The modification of the Phi-coefficient reducing its dependence on the marginal distributions

Peter V. Zysno
RWTH Aachen University

Methods of Psychological Research Online 1997, Vol.2, No.1
Internet: http://www.pabst-publishers.de/mpr/

The Modification of the Phi-coefficient Reducing its Dependence on the Marginal Distributions

Peter V. Zysno

Abstract

The Phi-coefficient is a well known measure of correlation for dichotomous variables. It is worthy of remark, that the extreme values ±1 only occur in the case of consistent responses and symmetric marginal frequencies. Consequently low correlations may be due to either inconsistent data, unequal response frequencies or both. In order to overcome this somewhat confusing situation various alternative proposals were made, which generally, remained rather unsatisfactory. Here, first of all a system has been developed in order to evaluate these measures. Only one of the well-known coefficients satisfies the underlying demands. According to the criteria, the Phi-coefficient is accompanied by a formally similar modification, which is independent of the marginal frequency distributions. Based on actual data both of them can be easily computed. If the original data are not available – as usual in publications – but the intercorrelations and response frequencies of the variables are, then the grades of association for assymmetric distributions can be calculated subsequently.

Keywords: Phi-coefficient, independent marginal distributions, dichotomous variables

1 Introduction

In the beginning of this century the Phi-coefficient (Yule 1912) was developed as a correlational measure for dichotomous variables. Its essential features can be quickly outlined. N elements of two variables i and j fall into the classes "+" and "−" with frequencies n_i, $u_i = N - n_i$, n_j and $u_j = N - n_j$, respectively. Consequently, within a pair of variables four dyadic events may occur: $++, +-, -+$ and $--$. The corresponding joint frequencies are denoted by the initial letters of the alphabet a, b, c and d. The parameters are usually presented in a fourfold table (tab.1).

The degree of coherence between the two variables is defined as the quotient of two frequencies. The numerator represents the difference of the products of

Table 1: Fourfold table of the Phi-coefficient for two binary variables i and j with classes + and −.

		j		
		+	−	
i	+	a	b	n_i
	−	c	d	u_i
		n_j	u_j	N

Table 2: Phi-correlations of 6 Items with 3 levels of difficulty

	1	2	3	4	5	6
1	-	1.00	0.43	0.43	0.18	0.18
2	1.00	-	0.43	0.43	0.18	0.18
3	0.43	0.43	-	1.00	0.48	0.48
4	0.43	0.43	1.00	-	0.48	0.48
5	0.18	0.18	0.48	0.48	-	1.00
6	0.18	0.18	0.48	0.48	1.00	-

the congruent $(++,--)$ and the incongruent $(+-,-+)$ dyads, the denominater representing the corresponding expected value:

$$\phi = \frac{ad - bc}{\sqrt{n_i u_i n_j u_j}} \tag{1}$$

Guilford (1956[3] p.511) has shown that this is a product-moment-coefficient. Therefore, it may also be described as the relationship of covariance and total variance, i.e. $\phi = cov/(s_i s_j)$, with $cov = (ad - bc)/N^2$, $s_i = [n_i u_i/N]^{1/2}$ and $s_j = [n_j u_j/N]^{1/2}$. Using relative frequencies $(p_i = n_i/N, q_i = u_i/N, p_j = n_j/N, q_j = u_j/N, p_{ij} = a/N)$ these terms can be replaced by $cov = p_{ij} - p_i p_j, s_i = [p_i q_i]^{1/2}$ and $s_j = [p_j q_j]^{1/2}$. In short, this coefficient exhibits advantages such as evidence of interpretation, simple applicability, statistical testability and equivalence to Pearson's product-moment-correlation. However, there is one property which has repeatedly given rise to discussion. Measures of association (cf. Kubinger, 1990) in general, owe their creation to empirists' desire to make covariation more tangible and to more precisely accomodate some coincidental events or facts which cannot be grasped adequately by graphic or linguistic means. For this reason it is not only an objectification function that is desired, but also a descriptive one. The measure should allow substantial interpretation. For this purpose the degree of association is usually limited to the interval $[-1, +1]$, one denoting perfect association and zero indicating the absence there of. The Phi-coefficient satisfies these claims. However, the extreme values can only be reached within symmetric distributions. This fact may lead to potentially significant misjudgment. Exemplification of this can be found in the responses to six intelligence items answered consistently by 90 subjects; the analysis of the data may have resulted in the frequencies $n_1 = 80$, $n_2 = 80$, $n_3 = 50$, $n_4 = 50$, $n_5 = 20$, $n_6 = 20$. Because of the high conformity of the answers to the model, the matrix of intercorrelations is expected to show predominantly ones, except for small declines in the case of differing levels of difficulty. In fact, most of the values range from .48 to .18 (tab.2). Only in the case of symmetric frequencies, i.e. $n_i = n_j$ (if $ad > bc$) or $n_i = N - n_j$, (if $ad < bc$) are the entries equal to ±1.00. A factor analysis will show as many factors as there are levels of difficulty (Ferguson, 1941).

Obviously the Phi-coefficient is considerably diminished when differences between the difficulty levels increase. With real data this technical effect meshes with actual response inconsistencies of the subjects. The observer is usually unable to decide, to what degree a value below one is affected by the distribution asymmetry of the variables or by some incontigency within the observed events. Hence, generally the comprehension of this measure of association is complicated in that values below one may be caused by response inconsistencies, levels of difficulty or both. These problems are not restricted to graduations of theoretical test constructs. A symmetric partition of two sets of events will also be the exception in natural binary variables. For instance, for mammals there is a complete biological association between sex and capability of childbearing. Nevertheless there will not be a Phi-correlation of one. A possible explanation for this might be that some of the males show child bearing capability, too. In reality the reduced correlation originates from the fact that some females are unable to bear children because of sex-unspecific reasons (illness, accident). Actually, the two variables sex and childbearing capability do not generate equivalent classes. In this case the reasons for the diminished correlation will be readily elucidated as the result – contradictory to expectation – will call on inquiries. However, such previous knowledge will not often be available as a protective instance. In contrary, for the most part, measures of association support a primarily explorative procedure which allows for the acquisition of relational knowledge only after data analysis. As an example, a fictitious examination of the association between sex and subscription of a woman's journal can be used, where by a correlation of $\phi = .50$ may have been found. It might be concluded that, although this magazine is more often read by women, it is also read by men. Without access to the original data there is little reason to distrust this interpretation. However a more precise inspection of the data would reveal that the sample of 1000 individuals was made up on the one hand of 500 female and 500 male subjects, and on the other hand of 800 non-subscribers and 200 exclusively female subscribers. (The cells of the fourfold table might contain the following frequencies: a=200, b=300, c=0, d=500). Actually a very strong association between sex and magazine subscription is given: The subscribers are exclusively female, male subjects do not take the journal. This fact should be expressed by a $\phi^* = 1$, but the Phi-coefficient fails to do so.

Is it possible to formulate an alternative, perhaps complementary measure of association? Logic offers an access. A perfect Phi-correlation contains an equivalence proposition: $pa \iff pb$. If a person p has a feature a then he also has feature b and vice versa; if he solves task a then he will solve b as well, and if he does not solve a, then he will be unable to solve b. Here it is presupposed that the two variables are dichotomized symmetrically, i.e. on both sets of events exist equivalent binary partitions. What, however, can be done if the real conditions are not in accordance with these presumptions or with the test requirements? The sets of positive responses are now not in an equal, but in a partial set relation and the following implication holds: $pb \Longrightarrow pa$ (und $\neg pa \Longrightarrow \neg pb$). If subject p solves the more difficult item b then he will solve the easier item a (and if he does not solve the easier a, then he will not solve the more difficult b).

What follows from this distinction? The classical Phi-coefficient is appropriate only if a theoretically symmetric distribution of the variables can be assumed. Possible asymmetries indicate model adequacy deficiencies and consequently reduce the correlation. The assumption of symmetry however, is not a cogent condition. For this reason a distributionally independent measure ϕ^* would be desirable. The intercorrelation matrix (tab.2) with entries of exclusively one show that a completely homogeneous item collection is given. Generally, aside from equivalences, implicative connections in dichotomous variables could be revealed by technical means. But what would such a measure look like and what are the desired properties?

2 Demands on a Phi-coefficient independent from marginal distributions

A view into the literature teaches us that alternatives for the Phi-coefficient have been sought almost since its publication. Already in the beginning of this century attempts were made to find more suitable measures of association. The basic idea of the Phi-correlation was to relate the difference $ad - bc$ of the product of congruent

$(++, --)$ and incongruent $(+-, -+)$ response patterns to the complete set of expected response patterns. This expected value corresponds to the geometric mean of the products of the marginal frequencies in the symmetric case. For asymmetric marginal distributions this referential term exceeds the possible maximum number of different response patterns. The effect increases with the extent of asymmetry. The search for alternatives therefore, was principally directed toward more suitable reference terms. In general, coefficients were proposed which kept $ad - bc$ in the numerator while modifying the nominal term (nom). This type of coefficient is referred to as Phi-modification:

$$\phi^* = \frac{ad - bc}{nom} \quad \Big| \quad |ad - bc| \leq nom \leq \sqrt{n_i u_i \cdot n_j u_j} \tag{2}$$

The more detailed denominator properties will be illuminated in further discourse. However, the admissable domain can already be defined. On the one hand, its absolute value must not decrease the numerator ($|ad - bc| \leq nom$), because it would violate the convention of the ±1 limitation. On the other hand, ϕ^* should be at least as large as ϕ since the diminishing influence of different difficulty levels is reduced; hence the nominal term must not be larger than the denominator of the original Phi-coefficient. The restricted interval in the positive range implies equal signs for ϕ and ϕ^*.

Equation [2] can be regarded as a conceptual frame for any Phi-modification. More explicit specifications are directed to the appropiate profile of requirements which should be orientated to the Phi-coefficient as far as possible. Undoubtedly the following three conditions are part of this. First, the marginal freqencies must not be changed (A1). Secondly, all entries in the fourfold table are frequencies and therefore must not be negative (A2). Third, the inversion of a variable will change the sign but must not change the numerical value (A3).

A1: Invariance of the marginal distributions

A2: Non-negativity of the cell frequencies

A3: Invariance of the absolute value for inverted variables

While these three demands emphasize common ground with the Phi-coefficient, the before mentioned interval restriction [2] has to be refined. Here the idea of Torgerson (1958) and Cureton (1959) can be included to relate the Phi-coefficient to the possible maximum:

$$\phi^* = \frac{\phi}{\phi_{\max}} \tag{3}$$

A preliminary notion of the maximal Phi is provided by solving for $\phi_{\max} = \frac{\phi}{\phi^*}$ and inserting formulas [1] and [2] at the corresponding positions. The following definition results:

$$\phi_{\max} = \frac{nom}{\sqrt{n_i u_i \cdot n_j u_j}} \quad \Big| \quad |ac - bd| \leq nom \leq \sqrt{n_i u_i \cdot n_j u_j} \tag{4}$$

The developmental advantage of this expression as compared to formula [2], lies in the elimination of all cell parameters (a, b, c, d). For the purpose of research economy it is desirable to formulate a nominal term based solely on the marginal frequencies. In this way the maximal values of earlier published Phi-coefficients can be found even if the original data are unknown. In accordance with formula [3], the distributionally independent coefficient can also be determined. Surely, things have not yet come to this point. First, the formal requirements will be more precisely addressed. If the lower interval bound $|ac - bd|$ is taken for the nominal term, then $\phi_{\max}$ can not fall below the absolute value of ϕ. Doing the same for the upper bound yields a value of 1. Hence the maximal Phi is limited to the interval $[|\phi|, 1]$. This assumption is stated in A4. The conditions for reaching these interval bounds can be explicitly stated. The upper bound (supremum) 1 can only be reached if the marginal distributions are symmetric. According to formula [3] in this case, ϕ is related to one, hence ϕ^* equals the original Phi-coefficient. Technically, the symmetry of marginal distributions becomes apparent upon review of the fourfold table by the fact that one of the row sums n_i or u_i is equal to the column n_j (A4'). On the other hand, variables with asymmetric marginals are in a partial set relation if no incongruent dyads occur. Formally this appears in the fourfold table if one of the cells is equal to zero (A4"). The Phi-modification in this case should result in ϕ^*=1, consequently $\phi_{\max} = |\phi|$:

A4: Interval boundaries

$$\phi_{\max} \in [\,|\phi|, 1]$$

A4': Supremum for marginal symmetry

$$\left\{\begin{matrix} n_i = n_j & | & ad \geq bc \\ n_i = (N - n_i) & | & ad < bc \end{matrix}\right\} \Longleftrightarrow \phi_{\max} = 1$$

A4": Identity for the partial set relation

$$\exists|\{a, b, c, d\} = 0 \Longleftrightarrow \phi_{\max} = |\phi|$$

3 Previous Proposals

As already mentioned a considerable collection of Phi-modifications has been developed. To examine them all with respect to requirements A1 through A4 would be a rather arduous task. Fortunately, they can be roughly divided into two periods, namely up to 1948 and from 1949 onward. The early authors (Forbes 1907, Yule 1911, Hacker 1920, Ferguson 1941, Dice 1945, Loevinger 1947) proposed nominal terms which were related more or less explicitly to a fixed data constellation, for example $nom = n_i \cdot n_j$ (Forbes), $nom = ac + bd$ (Yule) or $nom = n_j(N - n_i)$ (Loevinger). They do not satisfy the invariance of item inversion and the interval restriction.

Beginning with Cole (1949) it was attempted to meet the perceived problems with suitable constraints. These attempts, in short, will be outlined in the following. Table 3 shows the maximal Phi for a set of examplary data consisting of two groups. In the first two homogeneous asymmetric variables are inverted systematically and have a correlation of ±.33. The maximal Phi, then, must also be equal to .33. In the second group, two moderately correlated variables with symmetric marginals are systematically inverted. The Phi-coefficient is $\phi = \pm.25$, for the maximal pendant a value of 1 should result. Cole's (1949) formula with equality constraints read:

$$nom = \begin{cases} n_i(N - n_j) & | \quad ad \geq bc \\ n_i n_j & | \quad bc > ad, \quad c \geq a \\ (N - n_i)(N - n_j) & | \quad bc > ad, \quad a > c \end{cases} \tag{5}$$

Example (group 1, data set 1): $\phi_{\max} = 9 \cdot 9/(3 \cdot 9) = 3$

Cureton's (1959) proposal was based on the idea of relating Phi to the possible maximum. He presented two nominal expressions, one of which comprises two

Table 3: Invariance of absolute value of the Phi-correlation if items are inverted

Data group 1									Data group 2							
i	j	i	j	i	j	i	j		i	j	i	j	i	j	i	j
1	1	1	0	0	1	0	0		1	1	1	0	0	1	0	0
1	1	1	0	0	1	0	0		1	1	1	0	0	1	0	0
1	1	1	0	0	1	0	0		1	1	1	0	0	1	0	0
1	0	1	1	0	0	0	1		1	1	1	0	0	1	0	0
1	0	1	1	0	0	0	1		1	1	1	0	0	1	0	0
1	0	1	1	0	0	0	1		1	1	1	0	0	1	0	0
1	0	1	1	0	0	0	1		1	0	1	1	0	0	0	1
1	0	1	1	0	0	0	1		1	0	1	1	0	0	0	1
1	0	1	1	0	0	0	1		0	1	0	0	1	1	1	0
0	0	0	1	1	0	1	1		0	1	0	0	1	1	1	0
0	0	0	1	1	0	1	1		0	0	0	1	1	0	1	1
0	0	0	1	1	0	1	1		0	0	0	1	1	0	1	1
+.33		−.33		−.33		+.33		ϕ	+.25		−.25		−.25		+.25	

alternatives dependent on the sign. In the case of negative associations, regard should be given to the secondary condition that for $(n_i + n_j) < N$ the term in brackets is put to zero:

$$nom = \begin{cases} N \cdot \min(n_i, n_j) - n_i n_j & \mid \quad ad \geq bc \\ n_i n_j - N(n_i + n_j - N) & \mid \quad ad < bc, \quad n_i + n_j < N \Longrightarrow 0 \end{cases} \tag{6}$$

This modification was developed by Cureton in a somewhat curious style of argumentation and was finally presented in formulas which appear rather involved and difficult to interpret in terms of correlations or variances. It will be difficult for the user to assess the performance of the formalism. Presumably uncertainties such as these are primarily attributable to oversight of this modification. Actually, the above conditions appear to be met. A confirmation of this lies in that the expected values are attained. A thorough examination of the requirements would be difficult to conduct at this point and will be taken up later after the clarification of additional essentials which will help to provide for a better understanding.

A third alternative has been presented by Guilford (1965). He proposed that $\phi_{\max} = [(p_j q_i)/(p_i q_j)]^{1/2}$ in the case of positive and $\phi_{min} = [(q_i q_i)/(p_j p_j)]^{1/2}$ in the case of negative correlations. The symbols p_i and q_i correspond to the relative frequencies n_i/N und u_i/N. Unfortunately, again there is no invariance of inversion (group 1, column 4): $\phi_{\max} = [(.75 \cdot .75)/(.25 \cdot .25)]^{1/2} = 3$.

A fourth modification with three equality constraints was presented by Bortz, Lienert and Boehnke (1990, p.278) which is obviously influenced by Guilford's proposal:

$$\phi_{\max} = \sqrt{\frac{p_i q_j}{q_i p_j}} \qquad \mid \quad p_i \geq q_i, q_j \geq p_j, p_i \geq p_j \tag{7}$$

The following example illustrates the missing invariance of inversion (group 2, column1): $\phi_{\max} = (8 \cdot 8/4 \cdot 4)^{1/2} = 2$. The form with the constraints $p_i \leq q_i$ and $q_j \leq p_j$ (cf. Bortz 1993, p.211) is also not convincing; Hannöver and Steyer (1994) have warned of these deficiencies.

Examination of the previous Phi-modifications leads to the conclusion that, except for Cureton's proposal, the alternatives are not invariant with respect to inversions or, in case of symmetric marginals, are not identical with the Phi-coefficent. Most desirable, however, is a maximal Phi which satisfies the above requirements, is conceptually clear and does not presuppose informations other than the marginal frequencies.

4 Formal Development

The basic idea of the Phi-modification is to relate the real difference $ad - bc$ to the maximum possible difference or extreme difference. Indicating the joint frequencies by an asterisk, it is defined as: $\mathit{diff}_{\text{ext}} = a^*d^* - b^*c^*$. The sign is equal to that of the real difference. The index "ext" was chosen intentionally instead of "max" in order to make clear that it may be negative or positive just as the real difference. However, because of the restriction to the positive area, only the respective absolute value is applied. Hence, the neutral Phi is defined as follows:

$$\phi^* = \frac{\mathit{diff}}{|\mathit{diff}_{\text{ext}}|} = \frac{ad - bc}{|a^*d^* - b^*c^*|} \tag{8}$$

Obviously this is a coefficient of the type $\phi/\phi_{\max}$. The maximal Phi, according to formula [4], can be put into concrete terms by the expression $|a^*d^* - b^*c^*|$. A technical interpretation of variance is possible as well, i.e. as proportion of covariance to the maximal possible covariance: $\phi^* = \mathit{cov}/|\mathit{cov}_{\text{ext}}|$.

Now the task is to determine the extreme difference from the parameters a^*, b^*, c^* and d^*. To accomplish this only one of them must be known because the others can be computed via the marginals. But, which one should it be? A first clue is given by the above observation that two completely homogeneous variables with different levels of difficulty are accompanied by a zero cell in the fourfold table. Of course, for the most part, the response sets of the variables will not form a partial set relation, and therefore, a zero cell will not appear. However, in order to find the maximum difference, one of the elements should be interpreted as rate of violations of a perfect association ("errors"). It is important that only one cell assumes this role; otherwise a definite solution can not exist for this problem.

Any uncertainty as to the existence of this uniqueness can be quickly eliminated. If a cell is reduced to zero, the adjacent cell must increase and the diagonal cell must decrease equally, so that according to A1 the marginal frequencies remain constant. If the reduction of a cell implies the reduction of the diagonal cell, only the smaller of the two may become zero, otherwise the non-negativity of the frequencies (A2) would be violated. In which of the diagonal products should this smaller element be sought? Again it is the lesser of the two. If found in the larger product, the maximum difference $a^*d^* - b^*c^*$ would have the opposite sign of the real difference and would contradict the sign identity of real and maximum difference according to definition [8]. Therefore the error rate e corresponds to the lesser of the two cells b or c for positive and a or d for negative associations; it is always in the lesser of the two diagonal products:

$$e = \begin{cases} \min(b, c \mid ad \geq bc) \\ \min(a, d \mid bc \geq ad) \end{cases} \tag{9}$$

Stated so as to be easily remembered: *The error cell is the smaller element of the smaller diagonal product.* If the products are equal, either of the minimal values can be chosen. Since the Phi-coefficient in this case is equal to zero, it can not be upgraded anyway. The error rate is unimportant under this condition as long as it remains larger than zero. Technically, the above dual form is easy to handle. It is however, more difficult to integrate such expressions in more complex formulas. For convenience a convex combination of the minimum terms – here the weighted

Table 4: The nominal term in depencence of the error cell

		Nominal Term			
Diff	Error cell	I	II		III
$ad \geq bc$	$e = b, b^* = 0$	$\lvert a^*d^* - b^*c^*\rvert = a^*d^*$	$(a+b)(b+d)$	$n_i u_j$	$ad - bc + bN$
$\phi \geq 0$	$e = c, c^* = 0$	$\lvert a^*d^* - b^*c^*\rvert = a^*d^*$	$(a+c)(c+d)$	$n_j u_i$	$ad - bc + cN$
$ad \leq bc$	$e = a, a^* = 0$	$\lvert a^*d^* - b^*c^*\rvert = b^*c^*$	$(a+b)(a+c)$	$n_i n_j$	$bc - ac + aN$
$\phi \leq 0$	$e = d, d^* = 0$	$\lvert a^*d^* - b^*c^*\rvert = b^*c^*$	$(b+d)(c+d)$	$u_i u_j$	$bc - ad + dN$

sums – can be formed:

$$K = g \cdot \mathrm{Min}_{ad \geq bc} + (1-g) \cdot \mathrm{Min}_{ad \leq bc} \qquad \mid g = (1 + \mathrm{sgn}(\phi))/2 \tag{10}$$

The weight g takes the values 1 for positive and 0 for negative real differences. In this way one of the two sums will be 0, the other delivers the respective minimum. If the real difference equals zero, then the arithmetic mean of the two minimum terms is formed. Here the above described freedom is used to choose the rate of errors pragmatically in the case of zero differences.

Since it has now been determined that, in principal, a definite solution to the problem exists, one might try to find the extreme difference or the maximal Phi (cf. Bösser, 1979) by determining a^*, b^*, c^* and d^*. However, a more general solution on the basis of the marginal frequencies is desired. For this reason the asterisk-parameters are transformed into the corresponding marginal parameters. The formal step-by-step execution of this process is provided in table 4. Each cell of the fourfold table is successively regarded as potential error element. For example, for positive real differences either b or c will be zero; the nominal term is reduced to a^*d^*. In the next step this product is replaced by the corresponding real parameters (col. II in tab.4). The error element will be added to the neighbouring cells because of the marginal identity (A1). The resulting equivalences $(a+b)(b+d)$ and $(a+c)(c+d)$ correspond to the marginal products $n_i u_j$ and $n_j u_i$, respectively.

The ideal product a^*d^* therefore equals one of the diagonal products. Which of them is the right one? According to the interval restriction, the nominal term must not exceed the geometric mean of the marginal products. If the products $n_i u_j$ and $n_j u_i$ are different, one of them will be less than, the other larger than the geometric mean. Consequently the maximum product of the main diagonal is given by $a^*d^* = \min[n_i u_j, n_j u_i]$.

The minimum of the secondary diagonal is analogously determined. Hence the maximal absolute difference of the products is provided by the following equation:

$$|a^*d^* - b^*c^*| = \begin{cases} \min(n_i u_j, n_j u_i) & \mid \phi \geq 0 \\ \min(n_i n_j, u_i u_j) & \mid \phi \leq 0 \end{cases} \tag{11}$$

Substituting the two nominal terms as numerator in equation [4] yields the corresponding dual form for the maximal Phi:

$$\phi_{\max} = \begin{cases} \frac{\min[n_i u_j, n_j u_i]}{\sqrt{n_i u_i n_j u_j}} = \min\left(\sqrt{\frac{n_i u_j}{n_j u_i}}, \sqrt{\frac{n_j u_i}{n_i u_j}}\right) & \mid \phi \geq 0 \\ \frac{\min[n_i n_j, u_i u_j]}{\sqrt{n_i u_i n_j u_j}} = \min\left(\sqrt{\frac{n_i n_j}{u_i u_j}}, \sqrt{\frac{u_i u_j}{n_i n_j}}\right) & \mid \phi < 0 \end{cases} \tag{12}$$

With a little algebra each minimum term is converted into two inversely equivalent radical fractions. The numerator of the first equals the denominator of the second and vice versa. Hence one of the radicals is ≤ 1, the other ≥ 1. However, since $\phi_{\max}$ must not exceed one, a simple two step algorithm can be used for computer programs:

1. IF $\phi \geq 0$ THEN $\phi_{\max} = [n_i u_j/(n_j u_i)]^{1/2}$ ELSE $\phi_{\max} = [n_i n_j/(u_i u_j)]^{1/2}$.
2. IF $\phi_{\max} > 1$ THEN $\phi_{\max} = 1/\phi_{\max}$.

The maximal Phi follows a simple logic: It equals the smaller radical fraction of the cross over products $(n_i u_j, u_i n_j)$ for positive and the parallel products $(n_i n_j, u_i u_j)$ for negative real differences.

With formula [12] the primary objective is attained by determining the maximal Phi exclusively by the marginal frequencies. The frequencies n and u may be replaced by probabilities p and q. Of course the minimal term may also be entered into the convex form [10].

At this point the initial requirements should be viewed. The invariance of the marginal distributions (A1) and the non-negativity of the frequencies (A2) are not violated by design. The invariance of the absolute value in case of inversions (A3) is achieved since the substantial quantity for extreme differences – the "error cell" (the smaller element of the smaller product) – remains unchanged. The interval restriction (A4) exists since the absolute value of the extreme difference can never be less than the real difference and never larger than the geometric mean of the marginal products.

Finally, the neutral ϕ^* has to be determined. If Phi-coefficients are available they will be related to the maximal Phi. However, if the original data are available, the fraction of real difference and maximum difference [11] may be formed. A more elegant way of performing this is by replacing the marginal frequncies with the corresponding joint frequencies (last column of table 4). The smaller value of $\min[ad - bc + bN, ad - bc + cN]$ now only depends on the smaller cell b or c. The maximum difference simplifies to $ad - bc + N \cdot \min(b, c)$. By employment of a similar process for the minimum difference and substitution into the convex combination we obtain:

$$\begin{aligned} |a^* d^* - b^* c^*| &= g\Big[(ad - bc) + N \cdot \min[b, c]\Big] + (1 - g)\Big[(bc - ad) + N \cdot \min[a, d]\Big] \\ &= |ad - bc| + N\big[g \cdot \min[b, c] + (1 - g)\min[a, d]\big] \\ &= |ad - bc| + N \cdot e \end{aligned} \tag{13}$$

As the terms $g(ad - bc)$ and $(1 - g)(bc - ad)$ are always positive they are replaced by $|ad - bc|$ and then factored out. For the remaining quantity in brackets the symbol e was already introduced in formula [9]. A very condensed formula results:

$$\phi^* = \frac{ad - bc}{|ad - bc| + N \cdot e} \tag{14}$$

Here it becomes apparent that the extreme difference consists of the absolute real difference added by the "error component".

5 Discussion

The starting point of this discourse was the statement that the incongruence of the marginals diminishes the Phi-coefficient. This is plausible as long as the hypothesis of association proceeds from symmetric distributions: The more skewed the distribution, the lower the appropriateness and, consequently the correlation. This is different, however, if the structural objectives or the empirical facts require asymmetry. The reduction, then, on formal grounds, is a methodical artefact. This problem was recognized early on and was followed by a series of Phi-modifications which left the user somewhat helpless in that they were difficult to judge without a suitable frame of reference. Which one is logically correct, technically sure and substantially clear to interprete?

As comparison four formal requirements have been put forth. The first two, invariance of marginals and non-negativity of the cell frequencies, are in general uncritical. However, the invariance of the absolute value is nearly always violated. The only exception was presented by Cureton (1959). At this point the question, as to the adequacy of his modification, can now be rather easily answered. The nominal terms in formula [11] can be transformed algebraically:

$$\begin{aligned}
\phi \geq 0 : \min[n_i u_j, n_j u_i] &= \min[n_i N - n_i n_j, n_j N - n_i n_j] = N \cdot \min(n_i, n_j) - n_i n_j. \\
\phi \leq 0 : \min[n_i n_j, u_i u_j] &= \min[n_i n_j, N^2 - N n_i - N n_j + n_i n_j] \\
&= \min[n_i n_j, n_i n_j - N(n_i + n_j - N)].
\end{aligned}$$

For a positive Phi the concluding expression corresponds formally to Cureton's proposal (apart from the denotiation of the variables). Principally, this is true for a negative Phi as well. In fact, he only offers the second alternative of the minimum – i.e. the difference of the products $n_i n_j$ and $N(n_i + n_j - N)$ – and requires that the second product is set to zero if $(n_i + n_j) < N$; this corresponds to the term $n_i n_j$. Therefore, his modification adheres to the requirements.

Upon consideration of this statement one might side with Cureton by asking if an updated treatment of this problem is even necessary. The following offer some support for this:

1. Catalogue of requirements, examination of uniqueness, solution.

   As compared with earlier approaches, a more global problem analysis was adopted. In the first step a catalogue of requirements for a Phi-modification on a generalized level was developed; in the second step it was shown that a solution must exist for a coefficient with these properties since the rate of error for the nominal term is uniquely fixed. On this basis a measure was developed satisfying the requirements. Such an analytical procedure offers the possible classification of other coefficents. Only Cureton's approach stands the formal criteria. He developed his modification rather intuitively on technical considerations concerning the possible variations of the relative marginal frequencies in the fourfold table. The resulting formulas are rather clumsy and bound to more constraints. Insight into his procedure and the rather opaque formalism does not suggest itself. Therefore it will not be surprising that even in the recent past further Phi-modifications were proposed and discussed (cf. Guilford, 1965; Bösser, 1979; Kubinger, 1990; Bortz, 1993; Hannöver & Steyer, 1994; Kubinger, 1995; Bortz, 1995). With every advancement, considerations shift off track becoming obviously insufficient. But of course there are also old "true" elements gaining new meaning in a new framework of thinking.

2. The developed system helps to evaluate and choose suitable modifications.

   Up to now Phi-modifications were not easily discussed because principles of evaluation – for instance a list of requirements – existed at best subliminally. Hardly an author has explained in what respect his modification is superior to the previous ones; this holds for Cureton, too. Consequently, there was no sufficient reason for the user to prefer or reject a certain coefficient. Here the nominal expression turned out to have four slightly different appearances dependent on the inversion of the variables. The published solutions, however, except for one of them, contain only a partial set. This underscores the merit of the system: Not only the insufficiency of most of the proposals could be shown, but also the formal equivalence of one – nearly forgotten – modification. Above all, the danger of being seduced into making false statements of association and conclusions through the use of unsuitable Phi-modifications is prevented.

3. Better understanding and ease of reception through clarity of interpretation.

   The propagated modification can be adopted by the user in the most accepted form of interpretation. According to the definition, it may be regarded as the relation of real difference and maximum possible difference of the product of congruent and incongruent dyads. But it may also be read according to Loevinger's (1948) homogeneity as the ratio of covariance and maximal covariance, or according to Cureton (1959), as the quotient of real Phi and maximal Phi. The most substantial part of the Phi-modification, the nominal term, is easy to memorize. It equals the minimum of the cross over products $(n_i u_j, u_i n_j)$ for positive and the minimum of the parallel products $(n_i n_j, u_i u_j)$ for negative correlations.

4. Ease of handling.

   ϕ^* can be computed on the basis of the joint frequencies or by means of the Phi-coefficient and the marginal frequencies. Both versions can be easily implemented in computer programs.

However, in the end, a deficit has to be stated: Since little is known about the distribution of the coefficient, a statistical test of sigificance can not yet been recommended. Goodmans (1959) criterion may possibly help. For the present it will perhaps be best to decide conservatively by testing the significance of the corresponding classical Phi-coefficient.

Finally it should be reminded that ϕ^* should not be used in a factor analysis since the asymmetry contradicts the presupposed normal distribution of the data.

References

[1] Bortz, J. (1993^4). *Statistik Für Sozialwissenschaftler.* Berlin: Springer.

[2] Bortz, J., Lienert, G.A., Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

[3] Bösser, T. (1979). PHI-Interkorrelation seltener Symptome. *Psychologische Beiträge, 21*, 343-348.

[4] Cole, L.C. (1949). The measurement of interspecific association. *Ecology, 30*, 411-424.

[5] Cureton, E.E. (1959). Note on $\phi/\phi_{\max}$. *Psychometrika, 24*, 89-91.

[6] Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297-302.

[7] Ferguson, G.A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 41*, 323-329.

[8] Forbes, S.A. (1907). On the local distribution of certain Illinois fishes. An essay in statistical ecology. *Bulletin of the Illinois State Laboratory; Natural History, 7*, 273-303.

[9] Goodman, L.A. (1959). Simple statistical methods for scalogram analysis. *Psychometrika, 24*, 29-43.

[10] Guilford, J.P. (1965^{4}a). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

[11] Guilford, J.P. (1965b). The minimal PHI coefficient and the maximal PHI. *Educational and Psychological Measurement, 25*, 3-8.

[12] Hannöver, W., Steyer, R.(1994). Zur Korrektur des ϕ-Koeffizientenn. *Newsletter der Fachgruppe Methoden, II*, 4-5.

[13] Kubinger, K.D. (1990). Übersicht und Interpretation der verschiedenen Assoziationsmaße. *Psychologische Beiträge, 32*, 290-346.

[14] Kubinger, K.D. (1995). Entgegnung: Zur Korrektur des ϕ-Koeffizienten. *Newsletter der Fachgruppe Methoden,II*, 3-4.

[15] Loevinger, Jane A. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph, 61*, 1-49.

[16] Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507-529.

[17] Torgerson, W.S. (1958). *Theory and methods of scaling.* New York: Wiley.

[18] Yule, G.U. (1911). *An introduction to the theory of statistics.* London: Griffin & Co.

[19] Yule, G.U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society, 75*, 579-652.

First Submitted January 1, 97
Revised June 16, 97
Accepted June 17, 97