A multivariate runs statistic
by Roy Neal Byrd

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in Mathematics
Montana State University

Abstract:
In this work, a new multivariate statistic is introduced which is a generalization of the runs statistic as proposed by A.M. Mood and others. Only absolutely continuous distributions are considered. Some distribution theory in a certain special case is given in Chapter II for the multivariate runs statistic. In Chapter III, a large sample property, consistency, is shown to hold for test of the hypothesis H0: F = G based on runs, with F, G, p-variate, p > 1. In Chapter V, this idea is extended to the hypothesis H0: F1 = F2 = ... Fk, k > 2, Fi p- variate, p > 1. It is assumed in both of these sections that if Fi ≠ Fj, then at least one set of marginals differ also. The concept of runs when the sample numbers are random variables is introduced in Chapter VI. The distributions of the sample numbers are assumed to be binomial or multinomial, depending upon whether there are two, or more, populations sampled. In Chapter VII, consistency is proved for the multivariate runs test of H0: F = G when the sample numbers are binomially distributed. This result is extended to the runs test of H0: F1 = F2 = ... Fk, k > 2, in Chapter VIII. It is again assumed that if Fi ≠ Fj, then at least one set of the marginals differ. l ,1

A MULTIVARIATE RUNS STATISTIC

by

ROY NEAL BYRD

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematics

Approved:

Louis C. Barrett
Head, Major Department

K. J. Tiahrt
Chairman, Examining Committee

Henry L. Parsons
Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

August, 1971

ACKNOWLEDGMENT

The author wishes to express his gratitude to his thesis adviser, Dr. K.J. Tiahrt, for the helpful suggestions and criticisms he made toward the completion of this work.

The author also appreciates the financial support extended by the National Science Foundation through a National Science Foundation Traineeship and National Science Foundation Research Grant GP-8935.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF PROGRAMS

ABSTRACT

In this work, a new multivariate statistic is introduced which is a generalization of the runs statistic as proposed by A.M. Mood and others. Only absolutely continuous distributions are considered. Some distribution theory in a certain special case is given in Chapter II for the multivariate runs statistic. In Chapter III, a large sample property, consistency, is shown to hold for test of the hypothesis H_0: $F = G$ based on runs, with F, G, p-variate, $p > 1$. In Chapter V, this idea is extended to the hypothesis H_0: $F_1 = F_2 = \ldots F_k$, $k > 2$, F_i p-variate, $p > 1$. It is assumed in both of these sections that if $F_i \neq F_j$, then at least one set of marginals differ also. The concept of runs when the sample numbers are random variables is introduced in Chapter VI. The distributions of the sample numbers are assumed to be binomial or multinomial, depending upon whether there are two, or more, populations sampled. In Chapter VII, consistency is proved for the multivariate runs test of H_0: $F = G$ when the sample numbers are binomially distributed. This result is extended to the runs test of H_0: $F_1 = F_2 = \ldots = F_k$, $k \geq 2$, in Chapter VIII. It is again assumed that if $F_i \neq F_j$, then at least one set of the marginals differ.

I. INTRODUCTION, NOTATION, AND PREVIOUS RESULTS

Development of multivariate non-parametric theory was comparatively dormant until the mid-1960's. At this time, the appearance of two papers opened the way for further investigations into such problems.

Chatterjee and Sen [2] considered the matrix obtained by ranking a set of multivariate observations coordinatewise, and showed that statistics based on this matrix are conditionally distribution-free under the hypothesis of identical distributions. Hence distribution-free statistics became available in the general multivariate case. A further discussion of the rank matrix concept is given at the beginning of the second chapter.

Puri and Sen [7] extended the theorem of Chernoff and Savage, 1958, to the multivariate case. They considered the set of random variables $Z_{i\alpha}^{k} = 1$ if the α-th smallest observation in variate i is from the k-th population, and $Z_{i\alpha}^{k} = 0$ otherwise. Here there are samples from $c \geq 2$ p-variate populations, and the total sample size is N. Puri and Sen studied statistics of the form

$$T_{Ni}^{k} = \sum_{\alpha=1}^{N} E_{N\alpha}^{i} \, Z_{i\alpha}^{k}, \quad k=1,\ldots,c; \; i=1,\ldots,p, \quad \text{where}$$

$\{E_{N\alpha}^{i}, \; \alpha=1,\ldots,N; \; i = 1,\ldots,p\}$ are given numbers satisfying

certain regularity conditions. These statistics are conditionally distribution-free since they are linear functions of the $Z_{i\alpha}^{k}$ which depend solely on the rank matrix of the observations. These authors further examined the use of the above statistics to test the hypothesis of c identical distributions, and showed the test is quite powerful against the alternatives of differing means or variances.

The theory of runs for univariate populations started toward the end of the nineteenth century, but it was not until 1925 that an actual distribution function appeared. W.L. Stevens furthered the theory in 1939, and Wald and Wolfowitz [9] published the same distribution, proving asymptotic normality, in 1940. Mood [6] in 1940 published an extensive study of runs from more than two univariate populations in the cases of fixed sample numbers, or multinomially distributed sample numbers. The exact and asymptotic distributions connected with runs were derived in Mood's paper.

It is the purpose of the following study to extend the theory of runs in the multivariate case. A new test statistic, which is a quadratic function of the $Z_{i\alpha}^{k}$ and is

therefore based on the rank matrix will be developed. The discrete and asymptotic distributions of this statistic will be derived under conditions of independence, and it will be shown that the test of identical populations based on this statistic is quite powerful, in a given sense, against all alternative hypotheses for a certain family of distributions.

Definition 1: A run is a sequence of like objects preceded and/or succeeded by different objects. For example, the sequence aabbbabaa has three runs of a's and two runs of b's.

Definition 2: The length of a run is the number of objects in the run. The first run of a's above is of length two; the first run of b's is of length three.

It will be assumed that there is one sample from each of a given set of populations, and the following notation will be used throughout this study:

k = the number of populations;

n_i = the sample size from population i;

n = the total sample size = Σn_i;

p = the dimension of each population;

r_{ij}^{w} = the number of runs of length j from population i in variate w;

(1) r_i^w = the number of runs from population i in variate w;

$= \sum_j r_{ij}^w$;

r_i = total number of runs from population i;

$= \sum_w r_i^w$;

$r^w = \sum_i r_i^w$;

= the number of runs in variate w;

r_{ij} = the number of runs of length j in population i;

$= \sum_w r_{ij}^w$.

Wald and Wolfowitz [9] and Mood [6] have completely specified the discrete and asymptotic distributions of the above statistics when p = 1. These results are summarized below. Define

$F(x,y) = 2$ if $x = y$,

$= 1$ if $x - y = \pm 1$,

$= 0$ otherwise.

If k = 2, then

$$(2) \qquad P(r^W_{ij}) = \frac{\begin{bmatrix} r^W_1 \\ r^W_{1j} \end{bmatrix} \begin{bmatrix} r^W_2 \\ r^W_{2j} \end{bmatrix} F(r^W_1, r^W_2)}{\binom{n}{n_1}}$$

where $\begin{bmatrix} A \\ A_i \end{bmatrix} = \frac{A!}{A_1! \ldots A_s!}$, and $\sum_i A_i = A$.

Also

$$(3) \qquad P(r^W_1, r^W_2) = \frac{\binom{n_1-1}{r^W_1-1} \binom{n_2-1}{r^W_2-1} F(r^W_1, r^W_2)}{\binom{n}{n_1}} .$$

Further

$$(4) \qquad P(r^W) = \binom{n_1 - 1}{r^W/2 - 1} \binom{n_2 - 1}{r^W/2 - 1} \frac{2}{\binom{n}{n_1}} \quad \text{if } r^W \text{ is even,}$$

$$= \left[\binom{n_1 - 1}{r^W_{+1}/2 - 1} \binom{n_2 - 1}{r^W_{-1} - 1} + \binom{n_1 - 1}{\frac{r^W-1}{2} - 1} \binom{n_2 - 1}{\frac{r^W+1}{2} - 1} \right] \frac{1}{\binom{n}{n_1}}$$

if r^W is odd.

And

(5) $$P(r_1^W) = \frac{\binom{n_1 - 1}{r_1^W - 1}\binom{n_2 + 1}{r_1^W}}{\binom{n}{n_1}} .$$

If k is greater than 2, then

(6) $$P(r_{ij}^W) = \prod_{i=1}^{k} \frac{\begin{bmatrix} r_i^W \\ r_{ij}^W \end{bmatrix} F(r_i^W)}{\begin{bmatrix} n \\ n_i \end{bmatrix}}$$

where $$F(r_i) = \sum_{a_i, n_{ij}} \frac{\prod_{i=1}^{k} n_i - 1 \; n_{ij}}{\prod_{i=1}^{k} \binom{n_i - 1}{r_i - 1}} .$$

Also

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

Additional distributions have been obtained, but are so complicated as to be of little use.

The following moments exist:

$$E(r_{ij}^W) = (n - n_1 + 1)^{(2)} n_i^{(j)} / n^{(j+1)}$$

where $x^{(a)} = x(x - 1) \ldots (x - a + 1)$;

$$E(r_i^W) = n_i(n - n_i + 1)/n;$$

$$E(r^W) = \frac{\left(\Sigma n_i\right)^2 - \Sigma n_i^2}{n} + 1 ;$$

(8)

$$\mathrm{Cov}(r_i^W, r_j^W) = \frac{n_i^{(2)} n_j^{(2)}}{n n^{(2)}} ;$$

$$\mathrm{Var}(r_i^W) = \frac{n_i^{(2)} (n - n_i + 1)^{(2)}}{n n^{(2)}} ;$$

if $k = 2$, then

$$\mathrm{Cov}(r_{1i}^W, r_{ij}^W) = \frac{n_2^{(2)} (n_2 + 1)^{(2)} n_1^{(i+j)}}{n^{(i+j+2)}} - \frac{n_2^2 (n_2 + 1)^2 n_1^{(i)} n_1^{(j)}}{n^{(i+1)} n^{(j+1)}} ;$$

(9)

$$\mathrm{Var}(r_{1i}^W) = n_2^{(2)} (n_2 + 1)^{(2)} n_1^{(2i)} / n^{(2i+2)} + \frac{(n_2 + 1)^{(2)} n_1^{(i)}}{n^{(i+1)}} \left(1 - \frac{(n_2 + 1)^{(2)} n_1^{(i)}}{n^{(i+1)}}\right) ;$$

$$\mathrm{Var}(r^W) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} .$$

Other moments are also available, but are so complicated as to be of little practical value.

The following asymptotic distributions have been found:

Theorem I: (Wald and Wolfowitz): Let $n_i/n = e_i$, $i = 1, 2$, be fixed, $p = 1$, and $k = 2$. Then the variable

$$x = \frac{r^W - 2ne_1e_2}{2\sqrt{n}\ e_1e_2}$$

is asymptotically normally distributed with zero mean and unit variance.

Theorem II: (Mood): Let $p = 1$ and $n_i/n = e_i$ be fixed for $i = 1, 2, \ldots, k$. Then the variable

$$x = \frac{r^W - n(1 - \Sigma e_i^2)}{n^{\frac{1}{2}}}$$

is asymptotically normally distributed with zero mean and variance

$$\mathrm{Var}(x) = \Sigma e_i^2 - 2\,\Sigma e_i^3 + (\Sigma e_i^2)^2.$$

II. DISTRIBUTIONS OF MULTIVARIATE RUNS

It is desired to find a test of the hypothesis that k p-variate distributions have identical distributions. The test statistic proposed is $r = r^1 + \ldots + r^p$, the total number of runs observed in the p variates.

Definition 3: Two matrices are said to be permutationally equivalent if one may be obtained from the other by a finite number of column interchanges. For a matrix R, let S(R) be the set of matrices which are permutationally equivalent to R.

Consider the matrix R which is the matrix of coordinatewise ranks of the observations. Each row of R is a permutation of the integers 1, ..., n. It has been shown by Chatterjee and Sen [2], or Puri and Sen [7], that under the hypothesis of identical continuous distributions, $P(R = R^* \mid S(R)) = 1/n!$. But this is the discrete uniform distribution, and hence is independent of the distributions in question. Thus any statistic based on R is conditionally distribution-free under the above hypothesis.

Define $v_{ij}^w = 0$ if the j-th smallest observation in variate w is from population i, and $v_{ij}^w = 1$ if the j-th smallest observation in variate w is from population $t \neq i$. Then the number of runs in variate w from populat-

ion i is

$$r_i^w = \sum_{j=2}^{n} \frac{(v_{ij}^w - v_{ij-1}^w)^2 + \delta_{0v_{11}^w} + \delta_{0v_{1n}^w}}{2}$$

where δ_{ij} is the usual Kronecker delta. Hence $r = \sum_{w,s} r_s^w$ is based on R and is thus a conditionally distribution-free statistic for testing the hypothesis of identical continuous distributions. For example, suppose $k = 3$, $p = 2$, $n_1 = n_2 = 2$, and $n_3 = 3$. The seven observed vectors might be $\binom{1}{2}, \binom{-1}{0}; \binom{2}{12}, \binom{0}{1}; \binom{-2}{-1}, \binom{3}{3}, \binom{4}{-5}$.

Then

$$R = \begin{pmatrix} 4 & 2 & 5 & 3 & 1 & 6 & 7 \\ 5 & 4 & 7 & 3 & 2 & 6 & 1 \end{pmatrix},$$

where the first two columns are the component-wise rankings of the observations from distribution 1, the second two columns are related to distribution 2, and the last two columns are related to distribution 3. Consequently S(R) is the collection of 2 x 7 matrices each of which may be obtained by permuting the columns of the above R. A convenient notation for this class, S(R), which will be used is the "standardized" matrix, $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 4 & 3 & 5 & 7 & 6 & 1 \end{pmatrix}$

The complete set of values of v_{ij}^w is given in the following tables:

Values of v_{ij}^1

i \ j	1	2	3	4	5	6	7
1	1	0	1	0	1	1	1
2	1	1	0	1	0	1	1
3	0	1	1	1	1	0	0

Values of v_{ij}^2

i \ j	1	2	3	4	5	6	7
1	1	1	1	0	0	1	1
2	1	1	0	1	1	1	0
3	0	0	1	1	1	0	1

Using the definition for r_i^W one finds $r_1^1 = 1/2[(-1)^2 + (1)^2 + (-1)^2 + (1)^2 + (0)^2 + (0)^2 + 0 + 0] = 2$, indicating that there are two runs of observations in variate one from population one; $r_2^1 = 1/2[(0)^2 + (-1)^2 + (1)^2 + (-1)^2 + (1)^2 + (0)^2 + 0 + 0] = 2$, indicating there are 2 runs of observations in variate one from population two; $r_3^1 = 1/2[(1)^2 + (0)^2 + (0)^2 + (-1)^2 + (0)^2 + 1 + 1] = 2$, so there are 2 runs of observations in variate one from population three. Similarly $r_1^2 = 1$, $r_2^2 = 2$, and $r_3^2 = 2$.

It follows that $r_1 = r_1^1 + r_1^2 = 2 + 1 = 3$, $r_2 = 4$, $r_3 = 5$, and $r = r_1 + r_2 + r_3 = 12$.

It should also be noticed that the number of runs in the t-th variate is independent of the number of runs in the q-th variate if $q \neq t$, if the q-th and t-th variates are independent. This follows because, under the hypothesis of identical distributions, R is distribution-free; hence $P(v_{ij}^t = 0,1 \mid v_{ij}^q) = P(v_{ij}^t = 0,1) = n_t/n, (n - n_t)/n$ when $q \neq t$. But the number of runs in variate t is a function of the v_{ij}^t's, while the number of runs in variate q depends on the v_{ij}^q's. In the following independence between variates is assumed.

Definition 4: Define a p-partition of the positive integer n to be a set of p positive integers $(n_1, \ldots, n_p)$ such that $n_1 + \ldots + n_p = n$. Let $S_p(n)$ denote the set of all p-partitions of n.

Since v_{ij}^t is independent of v_{fs}^q for $t \neq q$,

$$P(r_{ij}) = \sum_{S_p(r_{ij})} \prod_{t=1}^{p} P(r_{ij}^t), \text{ i.e.}$$

$$S_p(r_{ij}) = \{(r_{ij}^1, \ldots, r_{ij}^p) \mid \sum_{t=1}^{p} r_{ij}^t = r_{ij} \text{ and the } r_{ij}^t \text{ are}$$

positive integers}, with $P(r_{ij}^t)$ as in equation (2) if k = 2 and as in (6) if k is greater than two;

$$P(r_i) = \sum_{S_p(r_i)} \prod_{t=1}^{p} P(r_i^t) \text{ with } P(r_i^t) \text{ as}$$

in (5) or (7) as k = 2 or k greater than two;

$$P(r) = \sum_{S_p(r)} \prod_{t=1}^{p} P(r^t) \text{ with } P(r^t) \text{ as in}$$

(4) or as could be derived from (7), as k = 2 or as k is greater than 2. Tables one, two, and three in appendix A give critical values for r at the 5% level of significance for k = 2 and p = 2, 3, and 4, respectively, with independence between variates assumed.

Also the following moments can be found:

$$E(r_{ij}) = \sum_{t=1}^{p} E(r_{ij}^t)$$

$$= p(n - n_i + 1)^{(2)} n_i^{(j)}/n^{(j+i)}$$

from (8);

$$E(r_i) = \sum_{t=1}^{p} E(r_i^t)$$

$$= pn_i(n - n_i + 1)/n \text{ from (8);}$$

$$E(r) = \sum_{t=1}^{p} E(r^t)$$

$$= p + \frac{(\Sigma n_i)^2 - \Sigma n_i^2}{n} p \text{ from (8);}$$

$$\text{Var}(r_i) = \sum_{t=1}^{p} \text{Var}(r_i^t)$$

$$= \frac{p n_i^{(2)} (n - n_i + 1)^{(2)}}{n n^{(2)}}$$

from (8);

$$\text{Var}(r) = \sum_{t=1}^{p} \text{Var}(r^t)$$

$$= \frac{2p n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

when $k = 2$, from (9).

As before, other moments are obtainable but are so cumbersome as to be of little or no practical value.

The asymptotic distributions of r for $k = 2$, and for k greater than two will now be examined, when the p variates are independent.

Theorem III: Let $n_i/n \doteq e_i$ be fixed, p be greater than one, and $k = 2$. Then

$$x = \frac{r - 2pne_1 e_2}{2n^{\frac{1}{2}} p^{\frac{1}{2}} e_1 e_2}$$

is asymptotically normally distributed with zero mean and unit variance.

Proof: $x^w = \dfrac{r^w - ne_1e_2}{n^{\frac{1}{2}}e_1e_2}$ is asymptotically normally distributed with zero mean and unit variance by Theorem I. Hence $p^{\frac{1}{2}}x = \sum_{w=1}^{p} x^w$ is asymptotically normally distributed with zero mean and variance p, since the r^w are independent. So x is asymptotically normally distributed with unit variance and zero mean.

Theorem IV: Let $n_i/n = e_i$ be fixed, k be greater than two, and p be greater than one; then

$$x = \frac{r - np(1 - \Sigma e_i^2)}{n^{\frac{1}{2}}}$$

is asymptotically normally distributed with zero mean and variance

$$\mathrm{Var}(x) = p(\Sigma e_i^2 - 2\Sigma e_i^3 + (\Sigma e_i^2)^2).$$

Proof: $x^w = \dfrac{r^w - n(1 - \Sigma e_i^2)}{n^{\frac{1}{2}}}$ is asymptotically normally distributed with zero mean and variance $\Sigma e_i^2 - 2\Sigma e_i^3 + (\Sigma e_i^2)^2$ by Theorem II. Hence $p^{\frac{1}{2}}x = \sum_{w=1}^{p} x^w$ is asymptotically normally distributed with zero mean and variance $p(\Sigma e_i^2 - 2\Sigma e_i^3 + (\Sigma e_i^2)^2)$ because the r^w are independent. So

x is asymptotically normally distributed with zero mean and variance Var(x).

An example of the use of the above asymptotic distributions will now be given. Assume the hypothesis to be tested is H_0: $f_1 = f_2 = f_3$ and that the alternative hypothesis is H_A: Not all the distributions are identical. Suppose $n_1 = n_2 = n_3 = 25$ and $p = 3$, and that independence between variates can be assumed. The observations from population one are:

3.22	2.02	2.36	2.32	3.45	2.40	2.57
3.70	4.24	4.72	4.59	3.90	3.14	4.74
9.85	9.76	8.51	9.15	7.66	9.21	8.08
3.78	3.71	2.39	2.04	2.68	2.45	3.10
4.16	3.89	4.84	3.15	3.69	4.44	4.99
9.35	8.12	9.78	7.93	9.33	7.17	8.97
2.62	3.85	2.89	3.28	2.77	3.27	3.05
4.98	4.09	4.75	4.79	3.74	3.30	4.15
8.01	9.88	7.81	9.42	9.75	8.14	8.00
2.78	2.86	3.09	2.14			
3.88	3.13	4.69	4.76			
7.21	9.77	7.80	8.06			

From population two, the observations are:

3.76	2.60	2.17	2.61	3.83	3.49	2.94
3.77	4.00	3.83	4.49	3.12	4.86	3.26
6.37	6.42	7.05	7.91	8.72	7.92	8.62
3.04	3.29	3.88	2.06	3.46	3.81	2.00
4.21	3.92	3.41	4.96	3.95	4.29	3.22
6.21	8.42	8.05	9.01	9.83	8.15	6.79

2.95	2.27	2.55	2.93	3.94	2.16	3.41
3.73	3.31	4.22	4.19	3.16	3.66	4.82
6.32	6.72	6.84	9.93	7.61	6.88	6.62
3.84	2.08	3.44	3.77			
4.68	4.81	4.83	3.18			
7.25	7.45	6.31	8.02			

And from the third population, the observations are:

3.86	2.50	3.90	3.40	2.79	2.22	2.65
3.23	3.07	3.39	4.70	3.43	4.32	3.72
6.33	7.29	5.11	7.87	5.02	5.73	5.15
3.89	3.96	3.69	2.66	2.67	3.65	3.03
4.01	4.25	4.35	4.38	3.24	3.67	4.39
5.03	6.54	8.41	6.05	6.71	5.63	7.04
3.55	3.02	2.85	3.72	2.01	3.11	3.12
3.60	3.91	4.71	4.05	3.08	4.97	4.28
5.77	9.65	6.36	5.54	8.16	7.20	7.47
2.03	3.61	2.49	2.56			
4.17	3.58	3.46	4.45			
5.60	6.08	6.55	8.50			

From the above data, $r^1 = 44$, $r^2 = 43$, and $r^3 = 38$. Hence $r = 125$. Using Theorem IV $x = (125 - 75(3)(1 - 1/3))/(75)^{\frac{1}{2}} = -25/(75)^{\frac{1}{2}}$, where $e_i = n_i/n = 1/3$ and $s = (\mathrm{Var}(x))^{\frac{1}{2}} = (3(1/3 - 2(1/9) + 1/9))^{\frac{1}{2}} = (2/3)^{\frac{1}{2}}$. Hence x/s is approximately normally distributed with zero mean and unit variance. The calculated value of x/s is -3.54. Hence from the standard normal table, P(r is less than or equal to 125 $\mid f_1 = f_2 = f_3$) is about .0001, so, since this is "small" there is evidence based on the above data to reject H_0.

III. CONSISTENCY IN THE CASE OF TWO SAMPLES

In general, statistical tests have two large sample properties of interest to the statistician. One of these is relative efficiency. This concept deals with the comparison of the powers of the two tests in the sense that it is the limit of the ratio of the sample sizes required for each test to obtain a given power as the sample numbers tend to infinity.

The large sample property of the runs test which will be examined in this study is called consistency. It will be shown that the proposed test is consistent, that is, the power of the runs test approaches one as the sample numbers increase without bound in size. This concept is now formalized.

Definition 5: Let $A_n(t)$ denote the rejection region for a test of a statistical hypothesis based on the statistic t_n from a random sample of size n. Then the test is called consistent if $\lim_{n \to \infty} P(t_n \in A_n(t) \mid H_0 \text{ is false}) = 1$.

Obviously, a minimal requirement for any good test would be consistency. Wald and Wolfowitz have shown the runs test to be consistent for the case of $p = 1$, and $k = 2$, for a general class of univariate distributions. The next purpose of this work will be to examine the con-

sistency of the runs test in the case of $k = 2$, and p greater than one.

Definition 6: A closed rectangle (in p space) is a cartesian product of (p) closed intervals.

Definition 7: Two multivariate distributions F and G satisfy condition A if for any arbitrarily small $\delta > 0$ there exists a finite number of closed rectangles such that the probability of the union R of these rectangles is greater than $1 - \delta$, according to at least one of the distributions F or G, and such that F and G have positive continuous first partial derivatives $\partial F/\partial x_i$ and $\partial G/\partial x_i$, $i = 1, \ldots, p$, in R.

Definition 8: C is the collection of all pairs of distribution functions (F,G) such that $F \neq G$ and $F_t \neq G_t$ for at least one $t = 1, \ldots, p$, where F_t, G_t denote the marginal distributions in the t-th variate.

In the following, n_1 and n_2 will be considered as variables with fixed ratio $n_1/n_2 = \alpha$. The following theorem will now be proved.

Theorem V: If F and G satisfy condition A and if (F,G) is in C, then

$$\lim_{n_1 \to \infty} P(r < r_0(n_1)) = 1,$$

where $r_0(n_1)$ denotes the significant value of r for the desired level of the test. In other words, the runs test is consistent for k = 2 and p greater than one for distributions in the above collection, C.

The proof will be given in several stages. Suppose the $p \times 1$ vectors X and Y are distributed respectively as F and G, and let $E(r/n_1)$ and $Var(r/n_1)$ denote the mean and variance, respectively of r/n_1.

Lemma 1: If the following are fulfilled:

a) $F_i(x_i) = 0$ if $x_i < 0$,

$= x_i$ if $0 \leq x_i \leq 1$,

$= 1$ if $x_i > 1$;

b) $G(y) = 0$ if $y_i < 0$ for any i,

$= 1$ if $y_i > 1$ for all i;

c) the density g(y) exists, is positive, and continuous everywhere is the rectangle $0 < y_i < 1$, i = 1, ..., p;

then

d) $$\lim_{n_1 \to \infty} E(r/n_1) = 2 \sum_{t=1}^{p} \int_0^1 \frac{g_t(x)}{\alpha + g_t(x)} \, dx;$$

e) and $$\lim_{n_1 \to \infty} Var(r/n_1) = 0.$$

Proof: Wald and Wolfowitz [9] have proved this lemma when p = 1. Now $r = \Sigma r_i$, hence $E(r) = \Sigma E(r_i)$, and d) follows. Further, $Var(r) = \Sigma Var(r_i) - \sum_{i \neq j} Cov(r_i, r_j)$ and $(Cov(r_i, r_j))^2 \leq Var(r_i)Var(r_j)$, so e) follows.

Lemma 2: If a), b) and c) of Lemma 1 are fulfilled and if (F,G) is in the set C, then

$$\sum_{t=1}^{p} \int_0^1 \frac{g_t(x)}{\alpha + g_t(x)} \, dx < \frac{p}{1 + \alpha}$$

Proof: Wald and Wolfowitz have proved this lemma when p = 1. Note that if $f_t = g_t$, then the integral is equal to $1/(1 + \alpha)$. But since (F,G) is in C, there is at least one t such that $f_t \neq g_t$, and the integral is less than $1/(1 + \alpha)$. Hence the sum of the integrals must be less than $p/(1 + \alpha)$.

Proof of Theorem V: Let $\delta_1 > \delta_2 > \ldots$ be an arbitrary but fixed sequence such that $\lim \delta_j = 0$. For each δ_j and for each t = 1, ..., p, let $I_{t1}, \ldots, I_{tk_t(j)}$ be a set of closed intervals with disjoint interiors, and within which, by condition A, $f_t(x)$ and $g_t(x)$ exist, are positive, and continuous. Let I_t be the union of the I_{ti}, i = 1, ..., $k_t(j)$. Let I_{t0j} be the complement of I_t with respect to the

real line. Let r_{ti}, $i = 1, \ldots, k_t(j)$ and r_{t0j} denote, respectively, the runs caused by the observations which fall in the intervals I_{ti} and I_{t0j}. Then

$$\left| r^t - \sum_{i=1}^{k_t(j)} r_{ti} - r_{t0j} \right| \leq 2k_t(j).$$

Note that for n_1 sufficiently large, $r_{t0j} < 3n_1\delta_j/p$ with probability arbitrarily close to one, for each j, since F and G satisfy condition A.

Consider the interval $I_{ti} = [a_{ti}, b_{ti}]$, and let n_{1ti} and n_{2ti} denote, respectively, the number of observations on X and Y which fall in I_{ti}. Now the probability that m of the x_t's are less than c is given by $\binom{n_1}{m} (F_t(c))^m (1 - F_t(c))^{n_1-m}$. This is the binomial distribution with parameters n_1 and $F_t(c)$. So $E(m/n_1)$ converges stochastically to $F_t(c)$. Hence n_{1ti}/n_1 must converge stochastically to $F_t(b_{ti}) - F_t(a_{ti})$; and similarly, n_{2ti}/n_2 converges stochastically to $G_t(b_{ti}) - G_t(a_{ti})$.

Within the intervals I_{ti}, perform the transformation $x_t^* = F_t(x_t)$, $y_t^* = F_t(y_t)$, $t = 1, \ldots, p$, $i = 1, \ldots, k_t(j)$. This leaves the number of runs invariant since by condition

A, F_t is strictly monotone in the I_{ti}. For n_{1ti}, n_{2ti} fixed, the derived distribution of x* satisfies a) of Lemma 1 and the derived distribution of y* satisfies b) and c) of Lemma 1, since F, G satisfy condition A. Further $E(r_{ti}/n_1) = E(r_{ti}/n_{1ti})(n_{1ti}/n_1)$ so $E(r_{ti}/n_1) \leq 2/(n_{1ti}/n_{2ti} + 1)(n_{1ti}/n_1) = 2n_{1ti}n_{2ti}/(n_1(n_{1ti} + n_{2ti}))$ with probability arbitrarily close to one for n_1 large enough, by Lemmas 1 and 2. Note that by satisfying Lemma 1, r_{ti}/n_{1ti} converges stochastically to its expected value.

$$\text{So } \lim_{n_1 \to \infty} E(r_{ti}/n_1) \leq \lim_{n_1 \to \infty} \frac{2(n_{1ti}/n_1)(n_{2ti}/n_2)}{n_{1ti}/n_1 + n_{2ti}/n_2} =$$

$$(10) \quad \frac{2(F_t(b_{ti}) - F_t(a_{ti}))(G_t(b_{ti}) - G_t(a_{ti}))}{\alpha(F_t(b_{ti}) - F_t(a_{ti})) + G_t(b_{ti}) - G_t(a_{ti})}$$

, in probability.

$$\text{Thus } \lim_{n_1 \to \infty} E\left(\sum_{t=1}^{p} r_{ti}/n_1\right) \leq$$

$$(11) \quad \sum_{t=1}^{p} \frac{2(F_t(b_{ti}) - F_t(a_{ti}))(G_t(b_{ti}) - G_t(a_{ti}))}{\alpha(F_t(b_{ti}) - F_t(a_{ti})) + G_t(b_{ti}) - G_t(a_{ti})}$$

with probability arbitrarily close to one for n_1 large enough.

Now notice that (10) above =

$$(12)\quad 2\,\frac{\dfrac{F_t(b_{ti}) - F_t(a_{ti})}{b_{ti} - a_{ti}}\;\dfrac{G_t(b_{ti}) - G_t(a_{ti})}{b_{ti} - a_{ti}}}{\alpha\dfrac{F_t(b_{ti}) - F_t(a_{ti})}{b_{ti} - a_{ti}} + \dfrac{G_t(b_{ti}) - G_t(a_{ti})}{b_{ti} - a_{ti}}}\,(b_{ti} - a_{ti})$$

so the I_{ti} can be chosen so that the sum over i of (12) is arbitrarily close to

$$(13)\qquad 2\int_{-\infty}^{\infty} \frac{f_t(x)g_t(x)}{g_t(x) + \alpha f_t(x)}\,dx$$

by the usual definitions of derivative and integral. Applying the transformation $x_t^* = F_t(x_t)$, $y_t^* = F_t(y_t)$ to the integral (13) gives

$$2\int_0^1 \frac{g_t(F_t^{-1}(y^*))}{\alpha f_t(F_t^{-1}(y^*)) + g_t(F_t^{-1}(y^*))}\,dy^* =$$

$$2\int_0^1 \frac{h_t(y^*)}{\alpha + h_t(y^*)}\,dy^* < 2/(1 + \alpha),$$ by Lemma 2, where

$h_t(y^*) = g_t(F_t^{-1}(y^*))/f_t(F_t^{-1}(y^*))$. (Notice that h_t is a density function.) So the I_{ti}'s can be chosen such that the sum over all i of (10) above is less than $2/(1 + \alpha)$.

Hence

$$\sum_{t=1}^{p} \sum_{i=1}^{k_t(j)} \lim_{n_1 \to \infty} E(r_{ti}/n_1) < 2p/(1+\alpha) \text{ with probability}$$

arbitrarily close to one. Further, since (F,G) is in C,

$$S = \sum_{t=1}^{p} \lim_{k_t(j) \to \infty} \sum_{i=1}^{k_t(j)} \lim_{n_1 \to \infty} E(r_{ti}/n_1) < 2p/(1+\alpha)$$

with probability arbitrarily close to one, since the integral in (13) is less than $2/(1+\alpha)$ and the limit S is the integral in (13) summed p times.

Now choose ϵ such that $0 < 3\epsilon < 2p/(1+\alpha) - S$ and j so that $3\delta_j < \epsilon$. Then $r/n_1 =$

$$\sum_{t=1}^{p} r^t/n_1 \leq \sum_{t=1}^{p} 2k_t(j)/n_1 + \sum_{t=1}^{p} \sum_{i=1}^{k_t(j)} r_{ti}/n_1$$

$$+ \sum_{t=1}^{p} r_{t0j}/n_1 < \epsilon + S + \epsilon = S + 3\epsilon - \epsilon < 2p/(1+\alpha) - \epsilon$$

with probability arbitrarily close to one for n_1 large enough.

Now consider the random variable $x = (r - E(r))/[\mathrm{Var}(r)]^{\frac{1}{2}}$. The critical value for a test based on this quantity is $x_0 = (r_0(n_1) - E(r))/[\mathrm{Var}(r)]^{\frac{1}{2}}$. Hence $r_0(n_1) = x_0[\mathrm{Var}(r)]^{\frac{1}{2}} + E(r) = x_0[\mathrm{Var}(r)]^{\frac{1}{2}} + 2pne_1e_2$. Dividing by n_1 and taking limits as n_1 increases in size given $0 + 2pe_2 =$

$2pn_2/n = 2p/(1 + \alpha)$, using Lemma 1. Hence, for n_1 sufficiently large, with probability arbitrarily close to one, $r/n_1 < r_0(n_1)/n_1$. So $\lim_{n_1 \to \infty} P(r < r_0(n_1)) = 1$. This completes the proof of Theorem V.

IV. CONSISTENCY OF THE RUNS TEST IN THE CASE OF MORE THAN TWO SAMPLES WITH P = 1

The consistency of the runs test will now be proven when k is larger than two. This property will be shown to hold first for univariate distributions in a certain general class.

Definition 9: The distribution f is said to satisfy condition B if for arbitrarily small positive δ there exists a finite number of closed intervals such that the union, I, of these intervals has probability greater than $1 - \delta$ under f, and I is such that f has positive and continuous first derivatives f' in I.

In the following, there are k samples of size n_i, $i = 1, \ldots, k$; let $n = \Sigma n_i$; it will be assumed that as the n_i, increase in size, the ratios $n_i/n_1 = \alpha_{i1}$ are constant. (Hence $\alpha_{ij} = n_i/n_j$ are constant.) Let $r_0(n_1)$ be the critical value of the total number of runs, r, for the desired level of the test. Consider the following theorem.

Theorem VI: If $f_i(x)$, $i = 1, \ldots, k$, satisfies condition B, and if $f_i(x) \neq f_j(x)$ for some $i \neq j$, then

$$\lim_{n_1 \to \infty} P(r < r_0(n_1)) = 1.$$

The proof will be given in several stages.

Lemma 3: If the following are fulfilled:

a) $f_1(x) = 0$ if $x < 0$,

$= x$ if $0 \leq x \leq 1$,

$= 1$ if $x > 1$;

b) $f_i(x) = 0$ if $x \leq 0$

$= 1$ if $x \geq 0$, for $i > 1$;

c) The derivatives $f_i'(x)$ of $f_i(x)$, i greater than one, each exist, are continuous, and positive everywhere in $(0,1)$;

then:

d) $$\lim_{n_1 \to \infty} E(r_1/n_1) = \int_0^1 \frac{\sum_{i=2}^{k} \alpha_{i2} f_i'(x)}{\sum_{i=1}^{k} \alpha_{i2} f_i'(x)} \, dx;$$

and

e) $$\lim_{n_1 \to \infty} \mathrm{Var}(r_1/n_1) = 0.$$

Proof: Notice that $h(x) = \sum_{i=2}^{k} n_i f_i(x)/(n - n_1)$ is a distribution function satisfying conditions b) and c) of Lemma 1, since $f_i(x)$, $i = 2, \ldots, p$, satisfy conditions b) and c) of Lemma 3. Hence there is a sample of size n_1 from f_1, and $n - n_1$ from h. So, from Lemma 1, and since $\lim_{n_1 \to \infty} E(r_1/n_1) = \frac{1}{2} \lim_{n_1 \to \infty} E(r/n_1)$, because r_1 and $r - r_1$

differ at most by one, $\lim_{n_1 \to \infty} E(r_1/n_1) =$

$\int_0^1 \frac{h'(x)}{n_1/(n - n_1) + h'(x)}\, dx$. But $h'(x) = \sum_{i=2}^{k} n_i f_i'(x)/(n - n_1)$ and $f_1'(x) = 1$. Hence $\lim_{n_1 \to \infty} E(r_1/n_1) =$

$\int_0^1 \frac{\sum_{i=2}^{k} n_i f_i'(x)}{\sum_{i=1}^{k} n_i f_i'(x)}\, dx$. Now divide numerator and denominator by n_2, and d) follows. Similarly, since $r = 2r_1 \pm c$, $c = -1$, 0, 1, and from Lemma 1, $\lim_{n_1 \to \infty} \mathrm{Var}(r_1/n_1) = 0$. This completes the proof of Lemma 3.

It is evident that $\lim_{n_1 \to \infty} \mathrm{Var}(r/n_1) = 0$ as a consequence of Lemma 3(e) and the fact that $r = \Sigma r_i$. Also under the assumption that $f_i = f_j$ for each i and j, from Lemma 3(d), one has $\lim_{n_1 \to \infty} \Sigma E(r_j/n_j) = \sum_{i \neq j} \alpha_{i2}/\Sigma\alpha_{i2}$. Since $E(r/n_1) = \Sigma E(r_i/n_1) = \Sigma\alpha_{i1} E(r_i/n_i)$, then $\lim_{n_1 \to \infty} E(r/n_1) =$

$\Sigma\alpha_{j1}\left(\sum_{i \neq j} \alpha_{i2}/\Sigma\alpha_{s2}\right)$. Now using the relation $\alpha_{s2}/\alpha_{i2} = \alpha_{si}$,

$\lim_{n_1 \to \infty} E(r/n_1) = \sum_j \sum_{i \neq j} \alpha_{j1}/\Sigma\alpha_{si}$.

Lemma 4: If conditions a), b), and c) of Lemma 3 hold, and if $f_1(x) \neq f_j(x)$ for some j larger than one,

$$\int_0^1 \frac{\sum_{i=2}^{k} \alpha_{i2} f_i'(x)}{\sum_{i=1}^{k} \alpha_{i2} f_i'(x)}\, dx < \frac{n - n_1}{n} .$$

Proof: From Lemma 2, with $h(x) = \sum_{i=2}^{k} n_i f_i(x)/(n-n_1)$,

$\int_0^1 \frac{h'(x)}{n_1/(n-n_1)+h'(x)}\, dx < \frac{1}{1+n_1/(n-n_1)}$. But $h'(x) = \sum_{i=2}^{k} n_i f_i'(x)/(n - n_1)$ and $\frac{1}{1+n_1/(n - n_1)} = \frac{n - n_1}{n}$. Hence

$$\int_0^1 \frac{\sum_{i=2}^{k} n_i f_i'(x)/(n-n_1)}{\sum_{i=1}^{k} n_i f_i'(x)/(n-n_1)}\, dx = \int_0^1 \frac{\sum_{i=2}^{k} n_i f_i'(x)}{\sum_{i=1}^{k} n_i f_i'(x)}\, dx < \frac{n - n_1}{n} ,$$

since $f_1'(x) = 1$, and the lemma is proved.

Lemma 5: Suppose $f_i(x)$, $i = 1, \ldots, k$, satisfy condition B and $f_1(x) \neq f_j(x)$ for some $j \neq 1$. Then r_1/n_1 is less than $(n - n_1)/n - \epsilon$ for some $\epsilon > 0$, with probability arbitrarily close to one for sufficiently large n_1.

Proof: Let $\delta_1 < \delta_2 < \ldots$ be an arbitrary but fixed sequence with limit zero. For δ_j, let $I_1, \ldots, I_{k(j)}$ be a set of closed intervals with disjoint interiors and within which, by condition B, $f_i'(x)$ exist, are positive,

and continuous. Let I_{0j} be that part of the real line not contained in the union, I, of the I_s, s = 1, ..., k(j). Let r_{1s} and r_{10j} denote, respectively, the runs of $f_1(x)$ caused by the observations which fall in the intervals I_s, I_{0j}. Then

$$| r_1 - \sum_{s=1}^{k(j)} r_{is} - r_{10j} | \leq 2k(j).$$

From condition B it follows that with probability arbitrarily close to one, for sufficiently large n_1,

$$r_{10j} < 2n_1 \delta_j, \; j = 1, 2, \ldots \; .$$

Let $I_s = [a_s, b_s]$, s = 1, 2, ..., and let n_{is} denote the number of observations on $f_i(x)$ which fall in the interval I_s. Notice that n_{is}/n_i converges stochastically with increasing n_i to $f_i(b_s) - f_i(a_s)$. (See the proof of Theorem V).

Within the interval I_s perform the transformation $x_1^* = f_1(x)$, $x_i^* = f_1(x_i)$, where x_i is distributed as $f_i(x_i)$ for i bigger than one. Then for n_{is} fixed, the derived distribution of x_1^* is uniform on [0,1] and the derived distribution of x_i^* satisfies conditions b) and c) of

Lemma 4. Hence from Lemma 3, r_{1s}/n_1 converges stochastically to

$$\lim_{n_1 \to \infty} E(r_{1s}/n_1) \leq$$

$$(14) \qquad \frac{(f_1(b_s) - f_1(a_s)) \sum_{i=2}^{k} \alpha_{i2}(f_i(b_s) - f_i(a_s))}{\sum_{i=2}^{k} \alpha_{i2}(f_i(b_s) - f_i(a_s))}$$

since by Lemma 4,

$$\lim_{n_1 \to \infty} E(r_{1s}/n_1) < (n_s - n_{1s})n_{1s}/n_1 n_s.$$

Hence

$$\lim_{n_1 \to \infty} E(r_{1s}/n_1) \leq \lim_{n_1 \to \infty} (n_s - n_{1s})n_{1s}/n_1 n_s$$

$$= \frac{(f_1(b_s) - f_1(a_s)) \sum_{i=2}^{k} \alpha_{i2}(f_i(b_s) - f_i(a_s))}{\sum_{i=1}^{k} \alpha_{i2}(f_i(b_s) - f_i(a_s))}, \text{ in prob-}$$

ability. But the I_s can be chosen such that the sum over all s of (14) above is less than $(n - n_1)/n$, since the I_s can be chosen to make

$$S_1 = \sum_{s=1}^{\infty} \frac{(f_1(b_s) - f_1(a_s)) \sum_{i=2}^{k} \alpha_{i2}(f_i(b_s) - f_i(a_s))}{\sum_{i=1}^{k} \alpha_{i2}(f_i(b_s) - f_i(a_s))} =$$

$$\sum_{s=1}^{\infty} \frac{\dfrac{(f_1(b_s) - f_1(a_s))}{b_s - a_s} \quad \sum_{i=2}^{k} \dfrac{\alpha_{i2}(f_i(b_s) - f_i(a_s))}{b_s - a_s}}{\sum_{i=1}^{k} \alpha_{i2} \dfrac{(f_i(b_s) - f_i(a_s))}{b_s - a_s}} (b_s - a_s)$$

arbitrarily close to $\displaystyle\int_{-\infty}^{\infty} \frac{f_1'(x) \sum_{i=2}^{k} \alpha_{i2} f_i'(x)}{\sum_{i=1}^{k} \alpha_{i2} f_i'(x)} dx$

using the usual definitions of derivative and Riemann integral. Now applying the transformation $x^* = f_1(x)$ to the above integral, S_1 can be made arbitrarily close to

$$\int_0^1 \frac{f_1'(f_1^{-1}(x^*)) \sum_{i=2}^{k} \alpha_{i2} f_i'(f_1^{-1}(x^*))}{f_1'(f_1^{-1}(x^*)) \sum_{i=1}^{k} \alpha_{i2} f_i'(f_1^{-1}(x^*))} dx^*$$

$$= \int_0^1 \frac{\sum_{i=2}^{k} \alpha_{i2} g_i'(x)}{\sum_{i=1}^{k} \alpha_{i2} g_i'(x)} dx < \frac{n - n_1}{n},$$

where $g_i(x) = f_i'(f_1^{-1}(x))$, by Lemma 4.

Now choose j such that $\delta_j < \epsilon$ where $0 < 3\epsilon < (n - n_1)/n - S_1$. Then, since $r_1/n_1 \leq 2k(j)/n_1 +$

$\sum_{s=1}^{k(j)} r_{1s}/n_1 + r_{10j}/n_1$, and since r_{1s}/n_1 converges stochastically to its expected value by Lemma 3, $r_1/n_1 < S_1 + 2\epsilon < (n - n_1)/n - 3\epsilon + 2\epsilon = (n - n_1)/n - \epsilon$.

Hence with probability arbitrarily close to one, r_1/n_1 is less than $(n - n_1)/n - \epsilon$, for n_1 large enough, and the lemma is proved.

Proof of Theorem VI: From Lemma 5, for large enough n_i, with probabilities arbitrarily close to one, r_i/n_i is less than $(n - n_i)/n - \epsilon/k$. Now $r/n_1 = \sum_{i=1}^{k} r_i/n_1 = \sum_{i=1}^{k} \alpha_{i1} r_i/n_i < \sum_{i=1}^{k} \alpha_{i1}(n - n_i)/n - \epsilon$. Now $(n - n_i)/n = \sum_{j\neq i} n_j/n = \sum_{j\neq i} 1/\Sigma_s (n_s/n_j) = \sum_{j\neq i} 1/\Sigma_s \alpha_{sj}$. Hence $r/n_1 < \sum_{i=1}^{k} \sum_{j\neq 1}^{k} \alpha_{i1}/(\sum_{s=1}^{n} \alpha_{sj}) - \epsilon$ with probability arbitrarily close to one for large enough n_1.

Now consider $X = (r - E(r))/[Var(r)]^{\frac{1}{2}}$. The critical value for a test based on this statistic is $X_0 = (r_0(n_1) - E(r))/[Var(r)]^{\frac{1}{2}}$. As a consequence of Lemma 3, the follow-

ing may be found;

$$\lim_{n_1 \to \infty} \frac{r_0(n_1)}{n_1} = \lim_{n_1 \to \infty} \frac{X_0[\mathrm{Var}(r)]^{\frac{1}{2}}}{n_1} + \lim_{n_1 \to \infty} \frac{E(r)}{n_1} = 0 +$$

$$\sum_j \sum_{i \neq j} \alpha_{j1} / \sum_s \alpha_{si}.$$

So for n_1 sufficiently large, the probability is arbitrarily close to one that r is less than $r_0(n_1)$, i.e. $\lim_{n_1 \to \infty} P(r < r_0(n_1)) = 1$, if $f_i \neq f_j$ for some $i \neq j$; thus the proof of Theorem VI is completed.