ON A MEASURE OF DEPENDENCE BETWEEN TWO RANDOM VARIABLES

By Nils Blomqvist

University of Stockholm and Boston University

1. Summary. The properties of a measure of dependence q' between two random variables are studied. It is shown (Sections 3–5) that q' under fairly general conditions has an asymptotically normal distribution and provides approximate confidence limits for the population analogue of q'. A test of independence based on q' is non-parametric (Section 6), and its asymptotic efficiency in the normal case is about 41% (Section 7). The q'-distribution in the case of independence is tabulated for sample sizes up to 50.

2. Introduction and definitions. In drawing conclusions from statistical data it frequently happens that it is unnecessary to utilize all the information given by the data. In such cases it seems desirable to use methods which are

1) valid under rather weak assumptions regarding the distribution of the population and

2) easy to deal with in practice.

Naturally such methods should always be used, but their applicability is, in most cases, limited by their small efficiency.

Concerning methods of measuring correlation and testing independence some so-called rank correlation coefficients have been defined [2, 3, 4, 6] which have the first property. In large samples these are, however, rather tiresome to calculate, and a simpler method might then be preferable. The coefficient studied here has in most cases both properties mentioned above and can be used whenever its efficiency is not too small.

Let $(x_1, y_1) \cdots (x_n, y_n)$ be a sample from a two-dimensional population with cdf $F(x, y)$, and consider the two sample medians $\tilde{x}$ and $\tilde{y}$. The cdf $F(x, y)$ is assumed to have continuous marginal cdf's $F_1(x)$ and $F_2(y)$ in order that the probability of obtaining two equal x-values or two equal y-values in the sample will be zero. Let the x, y-plane be divided into four regions by the lines $x = \tilde{x}$ and $y = \tilde{y}$. It is then clear that some information about the correlation between x and y can be obtained from the number of sample points, say n_1, belonging to the first or third quadrants compared with the number, say n_2, belonging to the second or fourth quadrants.

Before going further we shall explain what is meant here by 'belong to'. If the sample size n is an even number the calculation of n_1 and n_2 is evident. If, however, n is an odd number one or two sample points must fall on the lines $x = \tilde{x}$ and $y = \tilde{y}$. In the first case this sample point shall not be counted. In the other case one point falls on each of the lines. Then one of the points shall be said to belong to the quadrant touched by both points, while the other shall

not be counted. It is easy to verify that both n_1 and n_2 by this method will be even numbers.

As a measure of correlation we define

$$q' = \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1 \qquad (-1 \leq q' \leq 1). \tag{1}$$

The definition of q' is not new [5] but as far as is known, its statistical properties have never been studied completely.

3. The asymptotic distribution. It is known [1] that the median in a sample from a one-dimensional distribution under certain conditions is a consistent estimate of the population median and asymptotically normally distributed. Although it seems possible to weaken the requirements in our case, we shall not do so. We require that

a) the population medians are uniquely defined (and assumed to equal zero),

b) the marginal distributions of $F(x, y)$ admit density functions $f_1(x)$ and $f_2(y)$.

c) $f_1(x), f_2(y)$ and their first derivatives are continuous in some neighbourhood of the origin and

d) $f_1(0)$ and $f_2(0)$ are $\neq 0$.

In order to avoid trivial complications we shall assume here that the sample size $n = 2k + 1$.

Now define for every arbitrarily chosen point (x, y)

$$\begin{aligned} a(x, y) &= P\{\xi > x, \eta > y\}, \\ b(x, y) &= P\{\xi \leq x, \eta > y\}, \\ c(x, y) &= P\{\xi \leq x, \eta \leq y\}, \\ d(x, y) &= P\{\xi > x, \eta \leq y\}, \end{aligned} \tag{2}$$

where the measure P refers to the cdf $F(x, y)$ and evidently

$$a + b + c + d = 1.$$

As the number of sample points belonging to the first and third quadrants around $(\tilde{x}, \tilde{y})$ must be equal, the probability of the combined event

$$\{n_1 = 2r; \tilde{x} \epsilon (x, x + dx), \tilde{y} \epsilon (y, y + dy)\}$$

is

$$p_k(2r; x, y) = \frac{(2k + 1)!}{r!^2 \cdot (k - r)!^2} \cdot (ac)^r \cdot (bd)^{k-r} \cdot S, \tag{3}$$

where

$$S = \frac{r}{a} \cdot d_x a \cdot d_y a - \frac{k - r}{b} \cdot d_x b \cdot d_y b + \frac{r}{c} \cdot d_x c \cdot d_y c - \frac{k - r}{d} \cdot d_x d \cdot d_y d + dF. \tag{4}$$

Each of the first four terms of the expression (4) refers to a case in which two sample points determine $(\tilde{x}, \tilde{y})$, and the last term refers to a case in which $(\tilde{x}, \tilde{y})$ is determined by only one point. From (3) it follows that the probability of obtaining n_1 at most equal to $2R$ is

$$P\{n_1 \leq 2R\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sum_{r=0}^{R} p_k(2r; x, y). \tag{5}$$

If we introduce the joint cdf $\Psi_k(x, y)$ of $\tilde{x}$ and $\tilde{y}$, (5) can be written

$$P\{n_1 \leq 2R\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\Psi_k(x, y) \frac{\sum_{r=0}^{R} p_k(2r; x, y)}{\sum_{r=0}^{k} p_k(2r; x, y)}, \tag{6}$$

as

$$d\Psi_k(x, y) = \sum_{r=0}^{k} p_k(2r; x, y).$$

Clearly the integrand in (6) is ≤ 1 everywhere it exists. In the points (x, y) where the denominator is equal to zero the integrand is undefined, but as the measure (Ψ) of the set of such points is zero, we need not have any trouble with them.

Under the conditions a)–d) $\tilde{x}$ and $\tilde{y}$ converge in probability to zero; that is

$$\lim_{k\to\infty} \Psi_k(x, y) = \begin{cases} 1 \text{ for } \{x \geq 0, y \geq 0\}, \\ 0 \text{ otherwise.} \end{cases}$$

Thus, when k and R tend to infinity such that $\frac{R}{k} \to \text{const}$, (6) becomes

$$\lim P\{n_1 \leq 2R\} = \lim \frac{\sum_{r=0}^{R} p_k(2r; 0, 0)}{\sum_{r=0}^{k} p_k(2r; 0, 0)}. \tag{7}$$

According to (3)

$$p_k(2r; 0, 0) = \frac{(2k+1)!}{r!^2 \cdot (k-r)!^2} \cdot (a_0 c_0)^r \cdot (b_0 d_0)^{k-r} \cdot S_0, \tag{8}$$

where the subscripts indicate the value at the point (0, 0). Because of (2),

$$c_0 = a_0, \qquad d_0 = b_0 \quad \text{and} \quad a_0 + b_0 = \tfrac{1}{2},$$

and the two parts of (8) are for large k

$$\frac{(2k+1)!}{r!^2 \cdot (k-r)!^2} \cdot a_0^{2r} \cdot b_0^{2(k-r)} \sim \frac{1}{2\pi a_0 b_0 \sqrt{2\pi k}} \cdot e^{-((r-2ka_0)^2/4ka_0b_0)}$$

and

$$S_0 \sim 2k\left[\left(\frac{\partial a}{\partial x}\right)_0\left(\frac{\partial a}{\partial y}\right)_0 - \left(\frac{\partial b}{\partial x}\right)_0\left(\frac{\partial b}{\partial y}\right)_0 + \left(\frac{\partial c}{\partial x}\right)_0\left(\frac{\partial c}{\partial y}\right)_0 - \left(\frac{\partial d}{\partial x}\right)_0\left(\frac{\partial d}{\partial y}\right)_0\right] dx\, dy.$$

The first of these expressions follows from the usual application of Stirling's approximation formula and we omit all details here.

Hence, after the introduction of

$$r = 2ka_0 + t\sqrt{2ka_0b_0},$$
$$R = 2ka_0 + T\sqrt{2ka_0b_0},$$

the expression (7) is transformed to

$$\lim P\left\{\frac{n_1 - 4ka_0}{\sqrt{8ka_0b_0}} \leq T\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{T} e^{-\frac{1}{2}t^2}\, dt. \tag{9}$$

From (9) it follows that n_1 is asymptotically normally distributed with mean $4ka_0$ and standard deviation $\sqrt{8ka_0b_0}$. Thus

$$q' = \frac{2n_1}{2k} - 1 = \frac{n_1}{k} - 1$$

is asymptotically normally distributed with mean $4a_0 - 1$ and standard deviation $2\sqrt{a_0(1 - 2a_0)/k}$.

4. Properties as an estimator. Suppose we measure the correlation between x and y by

$$q = 2\left[\int_{-\infty}^{0}\int_{-\infty}^{0} dF + \int_{0}^{\infty}\int_{0}^{\infty} dF\right] - 1 = 4a_0 - 1, \tag{10}$$

where, as before, (0, 0) are the coordinates of the population medians. Then q has the desired property of being equal to zero in the case of independence and equal to ± 1 in the case of linear relationship between x and y.

According to (9) q' is a consistent estimate of q when the conditions a)–d) are fulfilled. Furthermore, as the standard deviation of q' is, to a first approximation, independent of quantities other than q, it is possible to construct approximate confidence limits for q for large sample sizes. This is done in the following way. In terms of n and q we have, according to the last paragraph of section 3 and (10),

$$Eq' \sim q,$$
$$\sigma(q') \sim \sqrt{\frac{1 - q^2}{n}}.$$

Let $\Phi(x)$ be a standardized normal cdf and λ_1 and λ_2 two numbers such that

$\Phi(\lambda_2) - \Phi(\lambda_1) = 1 - \alpha$. According to (9) we then have

$$P\left\{\lambda_1 < \frac{q' - q}{\sqrt{1 - q^2}} \cdot \sqrt{n} < \lambda_2\right\} \sim 1 - \alpha, \tag{11}$$

which gives the desired result.

If we let $\lambda_2 = -\lambda_1 = \lambda$ and solve the inequality in (11) for q, the following symmetrical confidence interval is obtained

$$q' - \frac{\lambda}{n}\sqrt{\lambda^2 + n(1 - q'^2)} < q < q' + \frac{\lambda}{n}\sqrt{\lambda^2 + n(1 - q'^2)},$$

where we have used that $\lambda^2 \ll n$.

5. The normal case. If x and y are normally distributed with correlation coefficient ρ, we have

$$q = \frac{2}{\pi} \arcsin \rho. \tag{12}$$

This expression is the same as the mean of Esscher-Kendall's rank correlation coefficient τ [2, 4]. Hence, in the normal case q' and τ estimate the same quantity. The coefficient q' has, however, a much smaller efficiency. The asymptotic efficiency of q' relative to the afore mentioned coefficient is

$$\frac{\sigma^2(\tau)}{\sigma^2(q')} \sim \frac{\frac{4}{n} \cdot \left[\frac{1}{9} - \left(\frac{2}{\pi} \arcsin \frac{\rho}{2}\right)^2\right]}{\frac{1}{n} \cdot \left[1 - \left(\frac{2}{\pi} \arcsin \rho\right)^2\right]} = \frac{4}{9}$$

for $\rho = 0$.

6. Tests of independence based on q'. In testing independence between x and y it is in practice more convenient to use critical regions based on n_1 instead of q'. Since, under the null hypothesis, the measure of a critical region is independent of $F(x, y)$ ($F_1(x)$ and $F_2(y)$ are assumed to be continuous), any test based on n_1 is non-parametric. We have made exact calculations of the q'-distribution for sample sizes n up to 50. For larger sample sizes the normal approximation for n_1 does not seem to entail errors of practical importance.

To derive the exact distribution of n_1 under the null hypothesis we suppose that n equals $2k$. The probability that any k sample points shall have smaller x-values than the other k points is

$$\binom{2k}{k}^{-1}.$$

Hence, since any arrangement of the sample points according to their x-values does not affect the distribution of the y-values,

$$P\{n_1 = 2r\} = \frac{\binom{k}{r}^2}{\binom{2k}{k}}. \tag{13}$$

If $n = 2k + 1$ it is easily verified that the probability (13) remains unchanged, if we use the procedure in calculating n_1 and n_2 proposed in Section 2. This is, in fact, the main reason for the proposal.

Table of $P\{|\, n_1 - k \,| \geq \nu\}$

ν \ $2k$	4	8	12	16	20	24	28	32	36	40	44	48
0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	.333	.486	.567	.619	.656	.684	.706	.724	.740	.752	.764	.773
4		.029	.080	.132	.179	.220	.257	.289	.318	.343	.366	.387
6			.0022	.010	.023	.039	.057	.076	.094	.113	.131	.148
8				.0002	.0011	.0033	.0070	.012	.018	.026	.034	.042
10						.0001	.0004	.0011	.0022	.0038	.0060	.0087
12									.0002	.0004	.0007	.0012
14											.0001	.0001

ν \ $2k$	6	10	14	18	22	26	30	34	38	42	46	50
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	.100	.206	.286	.347	.395	.434	.466	.494	.517	.538	.556	.572
5		.0079	.029	.057	.086	.115	.143	.169	.194	.217	.238	.258
7			.0006	.0034	.0089	.017	.027	.038	.050	.063	.076	.089
9					.0003	.0012	.0028	.0053	.0086	.013	.017	.023
11							.0001	.0004	.0009	.0017	.0028	.0042
13									.0001	.0001	.0003	.0005
15												

$2k$ is the largest even number contained in the sample size.

The distribution of n_1 is symmetric about $n_1 = k$ with the variance

$$\frac{k^2}{2k - 1}.$$

Thus, in testing independence we can for large sample sizes use

$$\frac{n_1 - k}{\sqrt{k}} \cdot \sqrt{2k - 1}$$

as an approximately normally distributed random variable with mean zero and unit s.d.

7. The asymptotic efficiency of the q'-test. In the case that x and y are normally distributed with the correlation coefficient ρ, it is possible, but rather tedious, to calculate the power function of the q'-test. We will, therefore, restrict ourselves to considering only the asymptotic behavior of the power function.

Consider tests of independence ($\rho = 0$) against one-sided alternatives $\rho > 0$. Let $L_m^{(1)}(\rho)$ be the power function of the q'-test for the sample size m and $L_n^{(2)}(\rho)$ be the power function of the test based on the correlation coefficient r in a sample of size n. We assume that all tests have the same size, i.e.

$$L_m^{(1)}(0) = L_n^{(2)}(0) = \alpha \tag{14}$$

for all m and n. We shall say that the q'-test has the asymptotic efficiency ϵ if

$$\lim_{n\to\infty} \frac{\left(\frac{\partial L^{(1)}}{\partial \rho}\right)_{\rho=0}}{\left(\frac{\partial L^{(2)}}{\partial \rho}\right)_{\rho=0}} = 1 \tag{15}$$

when

$$m = \frac{n}{\epsilon}.$$

This means that the sample size in using the r-test need only be $100\epsilon\%$ of that in using the q'-test, in order to get the same derivative of the power functions at $\rho = 0$ (for large sample sizes). Since the definition of ϵ only concerns the behavior in the neighborhood of $\rho = 0$, it might perhaps be more correct to call ϵ the asymptotic local efficiency.

In order to calculate ϵ we define two sequences $\{q_m\}$ and $\{r_n\}$ such that $\{q' > q_m\}$ and $\{r > r_n\}$ are tests with the afore mentioned properties. According to (9) and (10) q' is asymptotically normally distributed with mean q and s.d. $\sqrt{(1-q^2)/m}$. Furthermore, r is asymptotically normally distributed with mean ρ and s.d. $(1-\rho^2)/\sqrt{n}$. Hence,

$$1 - L_m^{(1)}(\rho) = P\{q' \leq q_m \mid \rho\} \sim \Phi\left[\frac{q_m - q}{\sqrt{1-q^2}}\sqrt{m}\right],$$

$$1 - L_n^{(2)}(\rho) = P\{r \leq r_n \mid \rho\} \sim \Phi\left[\frac{r_n - \rho}{1-\rho^2}\cdot\sqrt{n}\right],$$

from which it follows

$$\begin{aligned}\left(\frac{\partial L^{(1)}}{\partial \rho}\right)_0 &\sim \Phi'(q_m\cdot\sqrt{m})\cdot\left(\frac{dq}{d\rho}\right)_0\sqrt{m},\\ \left(\frac{\partial L^{(2)}}{\partial \rho}\right)_0 &\sim \Phi'(r_n\cdot\sqrt{n})\cdot\sqrt{n}.\end{aligned} \tag{16}$$

According to (14) we have

$$\lim_{m\to\infty} q_m\cdot\sqrt{m} = \lim_{n\to\infty} r_n\cdot\sqrt{n} = \Phi^{-1}(1-\alpha).$$

Thus we conclude

$$\lim_{m,n\to\infty} \frac{\left(\frac{\partial L^{(1)}}{\partial \rho}\right)_0}{\left(\frac{\partial L^{(2)}}{\partial \rho}\right)_0} = \lim_{m,n\to\infty}\left(\frac{dq}{d\rho}\right)_0\cdot\sqrt{\frac{m}{n}}. \tag{17}$$

Clearly (17) is equal to 1 if

$$\frac{n}{m} = \left(\frac{dq}{d\rho}\right)_0^2.$$

Hence, according to (12) and (15)

$$\epsilon = \left(\frac{2}{\pi}\right)^2.$$

In other words, the asymptotic efficiency of the q'-test is about 41%.

8. Concluding remarks. An interesting similarity exists between the q'-test of independence and a test of equal location parameters in two distributions, constructed in the following way. Suppose that two samples of equal size, say k, are drawn independently from two distributions. Compute the number of individuals, say r, in the first sample, falling short of the median of the pooled samples. Then the distribution of $2r$ under the null hypothesis is the same as that of n_1 in the q'-test for sample size $2k$ (or $2k + 1$). The test based on r was discussed by F. Mosteller [7].

Another similarity is between the q'-test and a special case of the exact test of independence in a 2 x 2 table [8]. If in such a table the marginals happen to be cut at the 50% points the two test procedures become identical.

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.
[2] F. Esscher, "On a method of determining correlation from the ranks of a variate", *Skandinavisk Aktuarietidskrift*, Vol. 7 (1924), p. 201.
[3] W. Hoeffding, "A non-parametric test of independence", *Annals of Math. Stat.*, Vol. 19 (1948), p. 546.
[4] M. G. Kendall, "A new measure of rank correlation", *Biometrika*, Vol. 30 (1938), p. 81.
[5] F. Mosteller, "On some useful 'inefficient' statistics", *Annals of Math. Stat.*, Vol. 17 (1946), p. 377.
[6] C. Spearman, "The proof and measurement of association between two things", *Am. Jour. of Psych.*, Vol. 15 (1904), p. 88.
[7] F. Mosteller, "On some useful 'inefficient' statistics", unpublished thesis, Princeton University, 1946.
[8] R. A. Fisher, *Statistical Methods for Research Workers*, 8th Ed, Stechert & Co., 1941.