these various kinds of runs will be in general jointly asymptotically normally distributed. For example, if s_p and s_p' be the numbers of runs *up* of length p and of length p or more, respectively, and if t_p and t_p' are the same quantities referring to runs *down*, then, with l and k any positive integers,

$$s_1, s_2, \cdots, s_l, \qquad s_{(l+1)}', \qquad t_1, t_2, \cdots, t_k$$

are jointly asymptotically normally distributed. However, if $t'_{(k+1)}$ is included in this set, since

$$s'_1 = s_1 + s_2 + \cdots + s_l + s'_{(l+1)}$$

and

$$t'_1 = t_1 + t_2 + \cdots + t_k + t'_{(k+1)}$$

differ by at most one, the limiting distribution is degenerate, i.e., its covariance matrix is only semi-definite.

As another example, if we define a bizarre run as, say, the occurrence of a run up of length 5, followed, 17 elements later, by a run down of length 14, then the number of runs of this type is asymptotically normally distributed with expectation and variance of order n.

4. Additive functions of runs of all lengths. Combining the numbers of runs of all lengths greater than a given length generally involves a loss of information. The following theorem on additive functions of runs up and down may be of general interest and of utility in avoiding this undesirable situation.

THEOREM 2. *Let $f(i)$ be a function, defined for all positive integral values of i, which fulfills the following conditions:*

a) *There exists a pair of positive integers, a and b, such that*

$$\frac{f(a)}{f(b)} \neq \frac{a}{b} \tag{4.1}$$

b) *for any $\epsilon_1 > 0$ there exists a positive integer $N(\epsilon_1)$ such that, for all $n \geq N(\epsilon_1)$,*

$$\sum_{i=N(\epsilon_1)}^{i=n-1} |f(i)| \sigma(r_i) < \epsilon_1 n \tag{4.2}$$

where n, of course, has the same meaning as in the preceding sections. Let $F(S)$, a function of the chance sequence S, be defined as follows:

$$F(S) = \sum_{i=1}^{(n-1)} f(i) r_i . \tag{4.3}$$

Then the distribution of $\dfrac{F(S) - E[F(S)]}{\sigma[F(S)]}$ approaches the normal distribution as $n \to \infty$.

As an example, let $f(i) \equiv 1$. Then $F(S) \equiv r'_1$, whose limiting distribution is normal by Theorem 1.

This theorem is the exact analogue of Theorem 2 of [3] and the proof of the latter carries over without difficult changes except in one important respect. A difficulty in the proof of the theorem in [3] lay in proving Lemma 4, and this lemma has to be proved completely anew. We shall limit ourselves here to doing just that. Lemma 2 of Theorem 2 of [3], whose only role was to help in proving Lemma 4, has no analogue in our present problem, but all the others do. It will therefore be sufficient if we prove the following:

LEMMA. *There exists a constant* $c > 0$, *such that, for all* n *sufficiently large,*

$$\sigma^2[F(S)] > cn. \tag{4.4}$$

Condition (a) of the theorem is imposed simply in order that the result be not trivial. For, if (a) does not hold, we have that

$$f(i) \equiv if(1),$$

and

$$\begin{aligned} F(S) &\equiv f(1) \Sigma ir_i \\ &\equiv (n-1)f(1) = \text{a constant.} \end{aligned}$$

Suppose that

$$f(i) \equiv ui + v,$$

with u and v constants, and $v \neq 0$. Then by Theorem 1

$$\begin{aligned} F(S) &= u(n-1) + vr_1' \\ &= vr_1' + \text{a constant} \end{aligned}$$

is asymptotically normally distributed with variance of order n. Without loss of generality we may therefore assume that

$$f(i) \not\equiv ui + v. \tag{4.5}$$

From this it follows that there exists an integer $A \geq 2$ such that

$$f(A-1) + f(A+1) \neq 2f(A). \tag{4.6}$$

Our object is to prove that $\sigma^2[F(S)]$ is at least of order n. The basic idea of the proof will be to construct two sets, say L_1 and L_2, of sequences S, such that the (same) probability of each is not less than a positive lower bound independent of n, and such that there exists a one-to-one correspondence between the sequences of L_1 and those of L_2 so that, if S_1 is a member of L_1 and S_2 the corresponding sequence in L_2,

$$|F(S_1) - F(S_2)| \geq g\sqrt{n},$$

where g is a positive constant independent of n. It is easy to see that such a construction would prove the lemma.

We shall call the subsequence $(h_i, h_{i+1}, \cdots, h_{i+2A})$ of S, a run of type T_1 or simply a run T_1 (the notion will be used only for the proof of this lemma) if the following conditions are fulfilled:

(4.7) each of the signs of $(h_{i+1} - h_i)$ and $(h_{i+A+1} - h_{i+A})$ is the initial sign of a run of length A.

(4.8) if $i \neq 1$, the sign of $(h_i - h_{i-1})$ is not the final sign of a run of length A.

(4.9) if $i + 2A \neq n$, the sign of $(h_{i+2A+1} - h_{i+2A})$ is not the initial sign of a run of length A.

(4.10) after the transformation H, which interchanges h_{i+A-1} and h_{i+A}, has operated on the run, the sign of $(h_{i+1} - h_i)$ is the initial sign of a run of length $A - 1$, and the sign of $(h_{i+A-1} - h_{i+A})$, in the new ordering, is the initial sign of a run of length $A + 1$.

Thus, with $A = 2$ and $n = 7$, if $S = (7145326)$, then $R = (- + + - - +)$, and (1 4 5 3 2) is a run T_1, for after the transformation H has been applied we have (1 5 4 3 2) which gives $(+ - - -)$. The result of the operation H on a run T_1 will be called a run T_2.

The number r^* of runs T_1 and the number r^{**} of runs T_2 each have expected values and variances of order n, by considerations similar to those of [1]. Hence, for an arbitrarily small positive ϵ there exists a positive constant q such that, for all n sufficiently large, the probability $P\{r^* + r^{**} \geq qn\}$ of the set L^* of sequences S which satisfy the relation in braces, is not less than $1 - \epsilon$.

The set L^* can be divided into disjunct sets (families) as follows: Let $S(0)$ be any sequence S in L^* which has no runs T_2 (any doubt about the existence of such sequences will be soon removed) and let $r^*[S(0)] = m$. Hence $m \geq qn$. Operating with the transformation H on each of the m runs T_1 of $S(0)$ we get a set $S(1)$ of m different sequences for each of which $r^* = m - 1, r^{**} = 1$. Operating again with H on each of the pairs of runs T_1 of the sequence $S(0)$ we get a set $S(2)$ of $\binom{m}{2}$ distinct sequences for each of which $r^* = m - 2$, $r^{**} = 2$, etc. The process stops with $S(m)$, which contains a single sequence, for which $r^* = 0$, $r^{**} = m$. The set $S(i)$ contains $\binom{m}{i}$ different sequences for each of which $r^* = m - i, r^{**} = i$. The union of the sets $S(i)$ $(i = 1, 2, \cdots, m)$ will be called the family whose generator is $S(0)$. The sets $S(i)$ are obviously disjunct. Any sequence S in L^* belongs to one and only one family. For if we operate on *all* of its runs T_2 with H (which is its own inverse), we obtain the generator of the family to which it belongs. This also proves the existence of sequences in L^* for which $r^{**} = 0$.

Consider any family F whose generator is a sequence for which $r^* = m \geq qn$. It is easy to see that, when n is sufficiently large, the ratio of the total number of sequences S in the sets L_1^* and L_2^*, where

$$L_1^* = \sum_{i=0}^{i=\frac{1}{2}(m-\sqrt{m})} S(i),$$

and

$$L_2^* = \sum_{i=\frac{1}{2}(m+\sqrt{m})}^{i=m} S(i),$$

to the total number of sequences in F is greater than a fixed positive constant K'.

We are now ready to construct L_1 and L_2. The set L_1 is the union of the sets L_1^* of all the families in L^*, and the set L_2 is the union of the sets L_2^* of all the families in L^*. The probability of L_1 and of L_2 is therefore not less than $\frac{1}{2}K'(1 - \epsilon)$. The one-to-one correspondence is effected as follows: The subset $S\left(\frac{m}{2} - \frac{\sqrt{m}}{2} - j\right)$ of the set L_1^* of any family is to correspond to the

subset $S\left(\frac{m}{2}+\frac{\sqrt{m}}{2}+j\right)\left(j=0,1,2,\cdots,\frac{m}{2}-\frac{\sqrt{m}}{2}\right)$ of the set L_2^* of the same family. The individual sequences of either of the two subsets may be made to correspond to those of the other in any manner whatsoever. Any sequence S_1 in L_1 and its corresponding sequence S_2 in L_2 thus differ only in the numbers of runs T_1 and T_2, but are identical in the numbers of all other runs. They differ in at least $\sqrt{m}$ runs. Hence,

$$\begin{aligned}|\,F(S_1)-F(S_2)\,| &\geq \sqrt{m}\,|\,2f(A)-f(A-1)-f(A+1)\,|\\ &\geq \sqrt{qn}\,|\,2f(A)-f(A-1)-f(A+1)\,|\,.\end{aligned}$$

This is the required result with

$$g=\sqrt{q}\,|\,2f(A)-f(A-1)-f(A+1)\,|\,.$$

Hence the lemma and the theorem are proved.

The remarks of section 3 also apply to Theorem 2.

5. The distribution of long runs. Certain tests in use in quality control of manufactured products are based on the occurrence of long runs. Since the mean and variance of r_p, for any fixed p, are of order n, it follows that the probability that $r_p \neq 0$ approaches 1 (with increasing n). In order to base a test on the occurrence of a run of length p in long sequences it is therefore necessary to make p a function of n. This function must be a suitable one, because if p is, for example, of the order n, the probability that $r_p = 0$ approaches 1; p should, therefore, be neither too short nor too long.

The following theorem will help give the answer to this problem:

THEOREM 3. *Let p vary with n, so that*

$$\frac{(p+1)!}{n}=\frac{1}{K}$$

with K a fixed positive number. Then

$$\lim_{n\to\infty} P\{r_p=j\}=e^{-2K}\frac{(2K)^j}{j!} \qquad (j=0,1,2,\cdots)$$

i.e., *r_p has in the limit the Poisson distribution with mean $2K$.*

The proof will consist in showing that the moments of r_p approach the moments of a Poisson distribution with mean $2K$ as $n\to\infty$. This is sufficient (v. Mises [4]).

Let $x_i = 1$ if the sign of $h_{i+1}-h_i$ is the initial sign of a run of length p, and $x_i = 0$ otherwise. The probability that $x_i = 1$ is, by [1], Section [4], $\frac{2(p^2+3p+1)}{(p+3)!}$ for all i with a fixed number of exceptions.[2] Write $B=\frac{2}{(p+1)!}$; then

$$P\{x_i=1\}=B+o(B),$$

[2] Since these exceptions (at the ends of the sequence S) have no effect on the asymptotic theory, they will henceforth be ignored.

where the symbol $o(B)$ means that $\lim \frac{o(B)}{B} = 0$. Let y_i $(i = 1, 2, \cdots, n)$ be independent chance variables with the same distribution: $P\{y_i = 1\} = B$, $P\{y_i = 0\} = 1 - B$. Then it is easy to see that $Y = \sum_{i=1}^{n} y_i$ has in the limit the Poisson distribution with mean $2K$ and that its moments approach the moments of the same Poisson distribution. Hence it will be sufficient to show that in the limit Y and r_p have the same moments.

If $q, \alpha_1, \alpha_2, \cdots, \alpha_q$ and $i_1 < i_2 < \cdots < i_q$ are positive integers, we have that

$$E(y_{i_1}^{\alpha_1} y_{i_2}^{\alpha_2} \cdots y_{i_q}^{\alpha_q}) = E(y_{i_1} y_{i_2} \cdots y_{i_q}) = \prod_{j=1}^{q} E(y_{i_j}) = B^q \tag{5.1}$$

and

$$0 \leq E(x_{i_1}^{\alpha_1} x_{i_2}^{\alpha_2} \cdots x_{i_q}^{\alpha_q}) = E(x_{i_1} x_{i_2} \cdots x_{i_q}). \tag{5.2}$$

Also

$$E(r_p^l) = E\left[\sum_{i=1}^{n} x_i\right]^l. \tag{5.3}$$

After expansion of the right member of (5.3), we may replace, in accord with (5.2), each of the non-zero exponents of the x's by 1. The same operation on the terms of the expansion of the right member of

$$E(Y^l) = E\left[\sum_{i=1}^{n} y_i\right]^l, \tag{5.4}$$

is valid in accord with (5.1).

Let $i_1 < i_2 < \cdots < i_q$. In the expression

$$E(x_{i_1} x_{i_2} \cdots x_{i_q}), \tag{5.5}$$

let q be the "weight." A subsequence of consecutive x's in (5.5) (it may consist of a single x) which is such that the indices of two consecutive x's differ by less than $(p + 3)$, while the subsequence cannot be expanded on either side without violating this requirement, will be called a "cycle." Let c be the number of cycles in (5.5). By [1], Section 4, if x_i and x_j are in different cycles, i.e., $|i - j| \geq (p + 3)$, then x_i and x_j are independently distributed. If, therefore, $q = c$, we have that

$$E(x_{i_1} x_{i_2} \cdots x_{i_q}) = \prod_{j=1}^{q} E(x_{i_j}) = B^q + o(B^q). \tag{5.6}$$

If $q > c = 1$, we have, also from [1], Section 4, that

$$E(x_{i_1} x_{i_2} \cdots x_{i_q}) \leq E(x_{i_1} x_{i_2}) = o(B). \tag{5.7}$$

If $q > c$ and if there are two indices in the expression (5.5) which differ by less than p, then

$$E(x_{i_1}x_{i_2} \cdots x_{i_q}) = 0. \tag{5.8}$$

For x_i and x_j cannot both initiate runs of length p if $| i - j | < p$.

Let us now return to the expansions of the right members of (5.3) and (5.4), in which the exponents have been replaced as described before. Let the weight and cycle definitions also apply to terms of the type

$$E(y_{i_1}y_{i_2} \cdots y_{i_q}). \tag{5.9}$$

From (5.1) and (5.6) it follows that, in the limit, the contributions to $E(r_p^l)$ and $E(Y^l)$ of the sums of those terms for which $q = c$, are the same. Let W and W' be the sums of all the remaining terms in $E(r_p^l)$ and $E(Y^l)$, respectively. If we can show that

$$\lim W = \lim W' = 0 \tag{5.10}$$

we will have proven that

$$\lim E(r_p^l) = \lim E(Y^l) \tag{5.11}$$

and with it the theorem.

Let $B = O[f(n)]$ mean, as usual, that $| B | \leq Mf(n)$ for all n and a fixed $M > 0$. The number of terms in W' with fixed q and c ($c < q$, by definition of W') is $O(n^c p^{q-c})$. From (5.1) the value of the sum of all such terms is $O(B^q n^c p^{q-c})$. Now

$$nB = O(1)$$

by the hypothesis of the theorem. From the definition of p,

$$p = o(n)$$

and hence

$$pB = o(1).$$

Therefore

$$\begin{aligned} B^q n^c p^{q-c} &= (nB)^c (pB)^{q-c} \\ &= o(1). \end{aligned}$$

Since $q \leq l$, there are only a fixed number of such sums. Hence $\lim W' = 0$.

The number of terms in W with fixed q and c ($c < q$) is $O(n^c p^{q-c})$. However, most of these are of the type in (5.8) and therefore vanish. Those which do not vanish are $O(n^c)$ in number. Since $q > c$ we have by application of (5.7) that each term is $o(B^c)$. Hence the value of the sum of these terms is $o(n^c B^c) = o(1)$. Since $q \leq l$, there are a fixed number of such sums. Hence $\lim W = 0$.

This proves (5.10) and with it the theorem.

It is possible to generalize this result in a manner similar to that of Section 3.

The author is obliged to W. Allen Wallis who first drew his attention to problems in runs up and down, and to Howard Levene, who read the manuscript of this paper.

REFERENCES

[1] H. LEVENE and J. WOLFOWITZ, *Annals of Math. Stat.*, Vol. 15 (1944).
[2] HARALD CRAMÉR, *Random Variables and Probability Distributions*, Cambridge, 1937.
[3] J. WOLFOWITZ, *Annals of Math. Stat.*, Vol. 13 (1942), p. 247.
[4] R. v. MISES, *Zeitschrift fuer die angewandte Math. und Mechanik*, Vol. 1 (1921), p. 298.
[5] J. V. USPENSKY, *Introduction to Mathematical Probability*, New York, 1937.