ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
Hugo Steinhaus Jubilee Volume
X (1969)

Z. HELLWIG (Wrocław)

ON THE MEASUREMENT OF STOCHASTICAL DEPENDENCE

1. Introductory remarks. In 1956, at the request of the clothing industry, an anthropometric survey was carried out in Poland. The main goal of that survey was the determination of a set of statistical individuals, called "phantoms", which could serve as good patterns for fitting ready-made clothes.

The survey was carried out under Professor Hugo Steinhaus who acted as a general consultant and scientific supervisor of the whole operation.

During one of the seminars devoted to the optimal choice of statistical variates Professor Steinhaus pointed out some disadvantages of the correlation coefficient and expressed the opinion that any properly defined measure of stochastical dependence should meet the following requirements.

1. The measure of dependence, which will be called a coefficient of dependence, should be equal 0 when there is no stochastical dependence between the variables involved, and equal 1 when the dependence is perfect.

2. The coefficient should not lose its applicability in the non-linear case of dependence.

3. The knowledge of the distribution (or at least of the limit distribution) of the sample coefficient of dependence would have been extremely desirable.

4. The coefficient ought to preserve its validity in the case of discrete as well as continuous variables. In other words the coefficient should possess its applicability in the case of twofold classification (when the coefficient of association is usually applied), in the case of manifold classification, and in the case when the bivariate population is continuous.

5. The functional relationship between the *proposed* coefficient and the *classical* product correlation coefficient has to be examined, given the distribution of the universe is normal.

6. The proposed measure of dependence should possess two additional very important properties:

6a) the concept of this measure must be theoretically simple, liable to easy explanation, popularization and wide application;

6b) the numerical and computational aspects connected with the calculation of the coefficient of dependence should also be relatively simple.

Since 1955 the author has undertaken many attempts to find a measure of dependence possessing at least some of the properties listed above. The first result of this effort was presented in [1], the second, more developed, has taken the form of an unpublished paper "Coefficient of Dependence".[1]

2. Measures of dependence and their properties. Let us divide all measures of stochastical dependence into two groups: parametrical and non-parametrical ones, the first being related to the random variables themselves and the latter to the distributions of the random variables. To the first group belongs the correlation coefficient ϱ, the correlation ratio η, the "generalized correlation coefficient" R [1], Spearman's rank correlation coefficient R, Kendall's rank correlation coefficient τ; to the second group may be included Yule's coefficient of association Q, Pearson's coefficient of association η, Yule's "coefficient of colligation" Y, Pearson's contingency χ^2, Pearson's coefficient of mean square contingency C, and Tschuprow's coefficient T.[2]

The most important measure of dependence from the first group is beyond doubt Pearson's coefficient of correlation ϱ. This measure however has three important disadvantages:

1. The lack of correlation is not equivalent to the lack of stochastical dependence.[3]

2. It can be applied only when the regression function is linear.

3. The exact distribution of the sample correlation coefficient r is known only if the joint distribution of the random vector (X, Y) is normal.

[1] Prepared especially for the members of a seminar led by the author during his stay as a visiting senior lecturer at the University of Ibadan (Nigeria). By now the author has collected a large scope of empirical examples and practical experience providing him with a reason to believe that some of this results are worth publishing. The author's conviction is primarely due to the fact that the coefficient of dependence possesses at least one significant feature: it is easy to calculate and to apply.

[2] A detailed discussion of measures of stochastical relationship between random variables may be found in [2], where also numerous bibliographical references are included (worth readers' attention not only from a professional but also from a historical point of view).

[3] In other words, from the fact $\varrho(X,Y) = 0$ does not follow necessarily that $f(x,y) = f_1(x)\cdot f_2(y)$ (see p. 235).

The most important measure of stochastical dependence in the second group is Pearson's square contingency χ^2. This measure has the two following disadvantages:

1. It cannot be applied to measure the dependence between the two random variables X and Y if at least one of the variables is continuous.

2. It does confine within the limits 0 and 1.

It is worth mentioning that in exact sciences like physics, astronomy or chemistry only one measure of dependence is sufficient e.g. the coefficient of correlation because when proving the existence of dependence a *large scale experiment* can be applied. The situation is not so comfortable in biological and yet more in social sciences where artificial experimentation is very inconvenient, dangerous, costly, or sometimes impossible at all. In such a situation a statistical approach is of great importance, especially if there exist many competitive methods and if the indication of one method could be controlled by the other.

The measure of dependence we would like to present in this paper belongs to the second group. It is free of the disadvantages mentioned above. As will be shown by means of numerical examples, it is easy to compute and to apply.

3a. Coefficient of stochastical dependence (continuous case). Let (X, Y) be a random vector and $F(x, y)$ its distribution function. If there exists a density function $f(x, y)$ which is continuous almost everywhere then (X, Y) is said to be continuous. In this case the sufficient and necessary condition for X and Y being stochastically independent can be expressed in the form

$$f(x,y) = f_1(x)\cdot f_2(y) \tag{1}$$

where

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \quad \text{and} \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y)\,dx.$$

Starting from (1) we define the *coefficient of stochastical dependence beetwen X and Y*, called shortly the *coefficient of dependence* and denoted by d:

$$d = \sqrt{1 - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min[f(x, y), f_1(x)\cdot f_2(y)]\,dxdy} \tag{2}$$

It is easy to see that $0 \leqslant d \leqslant 1$.

3b. Coefficient of dependence (discrete case). In the previous section a two-dimensional random variable has been considered, assumed to be a continuous one. Now we are going to pay our attention to the case of a discrete random variable (X, Y) with a finite set of possible values. If X is stochastically independent of Y or, what is tantamount, Y is independent of X then

$$P(X=x,\, Y=y) = P(X=x)\cdot P(Y=y); \qquad x\in\mathfrak{X}; \qquad y\in\mathfrak{Y}$$

where $\mathfrak{X} = (x_1, x_2, \ldots, x_r)$ and $\mathfrak{Y} = (y_1, y_2, \ldots, y_s)$ are the sets of possible values of X and Y respectively.

For the sake of simplicity of notation we will apply the symbol p_{ij} instead of $P(X=x_i,\, Y=y_j)$ and the symbols p_i and q_j instead of $P(X=x_i)$ and $P(Y=y_j)$. Let

$$P_1 = \begin{bmatrix} p_{11} & \cdot & \cdot & \cdot & p_{1s} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{r1} & \cdot & \cdot & \cdot & p_{rs} \end{bmatrix} \tag{3}$$

and

$$P_2 = \begin{bmatrix} p_1q_1 & \cdot & \cdot & \cdot & p_1q_s \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ p_rq_1 & \cdot & \cdot & \cdot & p_rq_s \end{bmatrix}. \tag{4}$$

Suppose for a while that the coefficient of dependence has been defined by equation

$$D^2 = 1 - \sum_{i,j} \min(p_{ij}, p_iq_j). \tag{5}$$

This definition is an analogue of (2) from the previous section. We will now slightly change the external shape of (5) to make this formula more convenient for the purpose of numerical computation.

Let us denote by M the set of pairs (i,j) for which $p_{ij} > p_iq_j$, and by K the set of pairs (i,j) for which $p_{ij} \leqslant p_iq_j$. Note that

$$1 - \sum_{(i,j)\in M} p_iq_j = \sum_{(i,j)\in K} p_iq_j.$$

Hence

$$D^2 = 1 - \sum_{i,j} \min(p_{ij}, p_iq_j) = 1 - \sum_{(i,j)\in M} p_iq_j - \sum_{(i,j)\in K} p_{ij}$$

$$= \sum_{(i,j)\in K} p_iq_j - \sum_{(i,j)\in K} p_{ij} \tag{6}$$

$$= \sum_{(i,j)\in M} p_{ij} - \sum_{(i,j)\in M} p_iq_j. \tag{7}$$

Without any loss of generality we may assume that $s \geqslant r$ and that every row and every column of the matrix P_1 contains at least one positive element (otherwise the matrix could have been reduced to a smaller size, but again with no zero rows and columns).

Let us denote by k the number of pairs (i,j) in K and by m the number of pairs in M. Hence

$$k+m = rs; \quad 0 < k \leqslant rs; \quad 0 \leqslant m < rs.$$

If $k = rs$ then

$$p_{ij} = p_i q_j \quad \text{for all} \quad i,j.$$

In such a case

$$D^2 = \sum_{(i,j)\epsilon M} p_{ij} - \sum_{(i,j)\epsilon M} p_i q_j = 0.$$

It is slightly more difficult to show that $D^2 \leqslant 1$. Let us examine the difference (7). We shall prove the following

THEOREM 1.

$$\max\Big(\sum_{(i,j)\epsilon M} p_{ij} - \sum_{(i,j)\epsilon M} p_i q_j\Big) = 1-\frac{1}{r}.$$

The proof of the theorem is based on

LEMMA 1. *If every column of* $\boldsymbol{P}_1$ *contains only one element, then*

$$\sum_{(i,j)\epsilon M} p_i q_j \geqslant \frac{1}{r}. \tag{8}$$

The equality in (8) *holds if and only if* $p_1 = p_2 = \ldots = p_r = 1/r$.

To prove the lemma let us notice first that in this case M consists of exactly s pairs $[i(j),j]$, $j = 1,2,\ldots,s$, where $i(j)$ is the number of the row having a positive element $p_{i(j),j}$ in the j-th column of $\boldsymbol{P}_1$. Hence $q_j = p_{i(j),j}$ and

$$\sum_{(i,j)\epsilon M} p_i q_j = \sum_j p_{i(j)} q_j = \sum_j p_{i(j)} p_{i(j),j} = \sum_i \sum_{j\epsilon J_i} p_i p_{ij} = \sum_i p_i \sum_{j\epsilon J_i} p_{ij}$$

where $J_i = \{j: i(j) = i\}$. There is also $J_i = \{j: p_{ij} > 0\}$ and

$$\sum_{j\epsilon J_i} p_{ij} = \sum_j p_{ij} = p_i.$$

So we have

$$\sum_{(i,j)\epsilon M} p_i q_j = \sum_i p_i^2.$$

Since $\sum_i p_i^2$ is a convex symmetric function of the arguments $p_1, p_2, \ldots, p_r$ then for $\sum_i p_i = 1$ and $p_i > 0$ the minimal value is obtained for $p_1 = p_2 = \ldots = p_r = 1/r$. This ends the proof of the lemma.

Let us now go back to arbitrary matrices $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. By comparing appropriate elements of both matrices we can determine the set M. Let $\boldsymbol{A}$ be an $r\times s$ matrix, the non-zero elements of which are such p_{ij} which meet the requirement $(i,j)\epsilon M$. Suppose

$$\sum_{(i,j)\epsilon M} p_{ij} = a; \qquad 0 < a \leqslant 1.$$

We find now in every column of $\boldsymbol{A}$ a maximum element and set up from these elements an $r\times s$ matrix $\boldsymbol{A}_1$, all other elements being zeros. In exactly the same way we determine matrix $\boldsymbol{A}_2$ composed of maximum elements taken out of the columns of $\boldsymbol{A}$ after having erased these elements which had been exploited to set up $\boldsymbol{A}_1$. Repeating this procedure as long as possible we obtain a sequence of matrices $\boldsymbol{A}_l, l = 1, 2, \ldots, n$, where $1 \leqslant n \leqslant r$.

It follows from the definition of M that

$$p_i = \sum_j p_{ij} \geqslant \sum_{\substack{j\\(ij)\epsilon M}} p_{ij} = p_i^*$$

and

$$q_j = \sum_i p_{ij} \geqslant \sum_{\substack{i\\(ij)\epsilon M}} p_{ij} = q_j^*.$$

Let us denote by $p_i^{(l)}$ the sum of the i-th row of matrix $\boldsymbol{A}_l (l = 1, 2, \ldots, n)$ and by $q_j^{(l)}$ the sum of the j-th column of that matrix. The following relations hold

$$p_i^* = \sum_{i=1}^{k} p_i^{(l)}; \qquad q_j^* = \sum_{l=1}^{k} q_j^{(l)}, \qquad p_i q_j \geqslant p_i^* q_j^* \geqslant \sum_{l=1}^{k} p_i^{(l)} q_j^{(l)}$$

for all i, j. Let $\sum_j q_j^{(l)} = a_l$. Then

$$\sum_{l=1}^{k} a_l = a;$$

and, in accordance with lemma 1, the following inequality holds

$$\sum_{(i,j)\epsilon M^{(l)}} p_i^{(l)} q_j^{(l)} \geqslant \frac{a_l^2}{r}.$$

where $M^{(l)} \subset M$, $\bigcup_l M^{(l)} = M$, $M^{(l)} \cap M^{(t)} = \emptyset$ for any $l, t = 1, 2, \ldots, n$.

Hence

$$\sum_{(i,j)\epsilon M} p_i q_j \geqslant \sum_{(i,j)\epsilon M} p_i^* q_j^* \geqslant \sum_{l=1}^{n} \sum_{(i,j)\epsilon M^{(l)}} p_i^{(l)} q_j^{(l)} \geqslant \frac{1}{r} \sum_{l=1}^{n} a_l^2 .$$

If so, then

$$\sum_{(i,j)\epsilon M} p_{ij} - \sum_{(i,j)\epsilon M} p_i q_j \leqslant a - \frac{1}{r} \sum_{l=1}^{n} a_l^2 .$$

But

$$a - \frac{1}{r} \sum_{l=1}^{n} a_l^2 \leqslant a - \frac{a^2}{nr},$$

and the function $f(a) = a - a^2/nr$, $0 \leqslant a \leqslant 1$, reaches its maximal value for $a = 1$. Hence

$$\sum_{(i,j)\epsilon M} p_{ij} - \sum_{(i,j)\epsilon M} p_i q_j \leqslant 1 - 1/nr .$$

Since

$$a = \sum_{(i,j)\epsilon M} p_{ij} = 1,$$

therefore $p_{ij} \geqslant p_i q_j \geqslant 1/rs$ for every $(i,j)\epsilon M$. On the other hand the smallest number ot elements in the matrix A, which, when cancelled out, turn A into A_1 with all elements zero, is equal s. Hence $\underset{(i,j)\epsilon M}{\Sigma} p_i q_j \geqslant 1/r$. This means that in the case when $\underset{(i,j)\epsilon M}{\Sigma} p_i q_j$ takes its minimal value, the number n in the inequality

$$\sum_{(i,j)\epsilon M} p_{ij} - \sum_{(i,j)\epsilon M} p_i q_j \leqslant 1 - 1/nr$$

must be equal 1. Hence we obtain finally

$$\max\Big(\sum_{(i,j)\epsilon M} p_{ij} - \sum_{(i,j)\epsilon M} p_i q_j \Big) = 1 - \frac{1}{r}$$

which ends the proof of the theorem 1.

Making use of the result of theorem 1 we introduce the following definition of the *coefficient of dependence for the discrete case*

$$(9) \qquad d = \frac{D}{\sqrt{1 - \dfrac{1}{\min(r,s)}}} = \sqrt{\frac{1 - \sum\limits_{ij} \min(p_{ij}, p_i q_j)}{1 - \dfrac{1}{\min(r,s)}}} .$$

It follows from theorem 1 that

$$0 \leqslant d^2 \leqslant 1 .$$

4. Basic properties of the coefficient of dependence. Two cases deserve a special attention: $d^2 = 0$ and $d^2 = 1$. We are going to examine both cases a bit closer.

THEOREM 2. *A necessary and sufficient condition for X and Y being stochastically independent is the equality $d^2 = 0$.*

The elementary proof will be carried out only for the case when the random variables are continuous. The discrete case is analogous.

1° Let us assume that X and Y are stochastically independent. Then

$$f(x,y) = f_1(x)f_2(y)$$

and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min[f(x,y), f_1(x)\cdot f_2(y)]dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dx\,dy = 1.$$

Hence

$$d^2 = 0.$$

2° Let $d^2 = 0$. Then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min[f(x,y), f_1(x)f_2(y)]dx\,dy = 1,$$

but

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x)f_2(y)dxdy = 1$$

and therefore

$$f(x,y) = f_1(x)f_2(y)$$

almost everywhere.

THEOREM 3. *If X and Y are continuous random variables then $d^2 = 1$ if and only if the whole mass of probability is spread out over an area with a plane measure equal to zero.*

Proof. Let us define the following sets

$$H = \{(x,y): f_1(x)f_2(y) > 0\},$$
$$G = \{(x,y): 0 < f_1(x)f_2(y) < f(x,y)\}.$$

Then

$$d^2 = 1 - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min[f(x,y), f_1(x)f_2(y)]dxdy$$
$$= 1 - \Big[\iint_G f_1(x)f_2(y)dxdy + \iint_{H-G} f(x,y)dxdy\Big].$$

Since both integrals in the last expression are nonnegative then for equality $d^2 = 1$ it is necessary and sufficient that

$$\iint_G f_1(x)f_2(y)dxdy = 0$$

and

$$\iint\limits_{H-G} f(x, y)\,dxdy = 0.$$

Hence $f(x, y) = 0$ almost everywhere, what ends the proof.

We know that if (X, Y) is a discrete random variable then $d^2 = 1$ if sums of all rows of the matrix P_1 are equal and if $\sum\limits_{(i,j)\in M} p_{ij} = 1$. We suspect that these requirements are also necessary conditions for d^2 being equal 1.'

Let us examine the case of special importance, when $r = s = 2$. It is easy to prove the following theorems:

1° $d^2 = 0$ if and only if

$$p_{ij} = p_i q_j \quad \text{for} \quad i, j = 1, 2;$$

2° $d^2 = 1$ if and only if sums of both rows are equal and if columns contain one and only one element. In other words $d^2 = 1$ if and only if the non-zero elements of the matrix are equal and belong to one of the diagonals of the matrix.

Let us now examine some numerical examples which will present not only all details of the computational technique but also will illustrate practical implications of indication of the coefficient of dependence.

EXAMPLE 1. Table 1 presents figures illustrating the distribution of alive birts in Poland in 1964 with regards to consecutive delivery and to the partition into country and towns (see [3], p. 51). Matrices P_1 and P_2 are both presented in one table (table 2). The upper figures are the elements of the matrix P_1, i.e. the upper figures stand for probabilities p_{ij}, whereas the lower ones are the elements of the matrix P_2 (in other words they are probabilities $p_i q_i$).

One should now compare numbers in every cell. The numbers of those cells where the inequalities

$$p_{ij} > p_i q_j$$

TABLE 1. Alive births in thousands by delivery number and area

Area	Consecutive delivery								Total
	1	2	3	4	5	6	7	8 and more	
Urban	99.9	75.3	33.2	14.9	7.1	3.7	1.9	2.1	238.1
Rural	91.5	83.3	61.3	38.4	22.1	12.4	7.2	8.5	324.7
Total	191.4	158.6	94.5	53.3	29.2	16.1	9.1	10.6	562.8

TABLE 2. Computational scheme for table 1

.1775 *.1439*	*.1338* *.1192*	.0590 .0710	.0265 .0401	.0126 .0219	.0066 .0121	.0034 .0068	.0037 .0080	.4231 .4230
.1626 .1962	.1480 .1626	*.1089* *.0969*	*.0682* *.0546*	*.0393* *.0299*	*.0220* *.0165*	*.0128* *.0093*	*.0151* *.0109*	.5769 .5769
.3401 .3401	.2818 .2818	.1679 .1679	.0947 .0947	.0519 .0518	.0286 .0286	.0162 .0161	.0188 .0189	1.0000 .9999

hold are printed in italics in table 2. One has

$$\sum_{(i,j)\in M} p_{ij} - \sum_{(i,j)\in M} p_i q_j = .5776-.4812 = .0964,$$

and thus

$$d^2 = \frac{.0964}{1-\frac{1}{2}} = .1928, \quad d = .44.$$

Example 2. A factory manufacturing radio and television sets has issued a pilot batch of television sets aiming to examine if newly designed boxes which television sets were equipped with would please potential buyers and eventually cause an increase of demand, despite a slight increase of the price of new sets. Table 3 shows results of this examination. The size of the tested lot was equal 1000 sets, 400 of which have been of the old type and the remainig 600 were equipped with the newly styled boxes.

TABLE 3. Demand for television-sets w ith respect to old and new types

	B new	$\bar{B}$ old	Σ
A sold	600	—	600
$\bar{A}$ unsold	—	400	400
Σ	600	400	1000

TABLE 4. Computational scheme for table 3.

	B	B	Σ
A	*0.60* *0.36*	0.00 0.24	0.60 0.60
$\bar{A}$	0.00 0.24	*0.40* *0.16*	0.40 0.40
Σ	0.60 0.60	0.40 0.40	1 1

It is obvious that a modern outlook of the television set could exert a sufficiently decisive influence on buyer and make them prefer new sets to the old ones.

One expects that any measure of stochastic dependence, if embraced within the limits $\langle 0, 1\rangle$, should in a situation, like the one described above, be close to 1. Let us check how the coefficient of dependence will behave in these circumstances. Simple calculation technique connected with the computation of the numerical value of this coefficient is presented in table 4.

All marked cells create a set M. Hence

$$d^2 = \frac{\sum\limits_{(i,j)\epsilon M} p_{ij} - \sum\limits_{(i,j)\epsilon M} p_i q_j}{1 - \dfrac{1}{\min(r, s)}} = \frac{1-0.52}{1-\dfrac{1}{2}} = .96.$$

REMARK. The round and easy to calculate figures of table 3 has been chosen on purpose, in order to facilitate computational work.˙ This, however, should nobody lead to the conclusion that the whole problem is just a classroom example. On the contrary, the problem itself stems from economic life and preserves its practical importance. We are now proceeding to present a slightly more sophisticated example based on true statistical data ([3], p. 50). One important feature of this example is worth emphasizing. It consists in the fact that both variables involved are continuous. Examples 1-3 show that the coefficient of dependence can be applied in all situations: when both variables are non-measurable, when one is measurable — the other not, when both are measurable and discrete, when both are measurable and continuous, and, finally, when one is discrete and one is continuous.

While presenting the next example no further explanatory comments will be made.

EXAMPLE 3. Table 5 presents figures illustrating a stochastical dependence between the age of the members of married couples in Poland in 1964.

TABLE 5. Members of married couples by age

Age of men	Age of women							
	19 years old or less	20-24	25-29	30-34	35-39	40-49	50 years old or more	Total
19 years old or less	7242	3030	400	54	19	3	1	10649
20-24	37724	47969	6843	1140	236	59	5	93976
25-29	17824	37629	14140	3501	971	218	13	74296
30-34	2351	7793	6823	4017	1627	522	34	23167
35-39	324	1607	2247	2472	1883	888	61	9482
40-49	82	410	861	1430	2114	2274	405	7576
50 years old or more	18	85	181	465	1055	3924	5777	11529
Total	65565	98523	31395	13079	7905	7912	6296	230675

Table 6 is a contingency table filled up with probabilities p_{ij} (upper part) and $p_i q_j$ (lower part). We call these figures "probabilities" although in fact they are not probabilities but relative frequencies, playing the role of estimates of unknown probabilities.

TABLE 6. Computational scheme for table 5

.0314 *.0131*	.0131 .0197	.0013 .0063	.0002 .0026	.0001 .0016	.0000 .0016	.0000 .0013	.0461 .0462
.1635 *.1158*	*.2080* *.1740*	.0297 .0554	.0049 .0231	.0010 .0140	.0003 .0140	.0000 .0111	.4074 .4074
.0773 .0915	*.1631* *.1376*	*.0613* *.0438*	.0152 .0183	.0042 .0110	.0009 .0110	.0001 .0088	.3221 .3220
.0102 .0285	.0338 .0429	*.0296* *.0137*	*.0174* *.0057*	*.0071* *.0034*	.0023 .0034	.0001 .0027	.1005 .1003
.0014 .0117	.0070 .0176	*.0097* *.0056*	*.0107* *.0023*	*.0082* *.0014*	*.0038* *.0014*	.0003 .0011	.0411 .0411
.0004 .0093	.0018 .0140	.0037 .0045	*.0062* *.0019*	*.0092* *.0011*	*.0099* *.0011*	*.0018* *.0009*	.0330 .0328
.0001 .0142	.0004 .0214	.0008 .0068	.0020 .0028	*.0046* *.0017*	*.0171* *.0017*	*.0250* *.0014*	.0500 .0500
.2843 .2841	.4272 .4272	.1361 .1361	.0566 .0567	.0344 .0342	.0343 .0342	.0273 .0273	1.0002 .9998

We have

$$\sum_{(i,j)\in M} p_{ij} = .7876 \quad \text{and} \quad \sum_{(i,j)\in M} p_i q_j = .5277,$$

thus giving

$$d^2 = \frac{.7876-.5277}{1-\dfrac{7}{49}} = \frac{.2599}{.857} = .3.$$

$$d = .55.$$

REMARK. The correlation coefficient would have taken in this case a much greater value, presumably close to 0.7 (see section 6, table 7).

5. Significance of the coefficient of dependence. Let us now suppose that the statistical data we have just used to calculate the value of the coefficient of dependence have been obtained from a sample drawn at random from a universe. In such circumstances the sample coefficient of dependence is a random variable. Unfortunately, we do not know the exact distribution of this variable. Nevertheless there exists a possibility to examine the significance of a sample coefficient of dependence, given the sample is sufficiently large, so that relative frequencies $\hat{p}_{ij}$ could serve as sufficiently accurate estimates of p_{ij}. If this condition has been observed then $\hat{p}_i = \Sigma\hat{p}_{ij}$ and $\hat{q}_j = \Sigma\hat{p}_{ij}$ are by far more accurate estimates of p_i and q_j. Hence we are fully justified to put

$$p_i = \hat{p}_i, \quad q_j = \hat{q}_j.$$

It is generally accepted in practice to consider the size n of the sample ω as being sufficiently large if the smallest frequency n_{ij} in the (i, j) cell of the $r \times s$ contingency table is not less than 5. It is worthwhile to take notice of the following equalities

$$\sum_{ij} n_{ij} = n; \quad \hat{p}_{ij} = \frac{n_{ij}}{n}; \quad \hat{p}_i = \sum_j \hat{p}_{ij} \; \hat{q}_j = \sum_i \hat{p}_{ij}.$$

Let us denote by $\hat{d}^2$ an estimate of d^2 determined on the basis of the data delivered by the sample ω, given the sample has been drawn out of the population Ω, the marginal distribution of which are the probabilities $p_i, q_j, i = 1, 2, \ldots, r; j = 1, 2, \ldots, s$.

Hence

$$(10) \qquad \hat{d}^2 = r \cdot \frac{\sum\limits_{(i,j)\epsilon M} \hat{p}_{ij} - \sum\limits_{(i,j)\epsilon M} p_i q_j}{r-1},$$

where $r = \min(r, s)$.

It should be emphasized that r, p_i, q_j are given constants and M is a given subset of pairs (i, j). This enables us to find the variance $V(\hat{d}^2)$. Let us put

$$\frac{r}{r-1} = c; \quad \sum_{(i,j)\epsilon M} \hat{p}_{ij} = \hat{Q}; \quad \sum_{(i,j)\epsilon M} p_i q_j = R.$$

Hence

$$(11) \qquad V(\hat{d}^2) = c^2 \cdot V(\hat{Q}) = c^2 \cdot \frac{Q(1-Q)}{n}$$

where $Q = \sum\limits_{(i,j)\epsilon M} p_{ij}$. If $n \to \infty$ then the distribution of the random variable $(\hat{d}^2 - E(\hat{d}^2))/\sqrt{[V(\hat{d}^2)]}$ tends to the normal distribution $N(0, 1)$.

We would like to draw the reader's attention to the fact that when applying formula (10) one should in the first step draw a large sample ω_1, determine numbers r, s, R the subset M; then, in the next step, draw the sample ω_2 and, using the previous partition of the contingency table into r rows and s columns, find the value of the random variable $\hat{Q}$, even if some (may be all) rows or columns of the table were empty. In practice we can make use of the data of the same sample ω in both steps given the sample ω is large.

Now we are able to write a confidence interval for the parameter d^2:

$$P\{\hat{d}^2 - t\sqrt{[V(\hat{d}^2)]} < d^2 < \hat{d}^2 + t\sqrt{[V(\hat{d}^2)]}\} = \alpha,$$

where α depends on t and t is a random variable normally distributed $N(0,1)$.

6. Mutual relation between the coefficient of dependence and the coefficient of correlation when the distribution of the random variable (X, Y) is normal. Let us assume that the density function $f(x,y)$ of the random variable (X, Y) is of the following form

$$(12)\quad f(x,y) = \frac{1}{2\pi\sqrt{(1-\varrho)^2}}\cdot\exp\left(-\frac{x^2+2\varrho xy+y^2}{2(1-\varrho)^2}\right).$$

It is easy to notice that (12) is the density function of a normal distribution, where ϱ stands for the coefficient of correlation, whereas $E(X) = E(Y) = 0$ and $V(X) = V(Y) = 1$. In such a case the coefficient of dependence is given by the formula

$$(13)\quad d^2 = \frac{1}{2\pi}\iint_G \left[\frac{1}{\sqrt{(1-\varrho^2)}}\times \times\exp\left(-\frac{x^2-2\varrho xy+y^2}{2(1-\varrho^2)}\right) - \exp\left(-\frac{x^2+y^2}{2}\right)\right] dxdy,$$

where G stands for an area such that $f(x,y) > f_1(x)f_2(y)$ if $(x,y)\epsilon G$.

It is easy to show that G is the area lying between two branches of a hyperbola. We would like to draw the reader's attention to the fact that the regression lines

$$y = \lambda\varrho x \quad \text{and} \quad x = \lambda\varrho y, \quad \text{where } \lambda = 1/(1+\sqrt{1-\varrho^2}),$$

play the role of asymptotes of this hyperbola.

TABLE 7. Coefficient of dependence *vs.* Coefficient of correlation in a two-dimensional normal population

ϱ	d^2	d
.05	.0160	.13
.10	.0321	.18
.15	.0485	.22
.20	.0654	.26
.25	.0828	.29
.30	.1010	.32
.35	.1201	.35
.40	.1402	.37
.45	.1616	.40
.50	.1846	.43
.55	.2095	.46
.60	.2367	.49
.65	.2669	.52
.70	.3008	.55
.75	.3397	.58
.80	.3856	.62
.85	.4417	.66
.90	.5143	.72
.92	.5522	.74

Formula (13) is an explicit expression of the mutual relationship between d and ϱ in the normal case. The integral (13) has been investigated by A. Smoluk. As yet all attempts, undertaken by him to change the double integral (13) into a form which could enable an analytic solution have failed. However A. Smoluk applied approximate methods and solved the problem numerically [4]. Similar results have been obtained by E. Trybuś, who applied the Monte-Carlo method for evaluating the integral (13). Table 7 gives the results of computations carried out by A. Smoluk.

Table 7 gives the possibility of evaluating the coefficient of correlation by means of the coefficient of dependence and vice versa.

References

[1] Z. Hellwig, *Aproksymacja stochastyczna* (Stochastic approximation), Warszawa 1965.

[2] M. G. Kendall, *The advanced theory od statistics*, London 1947.

[3] Rocznik Statystyczny (Polish Statistical Yearbook), Warszawa 1966.

[4] A. Smoluk, *O wskaźniku Hellwiga dla rozkładu normalnego* (On Hellwig's coefficient in the normal case), Prace Naukowe WSE Wrocław, no. 12 (1968).

DEPARTMENT OF STATISTICS AND ECONOMIC ACCOUNTING
GRADUATE SCHOOL OF ECONOMICS, WROCŁAW

Received on 17. 11. 1967