SOME TESTS BASED ON DICHOTOMIZATION

By Nils Blomqvist

University of Stockholm and Boston University

1. Summary. Some methods for testing independence between the components of a random vector are discussed. The basic principle in the construction of the tests is dichotomization of each component variable. The distributions are obtained under randomization. Other applications of the tests are mentioned (Section 2). Certain limiting distributions are derived (Section 4). The exact distribution of the test statistic in a special case is tabulated (Section 5). A brief study of an alternative test is made (Section 6).

2. Introduction.

Consider a random sample of n vectors from an m-dimensional population with unknown distribution. It is desired to have a nonparametric test of independence between the m random variables. Solutions applicable to this problem were given by E. J. G. Pitman [1], who studied the conditional distribution of a certain statistic under permutations of the actual observations, and by M. Friedman [2] and M. G. Kendall and B. Babington Smith [3] using the method of ranks. At least in the latter method the absence of ties is essential, so that the observations of each component variable can be ordered. The present paper deals with the opposite situation, where the n observations of each variable are so heavily tied that it is possible to distinguish only between two groups, one higher and one lower, say. The situation can also be described as a dichotomization of distinct observations in order to simplify calculations, in situations where such simplifications (and loss of efficiency) can be afforded.

If all observations belonging to a higher group are replaced by scores one and the others by scores zero, and if the numbers of score one are n_1, n_2, $\cdots$ n_m $(0 < n_i < n, i = 1, 2, \cdots m)$, respectively, then the observations may after the dichotomization be represented by the following matrix

$$(1)\qquad \begin{array}{lccccc} & & & & & \textit{totals} \\ & x_{11} & x_{12} & \cdots & x_{1m} & y_1 \\ & x_{21} & x_{22} & \cdots & x_{2m} & y_2 \\ & \cdot & \cdot & & & \cdot \\ & \cdot & \cdot & & & \cdot \\ & \cdot & \cdot & & & \cdot \\ & x_{n1} & x_{n2} & \cdots & x_{nm} & y_n \\ \textit{totals} & n_1 & n_2 & \cdots & n_m & \end{array}$$

where each x_{ij} is equal to either one or zero.

Since the sample is assumed to be random, all $\binom{n}{n_j}$ different assignments of scores in the jth $(j = 1, 2, \cdots, m)$ column of (1) have the same probability. Denote the common expectation of the y's by $P = p_1 + p_2 + \cdots + p_m$, where

$n_j = np_j$ ($j = 1, 2, \cdots, m$). Under the (null) hypothesis of independence we expect all y's to be in the neighborhood of P. It seems, therefore, appropriate to base a test of independence upon the deviations from these expected values. Accordingly, we define

$$S = \sum_{i=1}^{n} (y_i - P)^2$$

as a test function and consider large values significant.

It should be observed that the test based on S can be applied also in situations other than the one considered above. First, when independence between the column vectors is assumed, we make the null hypothesis that all assignments of scores in a column are equally probable. In such a case we are dealing with tests of homogeneity between the rows of matrix (1). This situation has been considered by W. G. Cochran [4], whose statistic Q is the same as the one in Theorem 4 of the present paper. Cochran gave a nonrigorous proof of this theorem. For the sake of completeness the rigorous proof is given here. Secondly, let the mn values be observations in a two-way classification with one observation in each cell and assume that there are no column effects, in the sense of the analysis of variance. After choosing appropriate p-values the test S can be used to test the absence of row effects. This problem was studied by A. M. Mood and G. W. Brown ([5], p. 399) in the important special case when all p-values equal $\frac{1}{2}$, that is, when each column is dichotomized by the median. Theorems 3 and 4 in this paper are generalizations of results already given in [5], p. 399.

The p-values are considered nonrandom. This seems to be a proper assumption in the continuous case, since it is always possible to have the columns in (1) dichotomized by fixed quantiles. In the discrete case the test will be conditioned by this assumption.

No attempts have been made in this paper to investigate the power of tests considered. Consequently, all statements refer to the situation when the null hypothesis holds true.

3. Basic covariances. Before entering into discussion of the tests based on S we will introduce some notation and give some results for the case $m = 2$.

Let

$$t_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik}$$

be the number of rows in the matrix (1) having score one both in the jth and in the kth column and define

$$q_{jk} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - p_j)\cdot(x_{ik} - p_k) = \frac{t_{jk}}{n} - p_jp_k \tag{2}$$

as the basic covariance between the two columns.

In a previous paper [6] the author has studied a test of independence between two random variables, based on the statistic q_{12}, for the special case $p_1 = p_2 = \frac{1}{2}$. Some of these results are generalized here.

Dichotomization of the jth and kth column is equivalent to constructing the following 2 × 2 table:

Number of scores

Column k Column j	0	1	Totals
0	$n - n_j - n_k + \nu$	$n_k - \nu$	$n - n_j$
1	$n_j - \nu$	ν	n_j
Totals..........	$n - n_k$	n_k	n

From this table the exact distribution of any t_{jk} and, consequently, q_{jk} is obtained:

$$P\{t_{jk} = \nu\} = \frac{\binom{n_j}{\nu} \cdot \binom{n - n_j}{n_k - \nu}}{\binom{n}{n_k}}, \tag{3}$$

where

$$\max \{0, n_j + n_k - n\} \leq \nu \leq \min \{n_j \, , n_k\}.$$

The first two moments of the distribution of q_{jk} are obtained from (3):

$$\begin{aligned} E(q_{jk}) &= 0, \\ \sigma^2(q_{jk}) &= p_j p_k (1 - p_j)(1 - p_k)/(n - 1). \end{aligned} \tag{4}$$

From formula (3) also some asymptotic forms of the distribution of q_{jk} (and t_{jk}) can be derived. The derivations are straightforward applications of Stirling's formula, wherefore we will give only the final results here.

THEOREM 1. *If p_j and p_k remain fixed as $n \rightarrow \infty$, then q_{jk} has in the limit a normal distribution with mean and standard deviation as given in* (4).

THEOREM 2. *If $\sqrt{n}p_j \rightarrow \lambda_j$ and $\sqrt{n}p_k \rightarrow \lambda_k$ as $n \rightarrow \infty$, then t_{jk} has in the limit a Poisson distribution with parameter $\lambda_j \lambda_k$.*

4. Limiting distributions of S. We shall proceed to study the asymptotic behavior of the S-distribution under various assumptions regarding m, n and the p-values. From a practical point of view the case of large n should be most important when we are dealing with tests of independence between the columns in (1). In the case of testing for homogeneity between rows, however, large m-values become of main interest. Accordingly, we shall investigate both these cases and also the case when the p-values are small and n large.

THEOREM 3. *If m and all p-values remain fixed as $n \rightarrow \infty$, then S is asymptotically normally distributed with mean*

$$n \sum_{j=1}^{m} p_j(1 - p_j)$$

and variance

$$\frac{4n^2}{n-1}\sum_{j<k} p_j p_k(1-p_j)(1-p_k).$$

PROOF. It follows from the definition of S and q_{jk} that

$$\begin{aligned} S &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{m}(x_{ij}-p_j)\right\}^2 \\ &= n\sum_{j=1}^{m} p_j(1-p_j) + 2n\sum_{j<k} q_{jk}, \end{aligned} \tag{5}$$

which proves that S essentially depends only upon the sum of the basic covariances. Furthermore,

$$\begin{aligned} L(q_{12}, \cdots q_{m-1,m}) &= \sum_{j<k}\sqrt{p_jp_k(1-p_j)(1-p_k)}\cdot\frac{q_{jk}\sqrt{n-1}}{\sqrt{p_jp_k(1-p_j)(1-p_k)}} \\ &= \sqrt{n-1}\,.\sum_{j<k} q_{jk} \end{aligned}$$

is a fixed (in n) linear form in the random variables

$$\frac{q_{jk}\sqrt{n-1}}{\sqrt{p_jp_k(1-p_j)(1-p_k)}} \qquad (j, k = 1, 2, \cdots m; j \neq k).$$

For large n these variables, according to Theorem 1, tend in probability to standardized normal variables. The linear form L tends to the same linear form in the limiting variables. Hence L is in the limit normally distributed. It is readily seen that the variables q_{jk} are pairwise independent, from which it follows that also the limiting variables are pairwise independent. Consequently,

$$\begin{aligned} E(L) &= 0, \\ \sigma^2(L) &= \sum_{j<k} p_jp_k(1-p_j)(1-p_k) \end{aligned}$$

also in the limit. The theorem follows from (5).

THEOREM 4. *If n and all p-values remain fixed as $m \to \infty$, then*

$$\frac{n-1}{n}\cdot\frac{S}{\sum_{j=1}^{m} p_j(1-p_j)}$$

has in the limit a χ^2-distribution with $n-1$ degrees of freedom, subject to the condition that

$$\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m} p_j(1-p_j) \neq 0.$$

PROOF. The random vector $(y_1 - P, y_2 - P, \cdots, y_n - P)$ is a sum of m independent vectors with zero mean vectors and variances and covariances

$$\mu_{ii}^{(j)} = E(x_{ij}-p_j)^2 = p_j(1-p_j) \qquad (i = 1, 2, \cdots, n),$$

$$\mu_{ik}^{(j)} = E(x_{ij}-p_j)(x_{kj}-p_j) = -\frac{p_j(1-p_j)}{n-1} \qquad (i, k = 1, 2, \cdots, n; i \neq k; \; j = 1, 2, \cdots, m).$$

Since all these vectors are uniformly bounded with probability one, the Lindeberg condition for the generalized central limit theorem ([7], p. 113) is fulfilled. It follows that the vector

$$\frac{1}{\sqrt{m}}(y_1 - P, y_2 - P, \cdots, y_n - P)$$

has a limiting n-dimensional normal distribution with zero mean vector and variances and covariances

$$\mu'_{ii} = A \qquad (i = 1, 2, \cdots, \mathrm{n}),$$
$$\mu'_{ik} = -A/(n-1) \qquad (i, k = 1, 2, \cdots, n; i \neq k),$$

where

$$A = \lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} p_j(1 - p_j).$$

(From the assumptions made in Section 2 it follows that $A \neq 0$.) Hence the vector

$$\sqrt{\frac{n-1}{Amn}}(y_1 - P, y_2 - P, \cdots, y_n - P)$$

has a limiting normal distribution with zero mean vector and variances and covariances

$$(6) \qquad \begin{aligned} \mu_{ii} &= (n-1)/n && (i = 1, 2, \cdots, n),\\ \mu_{ik} &= -1/n && (i, k = 1, 2, \cdots, n; i \neq k). \end{aligned}$$

The covariance matrix constructed from (6) has $n - 1$ characteristic numbers equal to one and one characteristic number equal to zero. Consequently ([8], p. 314),

$$\frac{n-1}{Amn} \sum_{i=1}^{n} (y_i - P)^2 = \frac{(n-1)S}{Amn}$$

has a limiting χ^2-distribution with $n - 1$ degrees of freedom, which proves the theorem.

The next theorem concerns the case when all p-values tend to zero as n approaches infinity. A practical example where this situation becomes of interest is the following. Suppose that a large number (n) of persons are given some (m) psychological tests and we want to investigate whether or not these tests are independent of each other. For each person (i) and test (j), passing ($x_{ij} = 1$) or not passing ($x_{ij} = 0$) is registered. If then the number (n_j) of persons passing the jth test ($j = 1, 2, \cdots, m$) is small as compared with n, it seems preferable to use the approach of the following theorem to that of Theorem 4.

Since now most of the y-values in the sum column of matrix (1) will equal zero or one, it seems intuitively desirable to transform the test function S in such a way that the few y-values exceeding one are separated. Putting

$$T = \sum_{j<k} t_{jk},$$

we may write the expansion (5)

$$S = nP(1 - P) + 2T.$$

T depends only upon those y-values that exceed one, which follows from the fact that, if the number of y's equal to i is denoted by r_i, then

$$T = \sum_{i=2}^{m} \binom{i}{2} r_i, \tag{7}$$

a formula that can be used in computing T. We proceed to prove the following theorem on the limiting distribution of T, when the p-values tend to zero as $n^{-1/2}$.

THEOREM 5. *If m remains fixed and $\sqrt{n}\, p_j \to \lambda_j$ $(j = 1, 2, \cdots, m)$ as $n \to \infty$, then T has in the limit a Poisson distribution with parameter $\sum_{j<k} \lambda_j \lambda_k$*

PROOF. Let $T_k = \sum_{j=1}^{k-1} t_{jk}$. Then

$$T = T_2 + T_3 + \cdots + T_m. \tag{8}$$

The main part of the proof of the theorem will be to show that, if we add an extra column to the matrix (1), then T_{m+1} is in the limit Poisson distributed and independent of T.

As above, let $r_i(i = 0, 1, \cdots, m)$ be the number of y's equal to i. It is easily seen that T_{m+1} depends upon T only through the r-values. Since the columns of matrix (1) are independent, the conditional distribution of T_{m+1}, given $r_0, r_1, \cdots, r_m$, is

$$P\{T_{m+1} \mid r_0, r_1, \cdots, r_m\} = \sum \binom{n}{n_{m+1}}^{-1} \prod_{i=0}^{m} \binom{r_i}{x_i}, \tag{9}$$

where the summation is extended over all x's such that

$$\begin{aligned} &\sum_{i=0}^{m} x_i = n_{m+1}, \\ &\sum_{i=0}^{m} i x_i = T_{m+1}, \\ &0 \leq x_i \leq r_i \qquad\qquad (i = 0, 1, \cdots, m). \end{aligned} \tag{10}$$

We now let $n \to \infty$ and $n_j/\sqrt{n} \to \lambda_j(j = 1, 2, \cdots, m)$ while m and T_{m+1} remain fixed. Because of (10) $x_1, x_2, \cdots, x_m$ are bounded and $x_0 = n_{m+1} - \sum_{i=1}^{m} x_i$ is of the order $\sqrt{n}$.

Because of (7), $r_2, r_3, \cdots, r_m$ are bounded. Furthermore, since

$$\sum_{i=0}^{m} r_i = n,$$

$$\sum_{i=0}^{m} i r_i = \sum_{j=1}^{m} n_j,$$

it is true that r_0 is of the order n and r_1 of the order $\sqrt{n}$. Applying Stirling's formula to the summand in (9) we obtain after some calculations

$$\lim_{n\to\infty} \binom{n}{n_{m+1}}^{-1} \prod_{i=0}^{m} \binom{r_i}{x_i} = \begin{cases} e^{-\Lambda} \dfrac{\Lambda^{T_{m+1}}}{T_{m+1}!} & \text{for} \quad x_1 = T_{m+1}, x_2 = x_3 = \cdots = x_m = 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\Lambda = \lambda_{m+1} \sum_{j=1}^{m} \lambda_j .$$

Hence, according to (9),

$$\lim_{n\to\infty} P\{T_{m+1} \mid r_0, r_1, \cdots, r_m\} = e^{-\Lambda} \frac{\Lambda^{T_{m+1}}}{T_{m+1}!} .$$

Since the r-values completely determine T, it follows that T_{m+1} is in the limit independent of T and Poisson distributed with parameter Λ. Applying Theorem

TABLE I

$P\{S \geq S_0\}$

	$n=4$		
S_0 \ m	4	6	8
0	1.000	1.000	1.000
2	.931	.960	.973
4	.486	.651	.745
6	.356	.549	.663
8	.134	.302	.435
10	.0787	.225	.355
12		.117	.237
14		.102	.213
16	.0046	.0400	.109
18		.0315	.0971
20		.0109	.0563
22			.0379
24			.0299
26		.0032	.0239
30			.0078
32			.0030
34			.0025
36		.0001	.0017
38			.0013
40			.0005
50			.0001

	$n=4$		
S_0 \ m	3	5	7
1	1.000	1.000	1.000
3	.583	.761	.840
5	.361	.576	.696
9	.0278	.236	.399
11		.109	.237
13		.0471	.147
17		.0162	.0896
19			.0410
21			.0290
25		.0008	.0110
27			.0060
29			.0024
37			.0006

TABLE I—*Continued*

$n=6$	
S_0 \ m	4
0	1.000
2	.988
4	.813
6	.543
8	.358
10	.179
12	.0710
14	.0350
16	.0080
18	.0046
24	.0001

$n=6$		
S_0 \ m	3	5
1.5	1.000	1.000
3.5	.768	.907
5.5	.498	.752
7.5	.160	.489
9.5	.070	.383
11.5		.209
13.5	.0025	.132
15.5		.0675
17.5		.0360
19.5		.0113
21.5		.0090
23.5		.0031
25.5		.0008
29.5		.0003

$n=8$		
S_0 \ m	3	4
0	1.000	1.000
2	1.000	.998
4	.870	.947
6	.634	.759
8	.316	.572
10	.140	.353
12	.0296	.208
14	.0100	.0912
16		.0474
18	.0002	.0156
20		.0055
22		.0020
24		.0003
26		.0002

TABLE I—*Continued*

$n=10$	
S_0 \ m	4
0	1.000
2	1.000
4	.987
6	.903
8	.747
10	.562
12	.370
14	.221
16	.117
18	.0552
20	.0249
22	.0084
24	.0037
26	.0009
28	.0003
30	.0001

$n=10$	
S_0 \ m	3
2.5	1.000
4.5	.927
6.5	.747
8.5	.467
10.5	.241
12.5	.0847
14.5	.0280
16.5	.0043
18.5	.0012

$n=12$	
S_0 \ m	3
3	1.000
5	.959
7	.831
9	.600
11	.359
13	.165
15	.0639
17	.0172
19	.0045
21	.0005
23	.0001

$n=14$	
S_0 \ m	3
3.5	1.000
5.5	.977
7.5	.890
9.5	.709
11.5	.479
13.5	.264
15.5	.121
17.5	.0432
19.5	.0133
21.5	.0029
23.5	.0006
25.5	.0001

$n=16$	
S_0 \ m	3
0	1.000
2	1.000
4	1.000
6	.987
8	.929
10	.793
12	.591
14	.372
16	.197
18	.0858
20	.0314
22	.0091
24	.0023
26	.0004
28	.0001

2 for the case $m = 2$, we now proceed step by step to obtain the desired result that T is in the limit Poisson distributed with parameter

$$\sum_{k=2}^{m} \lambda_k \sum_{l=1}^{k-1} \lambda_l = \sum_{j<k} \lambda_j \lambda_k .$$

This completes the proof of the theorem.

5. The exact distribution of S in a special case. In the important special case when all p-values are equal to $\frac{1}{2}$ it follows from Theorem 3 and 4 that *S is asymptotically normally distributed with mean $mn/4$ and variance $n^2 \cdot m(m-1)/8(n-1)$ as $n \to \infty$, and that $4(n-1)S/mn$ is asymptotically χ^2-distributed with $n-1$ degrees of freedom as $m \to \infty$.* These limiting distributions may be used as approximations when n or m is large. For small values of n and m the exact distribution of S is needed. This is given in Table I for the following cases:

TABLE II

Comparison between exact and approximate distribution of S at the 5% point

n	m	S_0	$P\{S \geq S_0\}$		
			exact	χ^2-approx.	normal approx.
4	8	22	.038	*.044*	.017
6	5	17.5	.036	*.042*	.017
8	4	16	.048	*.050*	.029
10	4	20	.025	.038	*.014*
12	3	17	.017	.039	*.013*
14	3	17.5	.043	.063	*.037*
16	3	20	.031	.050	*.025*

n	4	6	8	10	12	14	16
m	3–8	3–5	3,4	3,4	3	3	3 .

The case $n = 2$ needs no consideration here since the test then reduces to the standard sign test. The case $m = 2$ is also excluded since it has already been tabulated in [6].

In Table II some comparisons are made between the exact distribution given in Table I and the approximations given in the paragraph above. The normal approximation has been applied after usual correction for continuity. For each pair (n, m) in Table II the best approximation has been underlined, which might serve as a guide in the choice of appropriate approximation.

6. Another test. Although it has not been mentioned explicitly against what alternative hypothesis the S-test is designed, it is clear that we have had in mind the case when all m component variables of the random vector studied are positively correlated. We do not intend to enter into a detailed discussion of the difficult question of alternatives. However, one case more shall be briefly mentioned. If about half of the variables are positively correlated with each other but negatively correlated with the rest of the variables, it is intuitively seen that the S-test will lose its power. Instead, a test should be used that is not based upon the algebraic sum of the basic covariances (5), but takes into account their absolute values. For example, a test based on the sum of the squares of the basic covariances might serve our purpose. In this connection we shall prove the following limiting theorem.

THEOREM 6. *If m and all p-values remain fixed as $n \to \infty$, then*

$$\sum_{j<k} \frac{(n-1)q_{jk}^2}{p_j p_k (1-p_j)(1-p_k)}$$

has in the limit a χ^2-distribution with $\binom{m}{2}$ degrees of freedom.

PROOF. In the proof of Theorem 3 it was stated that the random variables

$$\frac{q_{jk}\sqrt{n-1}}{\sqrt{p_j p_k (1-p_j)(1-p_k)}} \qquad (j, k = 1, 2, \cdots, m; j \neq k)$$

are pairwise independent and, as $n \to \infty$, normally distributed with zero mean and unit standard deviation. Hence, in the limit they are totally independent. The theorem follows.

The author expresses his indebtedness to Professor Frederick Mosteller of Harvard University for suggesting the original problem and for many helpful discussions. Miss Elizabeth Shuhany of Boston University has kindly assisted in the construction of tables.

REFERENCES

[1] E. J. G. PITMAN, "Significance tests which may be applied to samples from any populations; Part III. The analysis of variance," *Biometrika*, Vol. 29 (1938), p. 322–335.

[2] M. FRIEDMAN, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Jour. Am. Stat. Assn.*, Vol. 32 (1937), p. 675–701.

[3] M. G. KENDALL AND B. BABINGTON SMITH, "The problem of m rankings," *Annals of Math. Stat.*, Vol. 10 (1939), p. 275–287.

[4] W. G. COCHRAN, "The comparison of percentages in matched samples," *Biometrika*, Vol. 37 (1950), p. 256–266.

[5] A. M. MOOD, *Introduction to the Theory of Statistics*, McGraw-Hill Book Co., New York, 1950.

[6] N. BLOMQVIST, "On a measure of dependence between two random variables," *Annals of Math. Stat.*, Vol. 21 (1950), p. 593–601.

[7] H. CRAMÉR, *Random variables and probability distributions*, Cambridge University Press, 1937.

[8] H. CRAMÉR, *Mathematical methods of statistics*, Princeton University Press, 1946.