Measures of association for contingency tables

Sean Wallis, Survey of English Usage, University College London
ePublished: April 9 2012

1. Introduction

Often when we carry out research we wish to measure the degree to which one variable affects the value of another, setting aside the question as to whether this impact is sufficiently large as to be considered **significant** (i.e., significantly different from zero).

The most general term for this type of measure is **size of effect**. Effect sizes allow us to make descriptive statements about **samples**. Traditionally, experimentalists have referred to 'large', 'medium' and 'small' effects, which is rather imprecise. Nonetheless, it is possible to employ statistically sound methods for comparing different sizes of effect by estimating a Gaussian confidence interval (Bishop, Fienberg and Holland 1975) or by comparing contingency tables for the difference of differences (Wallis 2011).

In this paper we consider effect size measures for contingency tables of any size, generally referred to as "$r \times c$ tables". This effect size is the "measure of association" or "measure of correlation" between two variables. There are more measures applying to 2×2 tables than for larger tables.

Consider Table 1 below. A and B are dichotomous (two-way or Boolean) variables. We wish to find the best estimate that the value of $a \in A$ is dependent on the value of $b \in B$. We will refer to the ideal measure as the **dependent probability** of A given B, $dp(A, B)$.

A B	b_1	b_2	**Total**
a_1	45	5	50
a_2	15	35	50
Total	60	40	100

Table 1. An example 2×2 contingency table for two dichotomous variables *A, B*.

A B	b_1	b_2	**Total**
a_1	50	0	50
a_2	0	50	50
Total	50	50	100

Table 2. A **maximally dependent** contingency table.

It follows that if the data were arranged as in Table 2, we could conclude that the value of A completely depended on the value of B, and therefore our ideal dependent probability would be 1. Note that if we took any instance from the sample where $B = b_1$, then $A = a_1$ (and so forth).

This type of table is employed in conventional χ^2 tests of homogeneity (independence). Indeed, contingency tests may be considered as an assessment combining an observed effect size and the weight of evidence (total number of cases) supporting this observation.

Similar measures are used in other circumstances. Wallis (2012) discusses **goodness of fit** measures of association which measure the degree to which one categorical distribution correlates with another. Similarly, Pearson's r^2 and Spearman's R^2 are standard effect sizes (measures of correlation) for variables expressed as ratio (real numbers) and ordinal (ranked) data respectively.

However, with categorical data a multiplicity of alternate measures are available. Measures have often developed independently and their differences are rarely explored.

Several candidates for the size of the correlation (or association) between discrete variables have been suggested in the literature. These include the contingency coefficient C, Yule's Q and odds ratio o (Sheskin 1997). We can eliminate the odds ratio and Yule's Q from consideration because they only apply to 2×2 tables. The odds ratio is not probabilistic (it is the proportion between two

probabilities), although the logistic function can be applied to obtain the log odds. In this paper we consider three potential candidates: Cramér's ϕ, adjusted *C* and Bayesian dependency.

Given this range of potential measures for effect size, two questions arise.

> *Do they measure the same thing? And if not, how may we choose between them?*

2. Probabilistic approaches to dependent probability

2.1 Cramér's ϕ

Cramér's ϕ (Cramér 1946) is a probabilistic intercorrelation for contingency tables based on the χ^2 statistic.[1] We may compute ϕ from the formula

$$\phi \equiv \sqrt{\frac{\chi^2}{N \times (k-1)}} \tag{1}$$

where N is the total frequency, k is the minimum number of values of A and B, and χ^2 is the standard test for homogeneity. For 2×2 tables $k - 1 = 1$, so $\phi = \sqrt{\chi^2/N}$ is often quoted.

Note that this formula neatly defines a relationship between χ^2, ϕ, N and k. N represents the number of cases (weight of evidence). k standardises tables with different numbers of rows and columns.

An alternative formula, for 2×2 tables only, obtains a **signed** result, where a negative sign implies that the table tends towards the opposite diagonal.

$$\phi \equiv \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{2}$$

Equation (2) does not generalise to tables with more than one degree of freedom, however. We will therefore principally be concerned with absolute ϕ (equation 1).

2.2 Probabilistically adjusted C

A probabilistically adjusted *C* may be defined as follows (after Sheskin, 1997: 243),

$$C_{adj} \equiv \sqrt{\frac{\chi^2 \times k}{(\chi^2 + N) \times (k-1)}}. \tag{3}$$

No proof is required to show that this figure does not match ϕ!

2.3 Properties of ϕ

The following properties of ϕ are useful in interpreting scores, but are rarely summarised. For the mathematically inclined we present a number of novel proofs in the appendices.

1. **Limits.** In a maximally skewed table such as Table 2, χ^2 reaches a limit of $(k - 1) \times N$. As Cramér comments, the upper limit of ϕ, 1, is attained when and only when each row or column contains one single element other than zero. Hence in the case of non-square tables, k is the lower of the number of rows or columns.

[1] Sometimes called Cramér's *V*. Cramér (1946: 443) himself notes $f^2 = \chi^2 / N$ and $0 \le f^2 / (k-1) \le 1$ "thus $\chi^2 / N (k-1)$ may be regarded as a measure of the degree of association indicated by the sample."

$\phi = 0$

F	a	$\neg a$
b	$^1/_2$	$^1/_2$
$\neg b$	$^1/_2$	$^1/_2$

$\phi = p$

Φ	a	$\neg a$
b	$^{(p+1)}/_2$	$^{(1-p)}/_2$
$\neg b$	$^{(1-p)}/_2$	$^{(p+1)}/_2$

$\phi = 1$

I	a	$\neg a$
b	1	0
$\neg b$	0	1

Figure 1. ϕ measures the degree by which a flat matrix **F** is perturbed towards the identity matrix **I**.

ϕ has a value of 0 when there is no association between the two values, that is, when the probability of selecting a value of A is constant for any value of B.

2. **Interdependence.** In Appendix 1 we prove that Cramér's $\phi(A, B)$ ϕ measures the linear interpolation from flat to identity matrix (see Figure 1 for the idea). We can therefore refer to ϕ as the best estimate of the population **interdependent probability** $idp \equiv p(a \leftrightarrow b)$. This intercorrelation is robust, i.e. it 'gracefully decays' the further it deviates from this ideal. ϕ is also closely related to the Pearson product-moment coefficient. It is a root mean square measure of the orthogonal 'distance' from a maximal dependence line.

3. **Direction.** The concepts 'interdependent' and 'dependent' are distinct. Whereas dependent probabilities (e.g. $p(a \mid b)$) are directional ($p(a \mid b) \neq p(b \mid a)$), the interdependent probability is not. The contingency correlation $\chi^2(A, B) \equiv \chi^2(B, A)$, and therefore $\phi(A, B) \equiv \phi(B, A)$.

Another way of thinking about ϕ is to consider that the table represents the **flow of information** from B to A. If information about the value of B is irrelevant to the determination of the value of A, the score is zero. Information does not 'pass through' the matrix.

3. A Bayesian approach to dependent probability

An alternative to the stochastic approach to dependence $dp(A, B)$ can be constructed with a simple **Bayesian model**.

Church (2000) compares the probability of a Bayesian 'positive adaptation' probability, $p(+adapt)$, with the prior probability $p(prior)$ by measuring the probability of an event (the use of a selected word) in two parts of the same text. The absolute change in the probability of selecting a given b is, in Church's notation, $| p(+adapt) - p(prior) |$. We can rewrite this simply as

$$absolute\ dependent\ probability\ dp_A(a, b) \equiv | p(a \mid b) - p(a) |.$$

In Table 1 the chance of selecting a_1 is 0.5 (50 out of 100) by default. This increases to 0.75 (45 out of 60) if $B = b_1$, i.e., the probability has increased by 0.25. This is an absolute difference. The potential probability change is limited by $p(a)$, so if $p(a)$ is 0.5, $dp_A(a, b)$ can only reach 0.5.

This appears to be an underestimate. When the value of A is completely dependent on the value of B (cf. Table 2) the dependent probability should be 1. We simply scale the probability change by the available range:

$$relative\ dependent\ probability\ dp_R(a, b) \equiv \begin{cases} \dfrac{p(a \mid b) - p(a)}{1 - p(a)} & \text{if } p(a) < p(a \mid b) \\ \dfrac{p(a) - p(a \mid b)}{p(a)} & otherwise \end{cases} \quad (4)$$

Relative dependent probability will reach 1 for maximally dependent tables (such as Table 2).

We next consider summations. Where the value of B is known we could take the arithmetic mean of values of A.

A B	b_1	b_2	**Total**
a_1	45	5	50
a_2	15	35	50
Total	60	40	100

$p(a \mid b)$			$p(a)$	$dp_R(a, b)$		$dp_R \times p(b)$	
	0.75	0.125	**0.5**	0.5	0.75	0.3	0.3
	0.25	0.875	0.5	0.5	0.75	0.3	0.3
$p(b)$	**0.6**	0.4				½ Σ	**0.6**

B A	a_1	a_2	**Total**
b_1	45	15	60
b_2	5	35	40
Total	50	50	100

$p(b \mid a)$			$p(b)$	$dp_R(b, a)$		$dp_R \times p(a)$	
	0.9	0.3	0.6	0.5	0.75	0.25	0.375
	0.1	0.7	0.4	0.5	0.75	0.25	0.375
$p(a)$	0.5	0.5				½ Σ	**0.625**

Table 3. Summing Table 1 for the relative dependent probability of a given b (upper), and in the opposite direction, b given a (lower).

$$dp_R(A, b) \equiv \frac{1}{k}\sum_{i=1}^{k} dp_R(a_i, b), \tag{5}$$

Generalising across values of B we compute a weighted mean:

$$dp_R(A, B) \equiv \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{l} dp_R(a_i, b_j) \times p(b_j), \tag{6}$$

where k is the number of values of A and l the number of values of B.

When summed, relative change in probability is **directional** (absolute change is not). This is demonstrated in Table 3. We briefly touch upon the interpretation of a directional correlation measure in the next section.

In the case of a 2×2 table, equation (5) may be simplified to

$$dp_R(A, B) \equiv \frac{[p(a_1 \mid b_1) - p(a_1)]p(b_1)}{[1 - p(a_1)]p(a_1)}, \tag{7}$$

where a_1 and b_1 are the first Boolean values of A and B respectively. This simplification is possible because in a two-way table, $p(a_1) = 1 - p(a_2)$.

Unlike equation (5), this equation is *signed*, i.e., it obtains a positive value when the upper left and bottom right cells are greater than their alternates (in our case, when $b_1 \rightarrow a_1$ and $b_2 \rightarrow a_2$), and a negative value when the opposite ($b_1 \rightarrow a_2$, $b_2 \rightarrow a_1$) holds. The derivation is given in Appendix 2.

The main advantage of this formula is ease of use. The first computation in Table 3 is simply: $dp_R(A, B) = (0.75 - 0.5) \times 0.6 / (0.5 \times 0.5) = 0.6$.

4. Evaluating measures

In order to explore the behaviour of these formulae we construct idealised contingency tables defined in two steps. First, we interpolate a maximally dependent table (Figure 1) and an equivalent 'flat' matrix with a parameter x ranging from an even distribution for each row $\{N(a)/k, N(a)/k\ldots\}$ to the maximally dependent $\{N(a), 0\ldots\}$. Each row has values $\{x, (N(a) - x)/k\ldots\}$ where the frequency of the cell on the diagonal is x.

Second, each column in these tables is then weighted by a fixed prior probability $p(b)$. This introduces a directional bias into the model. We then plot a variety of dependent probability measures against x.

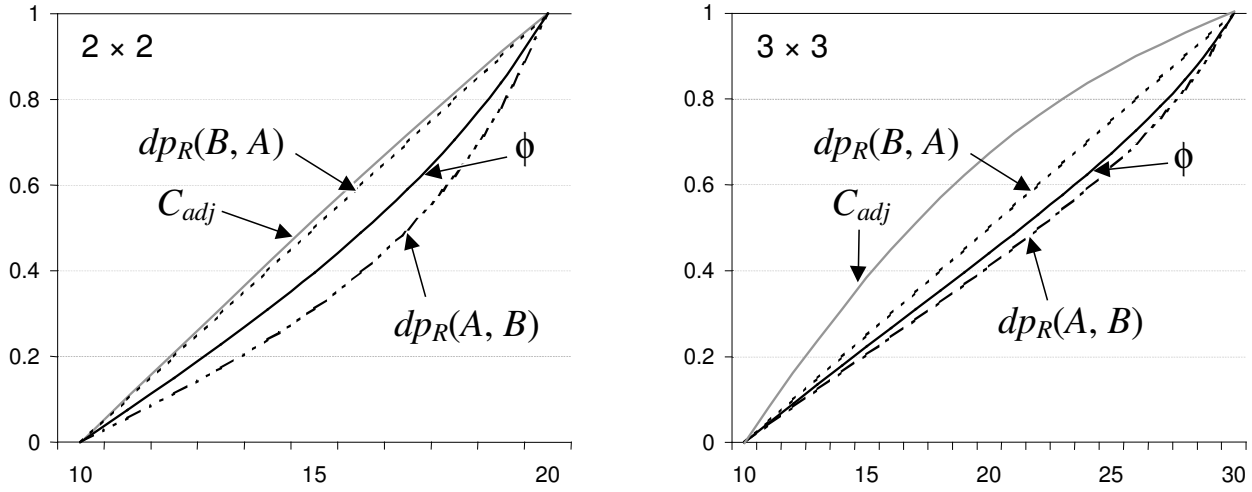

Figure 2: Effect of uneven priors. Measures of dp_R, C_{adj} and ϕ for idealised $k \times k$ tables with uneven prior probability $p(b)$ plotted against dependency factor x, $k = 2$ (left) and $k = 3$ (right).

Figure 2 visualises the relationship between ϕ, C_{adj} and Bayesian dp_R. The left graph plots dependency measures for a 2 × 2 table where $N(a)/k = 10$ and $p(b) = \{1/6, 5/6\}$. The right graph does the same for a 3 × 3 table where $p(b) = \{1/9, 4/9, 4/9\}$.

We note the following.

1. Relative dependency $dp_R(B, A)$ is linear with x when $p(a)$ is even.

2. Measures are ordered in size: $C_{adj} > dp_R(B, A) > \phi > dp_R(A, B)$.

3. $dp_R(A, B)$ (and therefore ϕ) converges to $dp_R(B, A)$ as $p(b)$ becomes more even (tends to $1/k$).

Whereas dp_R measures the distance on the first parameter from the prior (and is thus directional when a prior skew is applied to one variable only), ϕ is based on the root mean square distance of both variables. C_{adj} appears to behave rather differently to ϕ, as the right hand graph in Figure 2 shows. Given that the only other bi-directional measure, ϕ, measures the interdependence of A and B, there appears to be little advantage in adopting the less conservative C_{adj}.

Finally, for $k = 2$ the following equation also holds:

4. $\phi^2 = dp_R(A, B) \times dp_R(B, A)$.

We find an equality between a classical Bayesian approach to dependency and a stochastic approach based on Pearson's χ^2 for one degree of freedom. The proof is given in Appendix 3.

This raises the following question: what does 'directionality' mean here?

Note that dp_R differentiates by direction of inference if and only if there is a difference in the distributions of A and B (in our example above $p(a)$ is even but $p(b)$ varies). Given that we are analysing data from a natural experiment, we must resist the temptation to infer **causal** direction from a correlation. 'Directionality' here therefore simply means that the model accounts for the possibility that the distributions of the prior probabilities for A and B are different.

Is there an empirical advantage in employing dp_R over ϕ? Theoretically one might argue that dp_R captures more information than ϕ and is therefore a better estimate of the actual dependency of one case on another. However, this benefit remains to be demonstrated in practice.

5. Robustness and confidence intervals on ϕ

Note that there are typically three situations where it might be valuable to calculate confidence intervals on measures of association:

1. **To plot the measure with intervals.** Bishop, Fienberg and Holland (1975) identify a formula for the standard deviation for ϕ, and using this it is possible to plot approximate Gaussian intervals as $\phi \pm z.s$ (see Wallis 2011). See also the next section.

2. **To identify non-zero scores.** We discussed this question by way of introduction. Sheskin (1997) concurs that the equivalent χ^2 (or Newcombe-Wilson) test will determine whether the size of an effect is significantly different from zero.

3. **To compare values for statistical separability.** Wallis (2011) describes separability tests for comparing the outcomes of 2 × 2 tests of homogeneity, including by combining intervals using Bishop *et al.*'s formula.

Note that a confidence interval is concerned with limits of movement along a single degree of freedom. Consequently it is more difficult to interpret the results of comparing ϕ values across arbitrary $r \times c$ tables, and it is more useful to evaluate statistical separability (Wallis 2011) than to attempt to compare differences and intervals. Indeed, for comparing tables with a single degree of freedom the author recommends tests employing Wilson's score interval.

6. A worked example

Figure 3 provides a demonstration of plotting confidence intervals on ϕ. The figure summarises the results of an investigation into grammatical priming, i.e. the tendency to **reuse** a grammatical construction once it has already been employed.

Here *A* and *B* both represent the value of a linguistic choice of postmodifying clause type (**relative** vs. **non-finite** clause) at different locations in the same tree structure (i.e. sentence), using the ICE-GB corpus as a dataset.

In this experiment, ϕ measures the association between the decision to employ one type at point A in the tree and the decision to employ the same type at point B. Both A and B are found under a shared ancestor clause C. We aggregate data according to the distance between A and B, δ, counting phrases and clauses up to C. We distinguish between situations where C co-ordinates A and B or is a host clause. Coordinated cases are limited to even distances due to grammatical constraints.

To read the graph, note that ϕ = 0 means "no association" and ϕ = 1 means "completely determined". Numerically, the effect of coordination C obtains a higher ϕ score than the host clause at the same distance, but only a distance δ = 2 (one up, one down) obtains a **significantly** greater result.

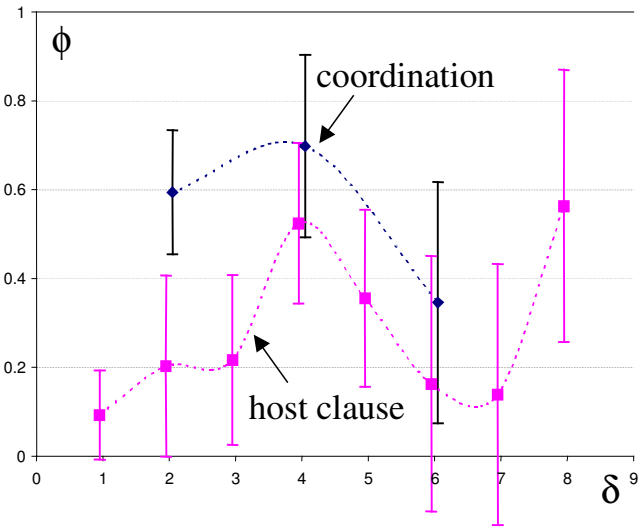

Figure 3: Association between two related decisions ϕ(*A*, *B*) over distance δ. Error bars based on $\sigma_\infty(\phi)$ (Bishop *et al.* 1975: 386).

References

BISHOP, Y.M.M., FIENBERG, S.E. & HOLLAND, P.W. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

CHURCH, K.A. 2000. Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2. *Proceedings of Coling-2000*. 180-186.

Cramér, H. 1946. *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

SHESKIN, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, Fl: CRC Press.

WALLIS, S.A. 2011. Comparing χ^2 tests for separability: Interval estimation for the difference between a pair of differences between two proportions. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf

WALLIS, S.A. 2012. Goodness of fit measures for discrete categorical data. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/gofmeasures.pdf

Appendix 1. The best estimate of population interdependent probability is Cramér's ϕ

Cramér's ϕ is not merely a 'measure of association'. It represents the best estimate of the dependent probability p for the population in a square matrix. Daniel (1990) notes that a perfect correlation is achievable between the two variables. Guilford (1965) comments that, for a 2×2 χ^2, ϕ measures the 'intercorrelation' between the two variables. Below we prove that this intercorrelation is the best estimate of a population interdependent probability for the general case $k \times k$.

Consider the identity matrix $k \times k$ such that $a_i = b_i$ and a_i represents the i-th value of A (etc.). This is defined with ones along the diagonal and zeros elsewhere. For all $i = 1 \ldots k$, this matrix maps as follows.

	b_1	b_2		b_k
a_1	1	0		0
a_2	0	1		0
a_k	0	0		1

This is a special case of the interdependent probability matrix where $a_i = b_i$ with probability $p = 1$. This is an *inter*dependent probability matrix because the transformation is *reversible*. We can summarise the general case as

	b_1	b_2		b_k
a_1	α	β		β
a_2	β	α		β
a_k	β	β		α

where $\alpha = p + (1 - p)/k$ and $\beta = (1 - \alpha)/(k - 1)$, such that $\alpha, \beta \in [0, 1]$ and rows and columns sum to 1 and $N = k$. Were this to be treated as a chi-square table the expected distribution of this is a constant $1/k$.

Note that we can multiply the table by any factor without affecting the relative skew. To simplify our notation, we first multiply the matrix by k. Rows and columns sum to k, each expected value is 1, $\alpha = pk + (1 - p)$, $\beta = (k - \alpha)/(k - 1)$ and $N = k^2$.

PROOF.

Consider the application of this table through $N = k^2M$ cases. We multiply the entire table by M, i.e. each observed cell **O** becomes αM or βM respectively. The expected values **E** are simply M. We can divide the summation into two distinct terms,

$$\chi^2 = \sum \frac{(\mathbf{O} - \mathbf{E})^2}{\mathbf{E}} = \boldsymbol{A} + \boldsymbol{B}$$

where $\boldsymbol{A} = k\frac{(\alpha M - M)^2}{M}$ and $\boldsymbol{B} = k(k-1)\frac{(\frac{k-\alpha}{k-1}M - M)^2}{M}$

$\boldsymbol{B}$ may be simplified as follows:

$$\begin{aligned}\boldsymbol{B} &= k(k-1)\frac{(\frac{k-\alpha}{k-1}M - M)^2}{M} \\ &= k(k-1)\frac{\left(\frac{kM - \alpha M - kM + M}{k-1}\right)^2}{M} = \frac{k}{k-1}\times\frac{(M-\alpha M)^2}{M} \\ &= \frac{\boldsymbol{A}}{k-1}.\end{aligned}$$

Substituting $\boldsymbol{A}$ and $\boldsymbol{B}$ back into χ^2,

$$\begin{aligned}\chi^2 &= (k + \frac{k}{k-1})\frac{(\alpha M - M)^2}{M} \\ &= (\frac{k^2}{k-1})\frac{(pkM + (1-p)M - M)^2}{M} = (\frac{k^2}{k-1})\frac{(pM(k-1))^2}{M} \\ &= k^2 p^2 M(k-1) \\ &= p^2 N(k-1).\end{aligned}$$

Therefore

$$p = \sqrt{\frac{\chi^2}{(k-1)N}} = \text{Cramér's } \phi.$$

Therefore for a square dependency matrix through which N random values pass, the optimum estimate of the population interdependent probability p is Cramér's ϕ. □

Corollary for 'goodness of fit' χ^2

In Wallis (2012) we consider employing ϕ in $r \times 1$ 'goodness of fit' conditions. Equation (1) may be rewritten accordingly:

$$\phi' = \phi(A, b) = \sqrt{\frac{\chi^2(A,b)}{(k-1)N(b)}}$$

where k is the number of values of A. Next, consider a model defined as below where α and β are defined as before ($\alpha = pk + (1-p)$ and $\beta = (k-\alpha)/(k-1)$, $N = k$) and the expected distribution $\mathbf{E} = \{1, 1, 1\ldots\}$.

	b
a_1	α
a_2	β
a_k	β

Multiplying this model by kM so that $N = k^2M$ permits the application of the proof above, i.e. that ϕ' equals the dependent probability p.

This proof depends on $\mathbf{E}$ being evenly distributed (constant) across k values. Where $\mathbf{E}$ is unevenly distributed it is necessary to reallocate variation to the observed frequency prior to computing χ^2 and thus ϕ'. See Wallis (2012).

Appendix 2. Deriving 2 × 2 rule dependency equation (7)

For a 2 × 2 table with a single degree of freedom, the following axioms hold.

A1. $p(a_2) = p(\neg a_1) = 1 - p(a_1)$; $p(a_2 \mid b_i) = p(\neg a_1 \mid b_i) = 1 - p(a_1 \mid b_i)$,

A2. $p(a_1 \mid b_i) - p(a_1) = p(a_2) - p(a_2 \mid b_i)$,

A3. $[p(a_1 \mid b_i) < p(a_1)] \leftrightarrow [p(a_2) < p(a_2 \mid b_i)]$.

A1 is a consequence of the Boolean definition of A, A2 can be demonstrated using Bayes' Theorem and A3 is a consequence of A2. A2 further entails that row sums are equal, i.e. $dp_R(a_1, B) = dp_R(a_2, B)$.

Equation (5) may therefore be simplified as follows

$$dp_R(A, B) \equiv \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} dp_R(a_i, b_j) \times p(b_j) = \sum_{j=1}^{k} dp_R(a_1, b_j) \times p(b_j)$$
$$= \frac{p(a_1 \mid b_1) - p(a_1)}{1 - p(a_1)} \times p(b_1) + \frac{p(a_1) - p(a_1 \mid b_2)}{p(a_1)} \times p(b_2).$$

Applying Bayes' Theorem ($p(a_1 \mid b_2) \equiv p(b_2 \mid a_1) \times p(a_1) / p(b_2)$) and axiom A1:

$$dp_R(A, B) = \frac{p(a_1 \mid b_1)p(b_1)}{1 - p(a_1)} - \frac{p(a_1)p(b_1)}{1 - p(a_1)} + \frac{p(a_1 \mid b_1)p(b_1)}{p(a_1)} - p(b_1).$$

The first and third terms then simplify to $[p(a_1 \mid b_1)p(b_1)] / [(1 - p(a_1))p(a_1)]$, so

$$dp_R(A, B) = \frac{p(a_1 \mid b_1)p(b_1) - p(a_1)^2 p(b_1) - (1 - p(a_1))p(a_1)p(b_1)}{(1 - p(a_1))p(a_1)}$$
$$= \frac{[p(a_1 \mid b_1) - p(a_1)]p(b_1)}{[1 - p(a_1)]p(a_1)}. \qquad = \text{equation (7)} \square$$

Appendix 3. For a 2 × 2 table, $\phi^2 \equiv dp_R(A, B) \times dp_R(B, A)$

The proof is in three stages.

STAGE 1. Simplifying the product $dp_R(A, B) \times dp_R(B, A)$.

Let Π stand for the product $dp_R(A, B) \times dp_R(B, A)$. From equation (7),

$$\Pi = \frac{[p(a_1 \mid b_1) - p(a_1)]p(b_1)}{[1 - p(a_1)]p(a_1)} \times \frac{[p(b_1 \mid a_1) - p(b_1)]p(a_1)}{[1 - p(b_1)]p(b_1)}$$
$$= \frac{[pr(a_1 \mid b_1) - pr(a_1)][pr(b_1 \mid a_1) - pr(b_1)]}{[1 - pr(a_1)][1 - pr(b_1)]}$$

Applying Bayes' Theorem to the second dependent probability term, $p(b_1 \mid a_1) \equiv p(a_1 \mid b_1) \times p(b_1) / p(a_1)$), and expanding, we have

$$= \frac{[p(a_1 \mid b_1) - p(a_1)]^2}{[1 - p(a_1)][1 - p(b_1)]} \times \frac{p(b_1)}{p(a_1)}.$$

STAGE 2. Converting to a+b+c+d notation.

The 2×2 χ^2 statistic, and thus ϕ, may be represented simply in terms of four frequencies in the table, a, b, c and d (note roman font to distinguish from a, a_1, etc). The table is labelled thus, and $N \equiv$ a+b+c+d:

	b_1	b_2	Σ
a_1	a	b	a+b
a_2	c	d	c+d
Σ	a+c	b+d	N

Probabilities are defined accordingly, thus $p(a1) \equiv$ (a+b) / N, $p(a_1 \mid b_1) \equiv$ a / (a+c), etc. Using this notation, the 2×2 ϕ^2 (with one degree of freedom) may be written as (cf. Sheskin, 1997: 244):

$$\phi^2 = \frac{\chi^2}{N} = \frac{(\mathrm{ad}-\mathrm{bc})^2}{(\mathrm{a}+\mathrm{b})(\mathrm{c}+\mathrm{d})(\mathrm{a}+\mathrm{c})(\mathrm{b}+\mathrm{d})}$$

Similarly,

$$\begin{aligned}\Pi &= \frac{(\frac{\mathrm{a}}{\mathrm{a}+\mathrm{c}}-\frac{\mathrm{a}+\mathrm{b}}{N})^2(\frac{\mathrm{a}+\mathrm{c}}{N})}{(\frac{\mathrm{c}+\mathrm{d}}{N})(\frac{\mathrm{b}+\mathrm{d}}{N})(\frac{\mathrm{a}+\mathrm{b}}{N})} \\ &= \boldsymbol{E}/(\mathrm{a}+\mathrm{b})(\mathrm{c}+\mathrm{d})(\mathrm{a}+\mathrm{c})(\mathrm{b}+\mathrm{d}),\end{aligned}$$

where $\boldsymbol{E} \equiv \left(\mathrm{a}^2 - 2\mathrm{a}(\mathrm{a}+\mathrm{b})(\mathrm{a}+\mathrm{c})/N + (\mathrm{a}+\mathrm{b})^2(\mathrm{a}+\mathrm{c})^2/N^2\right)N^2$.

STAGE 3. Algebraic reduction to prove $\boldsymbol{E} = (\mathrm{ad} - \mathrm{bc})^2$.

Substituting $N =$ a+b+c+d,

$$\begin{aligned}\boldsymbol{E} &= (\mathrm{a}+\mathrm{b}+\mathrm{c}+\mathrm{d})^2\mathrm{a}^2 - 2(\mathrm{a}+\mathrm{b}+\mathrm{c}+\mathrm{d})\mathrm{a}(\mathrm{a}+\mathrm{b})(\mathrm{a}+\mathrm{c}) + (\mathrm{a}+\mathrm{b})^2(\mathrm{a}+\mathrm{c})^2 \\ &= (\mathrm{a}^4+2\mathrm{a}^3\mathrm{b}+2\mathrm{a}^3\mathrm{c}+2\mathrm{a}^3\mathrm{d}+\mathrm{a}^2\mathrm{b}^2+2\mathrm{a}^2\mathrm{bc}+2\mathrm{a}^2\mathrm{bd}+\mathrm{a}^2\mathrm{c}^2+2\mathrm{a}^2\mathrm{cd}+\mathrm{a}^2\mathrm{d}^2) \\ &\quad - 2(\mathrm{a}^4+2\mathrm{a}^3\mathrm{b}+2\mathrm{a}^3\mathrm{c}+3\mathrm{a}^2\mathrm{b}+\mathrm{a}^2\mathrm{b}^2+\mathrm{ab}^2\mathrm{c}+\mathrm{a}^2\mathrm{c}^2+\mathrm{abc}^2+\mathrm{a}^3\mathrm{d}+\mathrm{a}^2\mathrm{bd}+\mathrm{a}^2\mathrm{cd}+\mathrm{abcd}) \\ &\quad + (\mathrm{a}^4+2\mathrm{a}^3\mathrm{c}+\mathrm{a}^2\mathrm{c}^2+2\mathrm{a}^3\mathrm{b}+4\mathrm{a}^2\mathrm{bc}+2\mathrm{abc}^2+\mathrm{a}^2\mathrm{b}^2+2\mathrm{ab}^2\mathrm{c}+\mathrm{b}^2\mathrm{c}^2) \\ &= \mathrm{a}^2\mathrm{d}^2 + \mathrm{b}^2\mathrm{c}^2 - 2\mathrm{abcd} = (\mathrm{ad} - \mathrm{bc})^2. \square\end{aligned}$$

The final stage completes the proof.

Note that this equality, $\phi^2 \equiv dp_R(A, B) \times dp_R(B, A)$, does not apply to tables with more than one degree of freedom (and Cramér's ϕ more generally).