THEORY OF RUNS

by

SHASHI N. SHARMA

B. S., Michigan State University, 1963

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

Approved by:

Major Professor

TABLE OF CONTENTS

1. INTRODUCTION

In studying a particular sample, the order in which the elements of the sample are drawn frequently is available. One reasonable mathematical method for handling this information is to make use of the distribution of runs. A run is defined as a succession of similar events preceded and succeeded by different events or no events. The number of elements in a run will be referred to as the length of the run. The total number of runs, the length of the longest run, and various other run statistics can be used as the sample information with which to test for randomness of arrangement against the alternative of sequence dependency. Also the above run statistics can be used to test whether two sampled populations are identical, whether trend exists in sampled population, and so forth.

The distribution theory of runs seems to have been started toward the end of the nineteenth century. The following history is given by Mood (1940). Karl Pearson (1897) regarded the distribution of runs as a special case of the multinomial distribution. An expression for the mean of the number of iterations of a given length from a binomial population was derived by Karl Marbe (1899). Grunbaum (1904) and Bruns (1906) derived the mean of the number of runs of given length from a binomial population by the multinomial method. In a book published in 1907, Von Borkviewicz correctly derived for the first time the mean and variance of runs from a binomial population using a method similar to that of Bruns. Von Mises (1921) showed the number of long runs of a given length was approximately distributed according to the Poisson law for large sample sizes. It was not until 1925 that an actual distribution function appeared when Ising (1925) gave the number of ways of obtaining a given total number of runs (without regard to length) from arrangements of two kinds of elements.

Stevens (1939) published the same distribution and described an X^2 criterion for significance. Wald and Wolfowitz (1940) published the same distribution and showed that it was asymptotically normal. Wald and Wolfowitz's paper described a very interesting application of the distribution to the problem of testing the hypothesis that two samples have come from the same continuous distribution. A. M. Mood (1940) in an interesting paper, investigated different problems concerning runs. He derived distributions of runs of given length from random arrangements of fixed numbers of elements of two or more kinds, and from binomial and multinomial populations. Wolfowitz (1944a) derived a distribution of runs up and down and also the asymptotic distribution of runs up and down. Levene and Wolfowitz (1944) obtained the covariance matrix of runs up and down and gave several methods for using runs up and down in tests of significance. Swed and Eisenhart (1943) derived the tables for testing randomness of grouping in a sequence of alternatives. The asymptotic properties of the Wald-Wolfowitz test of randomness were derived by Noether (1950). In later years Kruskal (1952), Mood (1954), Dixon (1954), Goodman (1957), Ferguson and Kraft (1955), and Weiss (1960) developed runs tests.

In this report the distribution theory of runs is developed in section 2. In section 3, runs tests are given. The Wald-Wolfowitz total number of runs test is discussed explicitly in section 3. Runs up and down are given in section 4. Total number of runs up and down and chi-square applied to run frequencies are also discussed in the same section. At the end in section 5 applications of runs tests are given.

2. DISTRIBUTION THEORY OF RUNS

2.1 Runs of Two Kinds of Elements

Consider a sample space constructed using n elements of two kinds, n_1 a's and n_2 b's with $n_1 + n_2 = n$. Any particular sample point is a sequence of a's and b's which consists of alternating runs of a's and b's. For each sample point let r_{1i} denote the number of runs of a's of length i and let r_{2i} denote the number of runs of b's of length i. Thus the sequence

aaabbaabaabbab

has $r_{11} = 1$, $r_{12} = 2$, $r_{13} = 1$, $r_{21} = 1$, $r_{22} = 2$, and all other r_{ij}'s are zero. Let $r_1 = \sum_i r_{1i}$ and $r_2 = \sum r_{2i}$ denote the total number of runs of a's and b's respectively.

Suppose a set of numbers $r_{ij}(i = 1, 2;\ j = 1, 2, \ldots, n_i)$ such that $\sum_j j\, r_{ij} = n_i$ is given. Then the numbers of ways of arranging the r_1 runs of a's and the r_2 runs of b's are

$\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix}^{*1}$ and $\begin{bmatrix} r_2 \\ r_{2j} \end{bmatrix}$ respectively. Thus the total number of ways of obtaining the set $\{r_{ij}\}$ is

$$N(r_{ij}) = \begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \begin{bmatrix} r_2 \\ r_{2j} \end{bmatrix} F(r_1 r_2) \qquad (2.1.1)$$

where $F(r_1 r_2)$ is the number of ways of arranging r_1 objects of one kind and r_2 objects of a second kind so that no two adjacent objects are of the same kind. Then

$$F(r_1 r_2) = \begin{bmatrix} 0 \ , & |r_1 - r_2| > 1 \\ 1 \ , & |r_1 - r_2| = 1 \\ 2 \ , & r_1 = r_2 \end{bmatrix} \qquad (2.1.2)$$

But there are $\binom{n}{n_1}^{*2}$ possible arrangements of a's and b's. If each of these arrangements is equally likely, then the probability of obtaining the set $\{r_{ij}\}$ is

$$P(r_{ij}) = \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \begin{bmatrix} r_2 \\ r_{2j} \end{bmatrix} F(r_1, r_2)}{\binom{n}{n_1}} \qquad (2.1.3)$$

The probability distributions of the set $\{r_{1j}\}$ and $\{r_1\}$ will be derived now. By summing $\begin{bmatrix} r_2 \\ r_{2j} \end{bmatrix}$ over all partitions of n_2, the

---

*1 $\begin{bmatrix} m \\ m_i \end{bmatrix} = \dfrac{m!}{m_1! m_2! \ldots m_s!}$ denotes the multinomial coefficient. When the multinomial coefficient is to be summed over the indices m_i, the following conditions

$$\sum m_i = m, \ m_i \geqslant 0$$

will always hold.

*2 $\binom{m}{k} = \dfrac{m(m-1) \ldots (m-k+1)}{k!}$ denotes the binomial coefficient.

probability distribution of the set r_{1j} can be obtained. The summation over all partitions of n_2 is aided by first finding the coefficient of x^{n_2} in

$$(x + x^2 + x^3 + \dots)^{r_2} = x^{r_2}(1 + x + x^2 + \dots)^{r_2}$$

$$= \frac{x^{r_2}}{(1 - x)^{r_2}} \qquad |x| < 1$$

$$= x^{r_2} \sum_{t=0}^{\infty} \binom{r_2 - 1 + t}{r_2 - 1} x^t$$

The term corresponding to $t = n_2 - r_2$ gives the desired result, so we may write the coefficient of x^{n_2} as $\binom{n_2 - 1}{r_2 - 1}$. Thus the desired summation over all partitions of n_2 is

$$\sum_{\sum_j r_{2j} = n_2} \begin{bmatrix} r_2 \\ r_{2j} \end{bmatrix} = \binom{n_2 - 1}{r_2 - 1} \tag{2.1.4}$$

Then

$$P(r_{1j}, r_2) = \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_2 - 1}{r_2 - 1} F(r_1 r_2)}{\binom{n}{n_1}} \tag{2.1.5}$$

Summing equation (2.1.5) over r_2, and simplifying,

$$P(r_{1j}) = \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_2 + 1}{r_1}}{\binom{n}{n_1}} \tag{2.1.6}$$

Summing (2.1.3) over r_{1j} and r_2, gives by means of (2.1.4)

$$P(r_1, r_2) = \frac{\binom{n_1 - 1}{r_1 - 1} \binom{n_2 - 1}{r_2 - 1} F(r_1, r_2)}{\binom{n}{n_1}} \tag{2.1.7}$$

The distribution (2.1.7) was derived by Wald and Wolfowitz in 1939.

The probability function of r_1, the total number of runs of a's, is obtained by summing (2.1.7) over r_2.

$$P(r_1) = \frac{\binom{n_1 - 1}{r_1 - 1}\binom{n_2 + 1}{r_1}}{\binom{n}{n_1}} \tag{2.1.8}$$

This distribution is discussed by Stevens (1939). A similar expression holds for the probability function of r_2.

Mood (1940) derived several distributions useful for applications. He added together long runs to form new variables, decreasing the number of variables compared with (2.1.3) and (2.1.6).

One of the marginal distributions is obtained by summing (2.1.6) over r_{1i} for $i \geqslant k$. Letting

$$s_{1j} = r_{1j}\ , \quad j < k$$

$$s_1 \ = r_1$$

$$s_{1k} = \sum_{k}^{n_1} r_{1j}\ , \quad A = \sum_{1}^{k-1} j\, r_{1j}$$

The multinomial coefficient

$$\frac{s_{1k}!}{r_{1k}! \cdots r_{1n_1}!}$$

must be summed over all partitions of n_1 - A such that every part is greater than k - 1. This can be obtained by the coefficient of $x^{n_1 - A}$ in

$$(x^k + x^{k+1} + \ldots)^{s_{1k}} = x^{ks_{1k}} \sum_{t=0}^{\infty} \binom{s_{1k} - 1 + t}{s_{1k} - 1} x^t$$

which gives

$$\sum_{(k)} \frac{s_{1k}!}{r_{1k}! \ldots r_{1n_1}!} = \binom{n_1 - A - (k-1)s_{1k} - 1}{s_{1k} - 1} \qquad (2.1.9)$$

where $\sum_{(k)}$ denotes summation over all positive integers r_{1k}, $r_{1k+1}, \ldots, r_{1n_1}$ such that $\sum_{k}^{n_1} j\, r_{1j} = n_1 - A$. This identity with (2.1.6) gives

$$P(s_{1i}) = \frac{\begin{bmatrix} s_1 \\ s_{1i} \end{bmatrix} \binom{n_2+1}{s_1} \binom{n_1-A-(k-1)s_{1k}-1}{s_{1k} - 1}}{\binom{n}{n_1}}, \quad i = 1,2,\ldots k \qquad (2.1.10)$$

Another marginal distribution can be derived by considering runs of both kinds of elements. Defining $s_{2j}(j = 1, 2, \ldots h)$ and B in terms of r_{2j} just as s_{1i} and A were defined before, it follows from (2.1.3) and (2.1.10) that

$$P(s_{1i}, s_{2j}) = \frac{\begin{bmatrix} s_1 \\ s_{1i} \end{bmatrix} \begin{bmatrix} s_2 \\ s_{2j} \end{bmatrix} \binom{n_1 - A - (k-1)s_{1k} - 1}{s_{1k} - 1} \times \binom{n_2 - B - (h-1)s_{2h} - 1}{s_{2h} - 1} F(s_1 s_2)}{\binom{n}{n_1}}$$

$$i = 1, 2, \ldots, k$$

$$j = 1, 2, \ldots, h$$

Here k and h in new variables s_{1k} and s_{2h} can be chosen so that the number of variables is appropriate for data in hand.

2.2 Moments of Runs r_1, r_2 of Two Kinds of Elements

Instead of dealing with ordinary moments, the easiest way to find moments of r_1 is given by means of factorial moments because the expressions are much more compact. For the g^{th} factorial moment $\mu'_{[g]}$, we get using (2.1.8)

$$\mu'_{[g]} = E(r_1^{[g]}) = \sum_{r_1=g}^{n_1} \frac{r_1^{[g]}\binom{n_1 - 1}{r_1 - 1}\binom{n_2 + 1}{r_1}}{\binom{n}{n_1}} \tag{2.2.1}$$

which can be written as

$$\mu'_{[g]} = \frac{(n_2 + 1)^{[g]}}{\binom{n}{n_1}} \sum_{r_1=g}^{n_1} \binom{n_1 - 1}{r_1 - 1}\binom{n_2 + 1 - g}{r_1 - g} \tag{2.2.2}$$

But it follows from the identity

$$\sum_{i=0}^{B} \binom{A}{C + i}\binom{B}{i} = \binom{A + B}{C + B} \tag{2.2.3}$$

that (2.2.2) can be written as

$$\mu'_{[g]} = \frac{(n_2 + 1)^{[g]}\binom{n - g}{n_1 - g}}{\binom{n}{n_1}} \tag{2.2.4}$$

from which the mean and the variance of r_1 can be found to be:

$$\mu(r_1) = \frac{n_1(n_2 + 1)}{n}, \quad \sigma^2(r_1) = \frac{(n_2+1)^{[2]}(n_1)^{[2]}}{n(n)^{[2]}} \tag{2.2.5}$$

Similar formulas exist for the mean and the variance of r_2.

Applying similar methods to (2.1.7), one can find the general factorial moments of $(r_1 - 1)$ and $(r_2 - 1)$,

$$E\left[(r_1-1)^{[g_1]}(r_2-1)^{[g_2]}\right] = \frac{(n_1-1)^{[g_1]}(n_2-1)^{[g_2]}}{\binom{n}{n_1}} \binom{n-g_1-g_2}{n_1-g_2} \qquad (2.2.6)$$

2.3 Distribution and Moments of Runs of k Kinds of Elements

Let $a_1, a_2, \ldots, a_k$ be k kinds of elements. Suppose there are n_i elements of the i^{th} kind. Then let

$$n = \sum_{i=1}^{k} n_i, \quad r_i = \sum_{j=1}^{n_i} r_{ij}$$

where r_{ij} denotes the number of runs of elements of the i^{th} kind of length j.

Using the same argument as in (2.1.3) gives

$$P(r_{ij}) = \frac{\prod_{i=1}^{k} \begin{bmatrix} r_1 \\ r_{ij} \end{bmatrix} F(r_1, r_2, \ldots, r_k)}{\begin{bmatrix} n \\ n_1 \end{bmatrix}} \qquad (2.3.1)$$

where the function $F(r_1, r_2, \ldots, r_k)$, which will be referred to hereafter simply as $F(r_i)$, represents the number of different arrangements of r_1 objects of one kind, r_2 objects of a second kind, and so forth, such that no two adjacent objects are of the same kind.

The exact expression for $F(r_i)$ can be found using generating functions and is given in Mood (1940).

The probability $P(r_i)$ is obtained by summing (2.3.1) over r_{ij} with r_i fixed by means of a generalization of identity (2.1.4), giving

$$P(r_i) = \frac{\prod_{i=1}^{k} \binom{n_i - 1}{r_i - 1} F(r_i)}{\begin{bmatrix} n \\ n_i \end{bmatrix}} \tag{2.3.2}$$

<u>Moments</u>. It is possible to find moments of r_i as distributed by (2.3.2). Since $\sum_{r_i} P(r_i) = 1$, the following is true:

$$\sum_{r_i} \prod_{i=1}^{k} \binom{n_i - 1}{r_i - 1} F(r_i) = \begin{bmatrix} n \\ n_i \end{bmatrix}. \tag{2.3.3}$$

From (2.3.3) the moments are derived by putting $u_i = n_i - r_i$. The factorial moments of u_i are derived below.

$$\begin{aligned}
\sum_{r_i} \prod u_i^{[a_i]} \prod \binom{n_i-1}{r_i-1} F(r_i) &= \sum_{r_i} \prod (n_i-r_i)^{[a_i]} \prod \binom{n_i-1}{r_i-1} F(r_i) \\
&= \sum_{r_i} \prod (n_i-1)^{[a_i]} \prod \binom{n_i-a_i-1}{r_i-1} F(r_i) \\
&= \prod (n_i-1)^{[a_i]} \sum \prod \binom{n_i-a_i-1}{r_i-1} F(r_i) \\
&= \begin{bmatrix} n-\sum a_i \\ n_i-a_i \end{bmatrix} \prod_{i=1}^{k} (n_i-1)^{[a_i]}
\end{aligned}$$

The summation involved in the last step is given by (2.3.3). The factorial moments of the u_i can be obtained by dividing the last equation by $\begin{bmatrix} n \\ n_i \end{bmatrix}$.

$$E\left(\prod_{i}^{k} u_i^{[a_i]}\right) = \frac{\begin{bmatrix} n - \sum a_i \\ n_i - a_i \end{bmatrix} \prod_{i=1}^{k} (n_i - 1)^{[a_i]}}{\begin{bmatrix} n \\ n_i \end{bmatrix}} \qquad (2.3.4)$$

From (2.3.4) the moments of the r_i may be found; the means, variances, and covariances are

$$E(r_i) = \frac{n_i(n - n_i + 1)}{n} \qquad (2.3.5)$$

$$\sigma_{ij} = \frac{n_i^{[2]} n_j^{[2]}}{nn^{[2]}} \qquad (2.3.6)$$

$$\sigma_{ii} = \frac{n_i^{[2]} (n - n_i + 1)^{[2]}}{nn^{[2]}} \qquad (2.3.7)$$

The moments of the variables r_{ij} in the distribution (2.3.1) are obtained by means of identities similar to (2.3.3).

2.4 Asymptotic Distribution

The distributions obtained in the previous section are asymptotically normal when the n_i become large in such a way that the ratio n_i/n, denoted by e_i, remains fixed.

The limit theorems for distributions such as (2.1.3) and (2.1.6) cannot be derived because the number of independent variables increases with n. The asymptotic character of the distribution (2.1.10) will be given in the following theorems. Other theorems of similar nature can be proved using the same procedure.

<u>Theorem</u>. The variables

$$x_i = \frac{s_{1i} - ne_1^{\,i}\, e_2^{\,2}}{\sqrt{n}} \qquad i < k \qquad (2.4.1)$$

$$x_k = \frac{s_{1k} - ne_1^{\,k}\, e_2}{\sqrt{n}}$$

are asymptotically normally distributed with zero means and variances and covariances

$$\sigma_{ij} = e_1^{\,i+j-1}\, e_2^{\,3}\Big[(i+1)(j+1)e_1e_2 - ije_2 - 2e_1\Big] \quad i, j < k,\ i \neq j$$

$$\sigma_{ii} = e_1^{\,2i-1}\, e_2^{\,3}\,\Big[(i+1)^2\, e_1e_2 - i^2e_2 - 2e_1\Big] + e_1^{\,2}e_2^{\,2},\ i < k \qquad (2.4.2)$$

$$\sigma_{ik} = e_1^{\,i+k-1}\, e_2^{\,2}\Big[\,(i+1)ke_1e_2 - ike_2 - e_1\,\Big] \quad i < k$$

$$\sigma_{kk} = e_1^{\,2k-1}\, e_2\,\Big[k^2(e_1-1)e_2 - e_1\Big] + e_1^{\,k}e_2$$

Proof. Let us make the substitutions

$$n_i = e_i n \qquad i = 1, 2$$

$$s_{1i} = ne_1^{\,i}e_2^{\,2} + \sqrt{n}\, x_i \qquad i = 1, 2, \ldots, k-1$$

$$s_{1k} = ne_1^{\,k}e_2 + \sqrt{n}\, x_k$$

$$s_1 = ne_1e_2 + \sqrt{n}\sum_1^k x_i$$

$$A = n(e_1 - e_1^{\,k} - ke_1^{\,k}e_2) + \sqrt{n}\sum_1^{k-1} ix_i$$

in equation (2.1.10) and estimate the factorial by means of Stirling's formula

$$m! = \sqrt{2\pi}\; m^{m+(1/2)}\; e^{-m}\,(1 + O(1/m)) \qquad (2.4.3)$$

First note that the exponential factors cancel out because the sum of lower indices of a binomial or multinomial coefficient is equal to the upper index. Also simplifying the expression by considering in detail only terms which involve the x_i, the normalizing constant can be determined from the final limit function. Any function of the parameters will be represented by the letter k. Thus in (2.4.3) we need consider only the factor $m^{m+(1/2)}$. All factorials will be of the form

$$m! = (na + \sqrt{n}\ L(x) + b)!$$

where $L(x)$ is a linear function of the x_i, and a and b are independent of n and x_i. Now

$$\begin{aligned} m^{m+(1/2)} &= (na+\sqrt{n}L(x)+b)^{na+\sqrt{n}L(x)+b+(1/2)} \\ &= (na)^{na+\sqrt{n}L(x)+b+(1/2)}\left(1+\frac{L(x)}{a\sqrt{n}}+\frac{b}{na}\right)^{na+\sqrt{n}L(x)+b+(1/2)} \\ &= k(na)^{\sqrt{n}L(x)}\left(1+\frac{L(x)}{a\sqrt{n}}+\frac{b}{na}\right)^{na+\sqrt{n}L(x)+b+(1/2)} \end{aligned}$$

and

$$\begin{aligned} \log m^{m+(1/2)} &= k+\sqrt{n}L(x)\log na+(na+\sqrt{n}L(x)+b+(1/2)) \\ &\quad \cdot \log\left(1+\frac{L(x)}{a\sqrt{n}}+\frac{b}{an}\right) \\ &= k+\sqrt{n}L(x)\log na+(na+\sqrt{n}L(x)+b+(1/2)) \\ &\quad \cdot \left(\frac{L(x)}{a\sqrt{n}}+\frac{b}{an}-\frac{L^2(x)}{a^2n}+0(1/n^{3/2})\right) \\ &= k+\sqrt{n}L(x)(1+\log na)+\frac{1}{2a}L^2(x)+0\left(\frac{1}{\sqrt{n}}\right) \qquad (2.4.4) \end{aligned}$$

so terms arising from b (and b + 1/2 in the exponent) will be neglected as they give rise only to terms independent of the x_i or of order $1/n^{1/2}$. Of course, $\log(1 + 0(1/m)) = 0(1/m)$. Thus keeping significant terms only, the result of the substitution (2.4.3) and (2.4.4) in (2.1.10) after taking logarithms and using (2.4.4) is

$$-\log P(s_{1i}) = k+\sqrt{n}\sum_1^{k-1} x_i(\log ne_1{}^i e_2{}^2+1) + \sum_1^{k-1}\frac{x_i{}^2}{2e_1{}^i e_2}$$

$$-\sqrt{n}(\sum_1^k x_i)(\log ne_2{}^2+1) + \frac{1}{2e_2{}^2}(\sum_1^k x_i)^2$$

$$+\sqrt{n}(\sum_1^{k-1} ix_i+(k-1)x_k)(\log ne_1{}^k+1)$$

$$-\frac{1}{2e_1{}^k}(\sum_1^k i\,x_i + (k-1)x_k)^2 \qquad (2.4.5)$$

$$+ 2\sqrt{n}\,x_k(\log ne_1{}^k e_2 + 1) + \frac{x_k{}^2}{e_1{}^k e_2}$$

$$-\sqrt{n}(\sum_1^k i\,x_i)(\log ne_1{}^{k+1} + 1)$$

$$+\frac{1}{2e_1{}^{k+1}}(\sum_1^k i\,x_i)^2 + O(1/\sqrt{n})$$

The coefficients of $x_i (i < k)$ and x_k are

$$\sqrt{n}(\log ne_1{}^i e_2{}^2+1 - \log ne_2{}^2-1+i\log ne_1{}^k+i-i\log ne_1{}^{k+1}-i) = 0$$

$$\sqrt{n}(-\log ne_2{}^2-1+k\log ne_1{}^k+k-\log ne_1{}^k-1 + 2\log ne_1{}^k e_2+2 - k\log ne_1{}^{k+1}-k) = 0$$

Hence only the quadratic terms remain and (2.4.5) may be written as

$$-\log P = k + 1/2\sum_{i,j}\sigma^{ij} x_i x_j + O(1/\sqrt{n})$$

where

$$\sigma^{ij} = \frac{1}{e_2{}^2} + \frac{ije_2}{e_1{}^{k+1}} \qquad i,\ j < k,\ i \neq j$$

$$\sigma^{ii} = \frac{1}{e_2{}^2} + \frac{1}{e_1{}^i e_2{}^2} + \frac{i^2 e_2}{e_1{}^{k+1}} \qquad i < k$$

$$\sigma^{ik} = \frac{1}{e_2{}^2} + \frac{i + i(k-1)e_2}{e_1{}^{k+1}} \qquad i < k$$

$$\sigma^{kk} = \frac{1}{e_2{}^2} + \frac{2}{e_1{}^k e_2} + \frac{k^2}{e_1{}^{k+1}} - \frac{(k-1)^2}{e_1{}^k}$$

It is merely a matter of straightforward multiplication of the two matrices to verify that $||\sigma^{ij}||$ is the inverse of $||\sigma_{ij}||$, hence is a positive definite matrix. Then

$$P = ke^{-1/2\sum \sigma^{ij}x_i x_j}\ (1 + O(1/\sqrt{n}))$$

In this equation k must necessarily contain the factor $(1/\sqrt{n})^k$ because there are k + 5 factorials in the denominator and 5 in the numerator of (2.1.10). Since $\Delta r_i = 1$, this factor, in view of (2.4.1), may be replaced by $\prod \Delta x_i$, so

$$P = ke^{-1/2\sum \sigma^{ij}x_i x_j} \prod \Delta x_i (1 + O(1/\sqrt{n}))$$

By restricting the x_i to any finite region R in the x-space, the function $O(1/n^{1/2})$ approaches zero uniformly as $n \longrightarrow \infty$. Thus if $A_i < B_i$ are any positive numbers such that the corresponding values of x_i, say a_i and b_i, obtained by substituting A_i and B_i for r_i in (2.4.1), determine a rectangular region $R'(a_i < x_i < b_i)$, which lies in R, then

$$\sum_{r_i=A_i}^{B_i} P(r_i) = \sum_{x_i=a_i}^{b_i} ke^{-1/2\sum \sigma^{ij}x_i x_j} \prod \Delta x_i (1 + O(1/\sqrt{n}))$$

$$\xrightarrow[n\to\infty]{} \int_{R'} ke^{-1/2\sum \sigma^{ij}x_i x_j} \prod dx_i$$

by the definition of a definite integral and Riemann's fundamental theorem.

Cor. 1. The variable

$$x = \frac{r - ne_1e_2}{\sqrt{n}\; e_1e_2}$$

where r is the total number of runs of one kind of element, is asymptotically normally distributed with zero mean and unit variance.

Cor. 2. The variable $Q = \sum \sigma^{ij} x_i x_j$ is asymptotically distributed according to the χ^2-law with k degrees of freedom.

2.5 Longest Run

The probability that the longest run of a's will be of length s can be obtained by taking n_1 and n_2 as fixed and summing the formula

$$P(r_{1j}) = \frac{\begin{bmatrix} r_1 \\ r_{1s} \end{bmatrix} \binom{n_2 + 1}{r_1}}{\binom{n}{n_1}} \qquad (2.5.1)$$

over all values of r_1 and over all sets of $r_{11}, r_{12}, \ldots r_{1(s-1)}, r_{1s}$ which satisfy

$$\sum_{j=1}^{n_1} jr_{1j} = n_1, \quad \sum_{j} r_{ij} = r_1, \text{ and } r_{1s} \geqslant 1$$

and such that r_1 exceeds neither $n_1 - s + 1$ nor $n_2 + 1$. The probability that the longest run of either a's or b's will be of length s can be obtained by an analogous attack upon the formula

$$P(r_{1j}) = \frac{\begin{bmatrix} r_1 \\ r_{1s} \end{bmatrix} \begin{bmatrix} r_2 \\ r_{2s} \end{bmatrix} F(r_1, r_2)}{\binom{n}{n_1}} \qquad (2.5.2)$$

with the proviso that both r_{1s} and r_{2s} cannot be zero at the same time.

3. STATISTICAL TESTS

The various formulas given above could be used as the bases for a variety of statistical tests of the hypothesis that a's and b's are arranged randomly. The particular formula used would depend upon the conditions given and upon the alternative hypothesis against which one wished the test to be most sensitive. However, calculations of probabilities generally become quite involved at any but the smallest sample size.

3.1 The Wald-Wolfowitz Total Number of Runs Tests

Wald and Wolfowitz (1940) developed the total number of runs test by using the distribution function for the total number of runs. Suppose the two samples have been drawn at random and independently of each other, each from a continuously distributed population. We wish to test whether or not the parent populations are identical. Let U stand for the total number of runs of both a's and b's. The number of runs of a's can be one less than, equal to, or one greater than the number of runs of b's, U can be an odd number in two ways but can be even in only one way.

The probability that the total number of runs will be some even number 2r, using (2.1.7), is

$$P(U = 2r) = \frac{2\binom{n-1}{r-1}\binom{m-1}{r-1}}{\binom{m+n}{n}} \tag{3.1.1}$$

where m and n are sizes of observations designated as a's and b's. The probability that it will be some odd number, 2r + 1, is

$$Pr(U = 2r + 1) = \frac{\binom{n-1}{r-1}\binom{m-1}{r} + \binom{n-1}{r}\binom{m-1}{r-1}}{\begin{matrix} m+n \\ m \end{matrix}} \tag{3.1.2}$$

Lehmann (1951), Dixon (1954), and Epstein (1955) discussed the efficiency of the above test. The Wald-Wolfowitz form of the run test has, relative to student's t-test, an asymptotic relative efficiency of zero and a small sample efficiency, which, when each sample contains five or less observations, generally exceeds .96 and may be as high as .995. The test compares poorly with other distribution-free tests. Epstein (1955) and Lehmann (1951) investigated the power of this test, the former author sampling from normal populations with homogeneous variances, the latter sampling from any continuously distributed population. They found the Wald-Wolfowitz runs test to be inferior in power to the following tests: Student's t, Lehmann's (1951) most powerful test, Mann-Whitney test, median test, and Epstein's (1955) exceedance test. If the ratio m/n of sample sizes remains constant as sample sizes m and n approach infinity, the Wald-Wolfowitz test is consistent. If the ratio m/n does not remain constant but approaches zero or infinity, the test is inconsistent.

Probabilities for U have been tabulated by Swed and Eisenhart (1943) for $m \leq n \leq 20$ and for certain other cases. David (1947) has provided tables appropriate when $m + n \leq 14$ and $2 \leq U \leq 14$.

3.2 Length of Longest Run As a Test for Randomness Against Trend Alternatives

This test has been proposed by Mosteller (1941). Suppose that a series of observations has been taken upon a continuously distributed variable and that the observations have been arranged in the order in which they were drawn, no two observations having been drawn simultaneously. If each observation is now labeled A or B, depending upon whether it is above or below the median for the entire series, the presence or absence of trend can be tested by using as the test statistic one of the following: The length of the longest run of A's (or B's), or the length of the longest run considering both A's and B's. If there is an odd number of observations, one of them will be the median and it should be discarded. Mosteller has published appropriate tables for the cases where $n_1 = n_2 = 5, 10, 15, 20,$ or 25. The use of only the length of the longest run ignores the "information" contained in the lengths of the less than longest runs. Bateman (1948) investigated the case where this statistic was found to be less powerful than the total number of runs.

3.3 The Sum of Squared Run Lengths

The total number of runs does not directly take account of the lengths of runs which are more explicit indices of tendency of like objects to cluster. Ramchandran and Ranganathan (1953) proposed a test which overcomes this objection. They found a new statistic, N, which is the sum of the squares of the lengths of runs, i.e.,

$$N = \sum_j j^2 r_{1j} + \sum_j j^2 r_{2j}$$

Thus all runs are taken account of, but each run is permitted to influence the test statistics in proportion to the square of its length. Ramchandran and Ranganathan recommend the test for the same situation dealt with by Wald and Wolfowitz, the test being used to decide if the two samples came from identical continuous populations. The authors, considering only the case where $n_1 = n_2$, have tabulated the values of N required for various levels of significance. The table values of N are exact for the cases $3 \leq n_1 \leq 5$ and approximate for $6 \leq n_1 \leq 15$, in the later case having been obtained by reading points from a Pearson type VI curve fitted to the true distribution of N.

3.4 Dixon Test

Dixon (1940) presented a criterion for testing the hypothesis that two samples have been drawn from populations with the same distribution function, assuming only that the cumulative distribution function common to the two population is continuous. Let the two samples O_m and O_n be of size m and n, respectively. Assume that $n \leq m$ without loss of generality. Suppose the elements $u_1, u_2, \ldots, u_n$ of O_n are arranged in order from the smallest to the largest, that is, $u_1 < u_2 < \ldots, u_n$.

This sequence can be represented by points along a line. The elements of O_m represented as points on the same line are divided into (n + 1) groups by the first sample, O_n. Let m_1 be the number of points having a value less than u_1, m_i the number lying between u_i and $u_{i+1} (i = 1, 2, \ldots, n)$ and m_{n+1} the number greater than u_n, $(m_{n+1} = m - m_1 - m_2 - \ldots, m_n)$. The criterion Dixon proposed is

$$c^2 = \sum_{i=1}^{n+1} \left(\frac{1}{n+1} - \frac{m_i}{m} \right)^2$$

The mean and variance of c^2 can be found to be as follows.

$$E(c^2) = \frac{n(n + m + 1)}{m(n + 1)(n + 2)}$$

$$\sigma_{c^2}^2 = \frac{4n(m - 1)(m + n + 1)(m + n + 2)}{m^3(n + 2)^2(n + 3)(n + 4)}$$

<u>Significance Values of c^2</u>. Let c_α^2 be defined as the smallest value of θ for which

$$P(c^2 \geq \theta) \leq \alpha$$

Then the values of c_α^2 can be computed for small values of m and n. The probability $P(c^2 \geqslant c_\alpha^2)$ will in general be less than α because the distribution of c^2 is not continuous.

For large values of m and n, Dixon (1940) fitted a gamma distribution to the distribution of nc^2 by the method of moments. By using the transformation $X^2 = nc^2$, nkc^2 is considered as distributed as chi-square with ν degrees of freedom, where ν is not necessarily an integer. Chi-square tables can be used for approximate values of the probability that nkc^2 will exceed certain values.

4. RUNS UP AND DOWN

4.1 Introduction

Let $s = (h_1, \ldots, h_n)$ be a random permutation of the n unequal numbers $a_1, \ldots, a_n$, and let R be the sequence of signs (+ or -) of the differences $h_{i+1} - h_i (i = 1, \ldots, n - 1)$. It is assumed that each of the n! sequences s is equally probable. A sequence of p consecutive plus signs not immediately preceded or followed by a plus sign is called a run up of length p; a sequence of p consecutive minus signs not immediately preceded or followed by a minus sign is called a run down of length p. The term "run" will denote 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. Let r_p and r'_p be the number of runs up and down in R of lengths p and p or more respectively. Levene and Wolfowitz (1944) found the exact values of $\sigma(r_p r_q)$, $\sigma^2(r_p)$, $\sigma(r'_p\, r'_q)$, $\sigma^2(r'_p)$, and $\sigma(r_p r'_q)$. The values of $E(r_p)$ and $E(r'_p)$ were also found. Certain misconceptions about applications of runs were also discussed by Levene and Wolfowitz.

J. Wolfowitz (1944a) established several theorems about the limiting distribution of a class of functions of runs up and down. These results apply to a large class of "runs". A new recursion formula was found by Olmstead (1946) to give the exact distribution of arrangements of n numbers, no two alike, with runs up and down of length p or more. These were tabled for n and p through n = 14. An exact solution is given for $p \geqslant n/2$. Olmstead (1946) also presented in simplified form the mean and variance determined by Levene and Wolfowitz (1944). Wolfowitz (1944) has shown that the limiting distribution for runs up and down is a Poisson distribution. Olmstead (1946) applied his derivation to the distribution of runs of length p or more and obtained identical conclusions for such runs. He gives tables for exact numbers of arrangements of n numbers with runs of length p or more and for the fraction of arrangement of n numbers with runs of length p or more based on the Poisson distribution.

4.2 Total Number of Runs Up and Down

The total number of runs is simply the number of runs of pluses or minuses of length 1 or greater; and Moore and Wallis (1943) showed that when n is greater than 2, it will have an expected value of $\frac{2n - 1}{3}$ and a variance of $\frac{16n - 29}{90}$. The total number of runs, r, is asymptotically normally distributed, so for large values of n the significance of the total number of runs can be tested by treating r as a normal deviate and referring the critical ratio

$$\frac{r - \frac{2n - 1}{3}}{\sqrt{\frac{16n - 29}{90}}}$$

to normal tables. By reducing the absolute value of the numerator by one-half, the critical ratio can be corrected for continuity.

There are (r - 1) turning points of the series, if r is the total number of runs. The test based on the total number of runs and a test based on the number of turning points are equivalent. The expected number of turning points, T, is $(2n - 4)/3$ and its variance is the same as that for the total number of runs. Therefore the significance of the number of turning points can be tested by forming the critical ratio analogous to the one given above, referring it to normal tables.

A. Stuart (1954) mentions in his paper that when all tests concerned are applied to samples from normally distributed populations, the turning point test has an asymptotic relative efficiency of zero with respect to the regression coefficient test and also with respect to each of the distribution free tests of randomness with which it was compared.

4.3 Chi-square Applied to Run Frequencies

Wallis and Moore (1941) suggested a chi-square test of significance applied in the usual way to the observed frequencies of "interior" runs (all runs except the runs at both ends) of like signs of length one, two, and over two, with the corresponding expected frequencies being $\frac{5(n-3)}{12}$, $\frac{11(n-4)}{60}$, and $\frac{4n-21}{60}$. There are two degrees of freedom, one degree having been expended by obtaining n from the sample. The test, however, is an approximate one if the significance of the calculated chi-square is obtained from the usual chi-square tables. This is the case because the run lengths are not entirely independent of one another, although the chi-square test assumes that they are. Various empirically obtained "corrections" are offered by Wallis and Moore for use when n exceeds 12. However, for $6 \leq n \leq 12$, they have provided a table of exact probabilities for the values of chi-square as calculated from the sample. These were obtained by means of a recursion formula and give, in effect, that proportion of the $n!$ permutations which yield a value of chi-square as great or greater than one tabled.

The test can be used as a test of randomness against either trend or correlation alternatives. In the later application, if two measurements a and b have been taken on each of n objects, the objects are arranged in order of increasing magnitude of one continuously distributed variable and the run test is applied to measurements on the other variable.

5. REMARKS

Suppose we have a random sample of m observations on one variable and a similar sample of n observations on another variable. 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 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 problems not only of two, but in general, of larger numbers of samples may arise. Runs up and down are widely used in quality control and have been applied to economic time series.

ACKNOWLEDGMENT

The writer wishes to express his appreciation to his major professor, Dr. W. J. Conover, for the assistance and advice extended to him during the preparation of this report.

BIBLIOGRAPHY

Bateman, G. (1948). On the power function of the longest run as a test for randomness in a sequence of alternatives. Biometrika, 35 97-112.

Bortkiewicz, L. V. (1917). Die Iterationen, Berlin.

Bruns, H. (1906). Wahrscheinlichkeitsrechnung und Kollektivmasslehre, Leipzig, page 216.

Dixon, W. J. (1940). A criterion for testing the hypothesis that two samples are from the same population. Annals of Mathematical Statistics. 11 199-204.

Dixon, W. J. (1954). Power under normality of several non-parametric tests. Annals of Mathematical Statistics, 25 610-614.

Epstein, B. (1955). Comparison of some non-parametric tests against normal alternatives with an application to life testing. Journal of the American Statistical Association, 50 894-900.

Ferguson, Thomas S., and Kraft, Charles H. (1955). A run test of the hypothesis that the median of a stochastic process is constant. Annals of Mathematical Statistics, 26 770-1112.

Goodman, L. A. (1957). Runs tests and likelihood ratio tests for Markov chains. Annals of Mathematical Statistics, 28 1072 (No. 50).

Grunbaum, H. (1904). Isolierte und reine Gruppen und die Marbe'sche Zahl "p", Wurzburg.

Ising, E. (1925). Deitrag zur theorie des Ferromagnetisms. Zeitschrift fur Physik, 31 253-258.

Krishner Iyer, P. V. (1950). The theory of probability distributions of points on a lattice. Annals of Mathematical Statistics, 21 198-217.

Kruskal, W. H. (1952). A non-parametric test for the several sample problems. Annals of Mathematical Statistics, 23 525-540.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain non-parametric tests. Annals of Mathematical Statistics, 22 165-179.

Levene, H. (1952). On the power function of tests of randomness. Annals of Mathematical Statistics, 23 34-56.

Levene, H. and Wolfowitz, J. (1944). The covariance matrix of runs up and down. Annals of Mathematical Statistics, 15 58-69.

Mann, Henry B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18 50-60.

Marbe, Karl (1899). Naturphilosophische Untersuchungen zur Wahrscheinlichkeitlehre, Leipzig.

Marbe, Karl (1916). Die Gleichformigkeit in der Welt, Munchen.

Marbe, Karl (1916). Mathematische Bemerkungen, Munchen.

Marbe, Karl (1934). Grundfragen der angewandten Wahrscheinlichkeitsrechnung, Munchen.

Mises, R. V. (1921). (Name of article not known.) Zeitschrift fur angew. Math. u. Mech., Vol. 1, page 298.

Mood, A. M. (1940). The distribution theory of runs. Annals of Mathematical Statistics, 11 367-392.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. Annals of Mathematical Statistics, 25 514 522.

Moore, G. H. and Wallis, W. A. (1943). Time series significance tests based on signs of differences. Journal of the American Statistical Association, 38 153-164.

Mosteller, F. (1941). Note on an application of runs to quality control charts. Annals of Mathematical Statistics, 12 228-232.

Noether, G. E. (1950). Asymptotic properties of Wald-Wolfowitz test of randomness. Annals of Mathematical Statistics, 21 231-246.

Olmstead, P. S. (1946). Distribution of sample arrangements for runs up and down. Annals of Mathematical Statistics, 17 24-33.

Pearson, Karl (1897). The chances of death and other studies in evolution, London, Vol. I, Chap. 2.

Ramchandran, P. S. and Ranganathan, J. (1953). A non-parametric two-sample list. Journal of Madras University, Sec. B, 35 202.

Stevens, W. L. (1939). Distribution of groups in a sequence of alternatives. Annals of Eugenics, 9 10-17.

Stuart, A. (1954). Asymptotic relative efficiencies of distribution free tests of randomness against normal alternatives. Journal of the American Statistical Association, 49 147-157.

Swed, Frieda S., and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. Annals of Mathematical Statistics, 14 66-87.

Wallis, W. A., and Moore, G. H. (1941). A significance test for time series analysis. Journal of the American Statistical Association, 36 401-409.

Wald, A., and Wolfowitz, J. (1940). On a test whether two samples are from the same population. Annals of Mathematical Statistics 11 147-162.

Weiss, Lionel (1960). Two sample tests for multivariable distribution. Annals of Mathematical Statistics, 31 159-164.

Wilks, S. S. (1962). Mathematical Statistics, New York, Wiley Publication, 144-150.

Wolfowitz, J. (1943). On the theory of runs with some applications to quality control. Annals of Mathematical Statistics, 14 280-288.

Wolfowitz, J. (1944a). Asymptotic distributions of runs up and down. Annals of Mathematical Statistics, 15 163-172.

Wolfowitz, J. (1944b). Note on runs of consecutive elements. Annals of Mathematical Statistics, 15 97-98.

THEORY OF RUNS

by

SHASHI N. SHARMA

B. S., Michigan State University, 1963

AN ABSTRACT OF
A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

This report is a summary of the principal papers published since 1940 on the theory of runs. Also presented in this paper are statistical tests derived from the main distributions in the theory of runs. All of these papers are discussed in brief except for Mood (1940) which has been discussed in detail.

The "two-sample" problem is examined by Mood (1940), using runs. Suppose there are n elements of two kinds, say, n_1, a's, and $n_2 = n - n_1$, b's, and that these are arranged at random in a row. If $r_{ij} (i = 1, 2)$ is the number of runs of j of elements of variety i, the probability of obtaining a given set of values of r_{ij} is obtained. Besides this basic distribution function, there are certain marginal distributions such as that for the occurrence of a given set of runs in the a's regardless of how the b's fall, or that for r_1 and r_2 if these are respectively the total number of runs of a's and of b's, or that of r_1 or r_2 alone. Factorial moments, mean, variances, and covariances are found. Similar results are obtained in case there are more than two kinds of elements.

Wald and Wolfowitz used the distribution function for the total number of runs (irrespective of length) to provide a test of the hypothesis that two samples have come from the same population. A test which takes into account the length of the runs is discussed. Dixon's criterion for testing the hypothesis that two samples have been drawn from populations with the same distribution function is also presented.

Runs up and down are also considered in brief. A chi-square test of significance applied in the usual way to the observed frequencies of "interior" runs of like signs of length one, two, and over two as suggested by Wallis and Moore (1941) is discussed.