The Annals of Statistics
1987, Vol. 15, No. 2, 694–711

THE AMALGAMATION AND GEOMETRY OF TWO-BY-TWO CONTINGENCY TABLES

BY I. J. GOOD[1] AND Y. MITTAL[2]

Virginia Polytechnic Institute and State University

If a pair of two-by-two contingency tables are amalgamated by addition it can happen that a measure of association for the amalgamated table lies outside the interval between the association measures of the individual tables. We call this the amalgamation paradox and we show how it can be avoided by suitable designs of the sampling experiments. We also study the conditions for the "homogeneity" of two subpopulations with respect to various measures of association. Some of the proofs have interesting geometrical interpretations.

1. Introduction. Consider a two-by-two contingency table whose entries in reading order are a, b, c, d with $a + b + c + d = N$, the sample size. We denote this table by $\mathbf{a} = [a, b; c, d]$ and we suppose that $abcd \neq 0$ and also that N is so large that sampling variation can be ignored.

We shall call the row categories T and $\bar{T}$ because often the first row corresponds to a treatment and the second row to a nontreatment (but it may be another treatment), and we call the column categories S and $\bar{S}$ because they might correspond to "success" and "failure." We use this notation consistently even if the causal relationship is reversed or even if there is no directed causal relationship as, for example, if the rows correspond to eye color and the columns to hair color.

Measures of association are used either to measure some aspect of a causal association or just an association between classifications in which neither classification is a (partial) cause of the other. A measure of association is a function of a, b, c and d, denoted by $\alpha = \alpha(\mathbf{a})$ with some desirable properties as described in Section 3. All the measures of association mentioned in this paper have the property that $\alpha(\mathbf{a}) = 0$ or 1 when the rows and columns are "independent," that is, when $ad = bc$. We usually call a measure of association simply a *measure*.

Let $\mathbf{a}_i = [a_i, b_i; c_i, d_i]$, $i = 1, 2, \ldots, n$, be the two-by-two contingency table corresponding to the ith of n mutually exclusive subpopulations, again with $a_i b_i c_i d_i \neq 0$. For example, when $n = 2$ the subpopulations might consist of men and women for testing a new drug. Let $N_i = a_i + b_i + c_i + d_i$ denote the sample size for the ith subpopulation and let $N = N_1 + N_2 + \cdots + N_n$, which is the total sample size from the whole population. If the n tables are added together, a process called *amalgamation*, we obtain a table $\mathbf{A} = [A, B; C, D] = [\Sigma a_i, \Sigma b_i;$

Received July 1985; revised June 1986.
[1]Research supported in part by an NIH grant.
[2]Research supported in part by an ONR contract.
AMS 1980 *subject classifications.* Primary 62H17; secondary 62A99.
Key words and phrases. Amalgamation paradox, contingency tables, homogeneity of subpopulations, geometry of contingency tables.

$\Sigma c_i, \Sigma d_i]$, where of course $A + B + C + D = N$. For the sake of simplicity we assume that N_i is proportional to the fraction p_i of the population that makes up the ith subpopulation. Taken literally this may not be a sensible way to sample; for example, we might want to sample people from New York and from Blacksburg separately, and then it would be inefficient to take sample sizes proportional to their populations. But we can scale the tables to force N_i to be proportional to p_i: This scaling is legitimate provided that the sample sizes are so large that sampling variation can be ignored for each subpopulation, as we are assuming. (It does not matter that the cell "frequencies" are not integers after scaling.) Our work is concerned with a paradox, essentially familiar, defined in the following manner:

DEFINITION 1.1. We say that the amalgamation (or aggregation) paradox, or simply the paradox, occurs if

$$\max_i \alpha(\mathbf{a}_i) < \alpha(\mathbf{A}) \quad \text{or} \quad \alpha(\mathbf{A}) < \min_i \alpha(\mathbf{a}_i).$$

We are here using "paradox" in the sense of an ostensible contradiction, which is one of its dictionary definitions.

Yule (1903) pointed out that "a pair of attributes does not necessarily exhibit independence within the universe [population] at large even if it exhibits independence in every subuniverse [subpopulation]"; that is, in our notation one can have $\alpha(\mathbf{a}_i) = 0$ (or 1) for all i, but $\alpha(\mathbf{A}) \neq 0$ [or $\alpha(\mathbf{A}) \neq 1$]. This was emphasized, for example, by Yule and Kendall (1950, pages 36–38) and presumably in all of the many editions and translations of that text including the ones before 1937, of which Yule was the sole author. Pearson (1899) had emphasized an analogous point regarding correlation measures for continuous (noncategorical) data, and Yule (1903) acknowledges Pearson. The paradox appears, using real data, in the slightly stronger form that $\alpha(\mathbf{A})$ can be negative (or less than 1) although $\alpha(\mathbf{a}_i) \geq 0$ [or $\alpha(\mathbf{a}_i) \geq 1$] for all i, in Cohen and Nagel (1934, page 499). [According to Cartwright (1979, page 422), Nagel suspects that he learned of the paradox from the 1904 edition of Yule (1911).] This stronger form of the paradox was discussed briefly by Simpson (1951, page 240), who stated that "the dangers of amalgamating two-by-two tables are well known" and he cited Kendall (1945, page 317). Thus the causal chain Pearson–Yule–Kendall–Simpson is fully documented. Blyth (1972) called the paradox "Simpson's paradox" in accordance with Stigler's law [Stigler (1980)] that eponymy is always wrong. [See Good (1985a) for a brief history of Stigler's law.] Messick and van de Geer (1981) called the paradox the "reversal paradox," but in Yule's formulation the subpopulations each had zero association in which case there is no reversal of sign. Hence we prefer the name "amalgamation paradox." Our formulation of the paradox is slightly more general than in these other papers.

When the paradox occurs it can be misleading or even dangerous. For example, a drug can be judged to be beneficial, as measured by α, for both men and women considered separately, but can seem to be harmful for the population

at large, by looking only at the amalgamated table. This can happen even though $N_i \propto p_i$. We claim that such a situation can arise only if not enough care is used in the design of the experiment. In Section 4, for each of several meaningful measures of association, we state what the designs should be in order to avoid the paradox. There is additional discussion of these measures in Goodman and Kruskal (1959) and Good (1985c).

In Section 5 we study a further property defined as follows:

DEFINITION 1.2. Two subpopulations, or the corresponding contingency tables $\mathbf{a}_1$ and $\mathbf{a}_2$ are *homogeneous* with respect to α if

$$\alpha(\mathbf{a}_1) = \alpha(\mathbf{a}_2) = \alpha(\mathbf{A}),$$

where $\mathbf{A} = \mathbf{a}_1 + \mathbf{a}_2$ is the amalgamated table.

We shall find necessary and sufficient conditions for two subpopulations to be homogeneous with respect to α, where α is any one of a certain set of measures that we define in Section 3.

2. Designs of relevant experiments. We now define two conditions under which the later theorems are proved and discuss situations under which these conditions can be built into the design of an experiment.

DEFINITION 2.1. An experimental design is said to be *row-uniform* or *row-fair* if, for some λ,

$$\frac{a_i + b_i}{c_i + d_i} = \lambda, \qquad i = 1, 2, \ldots, n. \tag{2.1}$$

Similarly we call the design *column-uniform* or *column-fair* if, for some μ,

$$\frac{a_i + c_i}{b_i + d_i} = \mu, \qquad i = 1, 2, \ldots, n. \tag{2.2}$$

Whether either or both these conditions can be attained depends on the nature of the tables and on the sampling procedures used. We first give definitions pertaining to the nature of the tables.

DEFINITION 2.2. A table is called row-causal or R-causal (column-causal or C-causal) if the row (column) categories can be regarded as (probabilistic) causes of the column (row) categories.

For example, the rows might correspond to smoking and nonsmoking, and the columns to high and normal blood pressure. An example of a "noncausal" table is one where the rows and columns correspond to attributes such as eye color and hair color.

For R-causal tables it is often practicable to use sampling procedures I and II_R, but less practicable to use procedures II_C or III, as defined in the

following manner:

DEFINITION 2.3. In sampling procedure I we sample at random from a population (or subpopulation). For two-by-two tables this could also be called tetranomial sampling.

DEFINITION 2.4. In sampling procedure II_R (or II_C) we control (fix) the row (column) totals, and then sample at random until these marginal totals are attained. These sampling procedures are sometimes called product-binomial sampling.

DEFINITION 2.5. In sampling procedure III, both the row and column totals are controlled.

A familiar kind of example of procedure III is where a subject tries to discriminate between two kinds of chewing tobacco knowing that there are four clods of each kind, that is, that the row totals are both 4, and she is forced to make each column total equal to 4 to give herself a chance of getting all eight "guesses" right. In other words as soon as one column total is 4 she has no further choice unless she is allowed to change her mind about previous guesses. Compare the famous experiment described by Fisher (1949, Chapter 2) who, however, did not express the matter in terms of contingency tables, and where the lady tasted tea instead of chewing tobacco.

Sampling procedure I can be used for both causal and noncausal tables. Procedures II_R and II_C can also be used for both, except that, for R-causal tables there would be less point in controlling the column totals alone. In the tasting experiment, which exemplifies procedure III, the true brands are potential causes of the "guesses." But procedure III is also exemplified by the sampling of a population of people classified according as they are above or below median height and above or below median income, and in this example the causal relationship is obscure.

An R-uniform (C-uniform) design is clearly possible under sampling procedure II_R (II_C). Under sampling procedure III it is easy to use a design that is both row and column uniform. Under sampling procedure I it is impossible to construct a design that could be guaranteed to be either row or column uniform. Similar comments apply to R-uniform (C-uniform) designs under sampling procedure II_C (II_R). However, R-uniformity (C-uniformity) can be *observed* (as opposed to being guaranteed in advance), at least approximately for a given group of contingency tables, as a consequence of either pure chance or the nature of the attributes studied. For example, consider a retrospective study of daughters with vaginal cancer (S) and mothers treated with DES (T) [Herbst, Ulfelder and Poskanzer (1971); Herbst, Poskanzer, Robboy, Friedlander and Scully (1975)]. If $P(S|T)$ and $P(S|\bar{T})$ are constant in various subpopulations, then in the experiment in which the total number of daughters in each category is fixed (sampling procedure II_C) we will observe that the design is approximately R-uniform. Thus in this experiment the design could be C-uniform and approximately R-uniform. For the effects of this approximation on our results, see the Appendix.

3. Association measures and properties.

Homogeneity of degree zero. A measure of association should be a homogeneous function of (a, b, c, d) of degree zero; that is, $\alpha(\mathbf{a}) = \alpha(\lambda\mathbf{a})$ for all positive λ, at least asymptotically for large samples, and we are only concerned with large samples in this paper. If this condition were not satisfied (asymptotically at least), then α would not be a consistent estimator of a population parameter. Some reasonable enough measures, such as $(a + \frac{1}{2})(d + \frac{1}{2})(b + \frac{1}{2})^{-1}(c + \frac{1}{2})^{-1}$, are not strictly homogeneous of degree zero, but all reasonable ones have the property asymptotically, and in this paper we consider only measures that are strictly homogeneous of degree zero. The term "homogeneous" is used in several senses and it is further discussed by Good (1986b).

Symmetry. A measure α is called *symmetric* if identically $\alpha(\mathbf{a}) = \alpha(\mathbf{a}')$, where $\mathbf{a}'$ denotes the transpose of $\mathbf{a}$, namely $[a, c;\ b, d]$. Symmetric measures are especially suitable for noncausal or attribute tables.

Row-scale and column-scale invariance. The measure α is called *row-scale* (*column-scale*) *invariant* if it is unchanged when the ith row (ith column) is multiplied by λ_i $(i = 1, 2)$.

When sampling procedure II_R (II_C) is used, then α should be row-scale (column-scale) invariant because otherwise α would depend on the arbitrarily chosen ratio of the row totals. When sampling procedure III is used, α should be both row-scale and column-scale invariant.

We now define the measures of concern in the body of this paper with some, but not exhaustive, indication of their practical interest. For a discussion of further properties of these measures and some history see Good (1985c).

(i) *Peirce's measure.* We write

$$\pi_R = \pi_R(\mathbf{a}) = \frac{a}{a+b} - \frac{c}{c+d}. \tag{3.1}$$

Under sampling procedure I or II_R, this expression is equal to

$$P(S|T) - P(S|\overline{T}), \tag{3.2}$$

which is a reasonable measure of superiority of treatment T over treatment $\overline{T}$ in an R-causal table. It can of course be negative and then $|\pi_R|$ measures an inferiority. When multiplied by 100, Peirce's measure is sometimes called the "percentage difference" [Somers (1978)]. It was introduced by Peirce (1884) and was cited by Goodman and Kruskal (1959, pages 129–130). For an interpretation of π_R that does not depend on a causal direction we refer to Somers (1978).

Under sampling procedure II_C, the interpretation (3.2) will be approximately correct if the ratio of the column totals is close to the ratio $P(S)/P(\overline{S})$ in all the subpopulations. The measure π_R, which is obviously row-scale invariant, is naturally designed for R-causal tables, but even for C-causal tables π_R can play a detective role. Similar comments apply to the other measures to be defined in this section.

We define analogously

$$\pi_C = \pi_C(\mathbf{a}) = \frac{a}{a+c} - \frac{b}{b+d}, \tag{3.3}$$

which can be interpreted as

$$P(T|S) - P(T|\bar{S}). \tag{3.4}$$

(ii) *Yule's measures.* Yule (1903) considered the excess of the first "cell probability" a/N over its expected value under independence of rows and columns. That is,

$$y = y(\mathbf{a}) = \frac{a}{N} - \frac{(a+b)(a+c)}{N^2} = \frac{ad-bc}{N^2}. \tag{3.5}$$

Apart from sign, it is the same for all four cells of the table. It is symmetrical between rows and columns and hence is especially suitable for noncausal tables. It is neither row- nor column-scale invariant.

(iii) *The odds ratio or cross-product ratio*

$$\kappa = \kappa(\mathbf{a}) = \frac{ad}{bc}. \tag{3.6}$$

Under any sampling procedure we can interpret κ as

$$\frac{O(S|T)}{O(S|\bar{T})} = \frac{O(T|S)}{O(T|\bar{S})},$$

where O denotes odds within the population.

Edwards (1963) showed that a measure of association α must be a function of κ if it satisfies both the following conditions: α is a function of $P(S|T)$ and $P(S|\bar{T})$ alone, and is also a function of $P(T|S)$ and $P(T|\bar{S})$ alone. Symmetry, as defined above, is a weaker condition than this one.

The odds ratio is both row-scale and column-scale invariant, and the only measures that are both row-scale and column-scale invariant are functions of κ.

PROOF. If $\mathbf{a} = \mathbf{a}(a, b; c, d)$ is row-scale and column-scale invariant, we must have

$$\alpha(a, b; c, d) = \alpha(\lambda_1\mu_1 a, \lambda_1\mu_2 b; \lambda_2\mu_1 c; \lambda_2\mu_2 d) \tag{3.7}$$

for all positive numbers λ_1, λ_2, μ_1 and μ_2. Take $\lambda_1^4 = d/(ab^2)$, $\lambda_2^4 = b/(cd^2)$, $\mu_1^4 = d/(ac^2)$ and $\mu_2^4 = c/(bd^2)$. Then the right side of (3.7) reduces to $\alpha(\kappa^{1/2}, \kappa^{-1/4}; \kappa^{-1/4}, 1)$, *so α is a function of κ alone.* This result can also be derived from Edwards (1963). □

Thus functions of κ are the only measures that are clearly reasonable when sampling procedure III is used. An example of such functions is Y_β [Good (1985c)], which generalizes two association measures used by Yule.

(iv) *Weight of evidence.* Let

$$W_C = W_C(\mathbf{a}) = \log\frac{a(b+d)}{b(a+c)}, \tag{3.8}$$

which, under sampling procedure I or II_C, is equal to

$$\log\left[\frac{P(T|S)}{P(T|\bar{S})}\right] \tag{3.9}$$

and can be interpreted as the prognostic weight of evidence (logarithm of the Bayes factor) in favor of S provided by the knowledge that the treatment was T. For the terminology "weight of evidence" in this (or near) sense, see, for example, a sentence in Peirce (1878, page 1345), a much fuller discussion in Good (1950, 1985b) and, for a careful justification of the terminology, Good (1984).

We write analogously,

$$W_R = W_R(\mathbf{a}) = \log\frac{a(c+d)}{c(a+b)}, \tag{3.10}$$

which, under sampling procedure I or II_R, is equal to

$$\log\left[\frac{P(S|T)}{P(S|\bar{T})}\right]. \tag{3.11}$$

(v) *Causal propensity.* Let

$$Q_R = Q_R(\mathbf{a}) = \log\frac{d(a+b)}{b(c+d)}, \tag{3.12}$$

which, under sampling procedure I or II_R, is

$$\log\left[\frac{P(\bar{S}|\bar{T})}{P(\bar{S}|T)}\right], \tag{3.13}$$

the *propensity of T to cause S rather than $\bar{S}$*, or the weight of evidence against T if $\bar{S}$ occurs [Good (1961, 1986c)]. It should be noted that it is possible for the direction of causality to be unclear or to be two-way. For example, the rows might represent higher and low I.Q. and the column high and low mathematical performance. In such cases Q_R and Q_C might both be suggestive of causal relationships. Without explicit reference to causality and without the logarithm, it was proposed earlier and independently by Sheps (1958) with the name "survival ratio," but "benefit ratio" or "effectiveness ratio" would be a more flexible name. Analogously, let

$$Q_C = Q_C(\mathbf{a}) = \log\frac{d(a+c)}{c(b+d)}, \tag{3.14}$$

which, under sampling procedure I or II_C, is equal to

$$\log\left[\frac{P(\bar{T}|\bar{S})}{P(\bar{T}|S)}\right]. \tag{3.15}$$

4. Avoidance of the paradox. Our query in this area arose from a conjecture of one of the authors and its simple geometrical proof by the other author. The conjecture was that if $(a_1 + b_1)/(c_1 + d_1) = (a_2 + b_2)/(c_2 + d_2)$

then $a_1d_1 > b_1c_1$ and $a_2d_2 > b_2c_2$ together imply $(a_1 + a_2)(d_1 + d_2) > (b_1 + b_2)(c_1 + c_2)$. Geometrically this says that, given two parallelograms, each with a vertex at the origin, and with vertices at (a_1, c_1) and (b_1, d_1) for one parallelogram, and (a_2, c_2) and (b_2, d_2) for the other, and if (i) these two parallelograms have collinear diagonals, (ii) the points (a_1, c_1) and (a_2, c_2) lie below the common diagonal (because the areas $a_1d_1 - b_1c_1$ and $a_2d_2 - b_2c_2$ are positive), then the point $(a_1 + a_2, c_1 + c_2)$ also lies below the diagonal. The theorems given in this section on avoidance of a paradox are stronger and their proofs are algebraic. A geometrical approach was pursued for some of these results and can be found in Good (1986a).

We now give conditions for avoiding the paradox for each of the measures of association defined in Section 3. We have n subpopulations with corresponding two-by-two contingency tables $\mathbf{a}_i$ amalgamated into a table $\mathbf{A}$, this and other notation being as in the Introduction.

The following obvious algebraic fact *will be used repeatedly* so we call it a lemma although we usually use it without citing it.

LEMMA 4.1. *If t, u, v and w are positive numbers then $(t + u)/(v + w)$ is a "convex combination" of t/v and u/w, namely*

$$\beta t/v + (1 - \beta)u/w,$$

where $\beta = v/(v + w)$.

THEOREM 4.1. *For row-uniform designs $\alpha(\mathbf{A})$ is the natural weighted average of the $\alpha(\mathbf{a}_i)$'s, that is,*

$$\alpha(\mathbf{A}) = \sum_{i=1}^{n} (N_i/N)\alpha(\mathbf{a}_i), \tag{4.1}$$

where α is π_R or y. Hence in particular

$$\min_i \alpha(\mathbf{a}_i) \le \alpha(\mathbf{A}) \le \max_i \alpha(\mathbf{a}_i) \tag{4.2}$$

and the paradox cannot occur.

COROLLARY 4.1. *For column-uniform designs,* (4.1) *and* (4.2) *hold when α is π_C or y.*

PROOF OF THEOREM 4.1. From (2.1) and (3.1) we have

$$N_i = (\lambda + 1)(c_i + d_i)$$

and

$$\pi_R(\mathbf{a}_i) = \frac{a_i/\lambda - c_i}{c_i + d_i}, \qquad i = 1, 2, \ldots, n.$$

Thus

$$\begin{aligned} \pi_R(\mathbf{A}) &= \frac{(\Sigma a_i)/\lambda - \Sigma c_i}{\Sigma(c_i + d_i)} = \frac{\Sigma(c_i + d_i)\pi_R(a_i)}{\Sigma(c_i + d_i)} \\ &= \sum_{i=1}^{n} \frac{N_i}{N}\pi_R(a_i). \end{aligned} \tag{4.3}$$

Since for row-uniform designs

$$\pi_R(\mathbf{a}) = y(a)(\lambda + 1)^2/\lambda,$$

the result for y follows from (4.3). □

THEOREM 4.2. *For row-uniform designs*

$$\exp[\alpha(\mathbf{A})] = \sum_{i=1}^{n} \delta_i \exp[\alpha(\mathbf{a}_i)] \tag{4.4}$$

for some $\delta_i > 0$ *with* $\sum_{i=1}^{n}\delta_i = 1$, *where* α *is* Q_R *or* W_R. *In particular*, (4.2) *holds for* $\alpha = Q_R$ *or* W_R *and the paradox cannot occur.*

COROLLARY 4.2. *For column-uniform designs*, (4.4) *and* (4.2) *hold for* $\alpha = Q_C$ *or* W_C *and the paradox cannot occur.*

PROOF OF THEOREM 4.2. For row-uniform designs we can write

$$\exp(Q_R(\mathbf{a}_i)) = \lambda d_i/b_i \quad \text{and} \quad \exp(Q_R(\mathbf{A})) = \lambda \sum d_i / \sum b_i.$$

Thus (4.4) holds with $\delta_i = b_i/(\Sigma b_i)$. Also

$$\exp(W_R(\mathbf{a}_i)) = \lambda^{-1} a_i/c_i \quad \text{and} \quad \exp(W_R(\mathbf{A})) = \lambda^{-1} \sum a_i / \sum c_i.$$

So (4.4) is true with $\delta_i = c_i/(\Sigma c_i)$. □

THEOREM 4.3. *If the design is both row-uniform and column-uniform then* (4.2) *holds for* $\alpha = \kappa$.

NOTE. The following example shows that just row-uniform or column-uniform design is not sufficient for Theorem 4.3. Let $\mathbf{a}_1 = [3,1;\ 1,9]$ and $\mathbf{a}_2 = [889,203;\ 381,2349]$. Then $\kappa(\mathbf{a}_1) = \kappa(\mathbf{a}_2) = 27$, and the design is row-uniform with $\lambda = 0.4$, but $\kappa(\mathbf{A}) = 26.991$.

PROOF OF THEOREM 4.3. We prove the theorem for $n = 2$. For $n > 2$, we can first amalgamate two tables and then add further tables one at a time to get the final result. Let

$$x_1 = a_1/a_2, \qquad x_2 = b_1/b_2, \qquad x_3 = c_1/c_2, \qquad x_4 = d_1/d_2.$$

For $n = 2$, we rewrite the conditions of row-uniform and column-uniform design as

$$x_1\delta + x_2(1 - \delta) = x_3\delta' + x_4(1 - \delta') \tag{4.5}$$

for $\delta = a_2/(a_2 + b_2) > 0$ and $\delta' = c_2/(c_2 + d_2) > 0$ (see Lemma 1);

$$x_1\eta + x_3(1 - \eta) = x_2\eta' + x_4(1 - \eta') \tag{4.6}$$

for $\eta = a_2/(a_2 + c_2) > 0$ and $\eta' = b_2/(b_2 + d_2) > 0$. Equation (4.5) shows that the intervals (x_1, x_2) and (x_3, x_4) must overlap (or, as a special case, $x_1 = x_2 = x_3 = x_4$), while (4.6) shows that the intervals (x_1, x_3) and (x_2, x_4) must overlap. Here the notation (x_1, x_2) is not intended to imply that $x_1 \le x_2$, etc.

Without loss of generality we shall assume that $\kappa(\mathbf{a}_1) \leq \kappa(\mathbf{a}_2)$, that is,

$$x_1 x_4 \leq x_2 x_3. \tag{4.7}$$

We shall show that

$$\kappa(\mathbf{a}_1) \leq \kappa(\mathbf{A}) \leq \kappa(\mathbf{a}_2). \tag{4.8}$$

First, let us assume that $x_1 \leq x_4$; then x_2 and x_3 cannot both lie in the interval $[x_1; x_4]$ since $x_1 \leq x_2 \leq x_3 \leq x_4$ violates (4.5), while $x_1 \leq x_3 \leq x_2 \leq x_4$ violates (4.6). The only exception is when $x_1 = x_2 = x_3 = x_4$, but the result (4.8) is then trivially true. For the same reasons it is not possible that one of x_2, x_3 is less than x_1 and the other greater than x_4. If both x_2 and x_3 are less than x_1, then (4.7) will be violated. Thus x_2 and x_3 both have to exceed x_4. Accordingly we are left with only two cases,

$$x_1 \leq x_4 < x_2 \leq x_3 \tag{4.9}$$

or

$$x_1 \leq x_4 < x_3 \leq x_2. \tag{4.10}$$

Arguing similarly for the case $x_4 \leq x_1$, we get two more cases, namely

$$x_4 \leq x_1 < x_2 \leq x_3 \tag{4.11}$$

or

$$x_4 \leq x_1 < x_3 \leq x_2. \tag{4.12}$$

For each one of the above four cases, we will show that (4.8) holds. We note that in all four of the cases (4.9), (4.10), (4.11) and (4.12) we have $x_1 \leq x_3$ and $x_4 \leq x_2$. This implies that $a_1/c_1 \leq a_2/c_2$ and $d_1/b_1 \leq d_2/b_2$. By use of Lemma 4.1 we see that $a_1/c_1 \leq (a_1 + a_2)/(c_1 + c_2) \leq a_2/c_2$ as well as $d_1/b_1 \leq (d_1 + d_2)/(b_1 + b_2) \leq d_2/b_2$, and (4.8) follows readily. □

5. Homogeneity of two subpopulations. As in Definition 1.2 we call the two subpopulations homogeneous with respect to α if

$$\alpha(\mathbf{a}_1) = \alpha(\mathbf{a}_2) = \alpha(\mathbf{A}), \tag{5.1}$$

where $\mathbf{a}_1, \mathbf{a}_2$ denote the two-by-two contingency tables of the subpopulations and $\mathbf{A}$ is the amalgamated table. [For a related discussion of homogeneity see Good, (1986b).] We shall give necessary and sufficient conditions for (5.1) when α is any one of the measures π_R, π_C, W_R, W_C, Q_R, Q_C or κ.

THEOREM 5.1. *For the odds ratio* κ, (5.1) *holds if and only if either*

$$a_1/c_1 = a_2/c_2 \quad \textit{and} \quad b_1/d_1 = b_2/d_2 \tag{5.2}$$

(for our example, this can be interpreted as "conditional on success, sex and treatment are independent") or

$$a_1/b_1 = a_2/b_2 \quad \textit{and} \quad c_1/d_1 = c_2/d_2 \tag{5.3}$$

(i.e., "conditional on treatment, sex and success are independent").

PROOF. By the definition of homogeneity with respect to κ we have

$$\frac{a_1 d_1}{b_1 c_1} = \frac{a_2 d_2}{b_2 c_2} = \frac{(a_1 + a_2)(d_1 + d_2)}{(b_1 + b_2)(c_1 + c_2)} = k, \quad \text{say.} \tag{5.4}$$

Substituting $a_1 = b_1 c_1 k/d_1$ and $a_2 = b_2 c_2 k/d_2$ in the last part of (5.4) we have

$$(b_1 c_1 k/d_1 + b_2 c_2 k/d_2)(d_1 + d_2) = k(b_1 + b_2)(c_1 + c_2).$$

This simplifies to

$$(b_1 d_2 - b_2 d_1)(c_1 d_2 - c_2 d_1) = 0.$$

Thus either $b_1/d_1 = b_2/d_2$ [which implies $a_1/c_1 = a_2/c_2$ in view of (5.4)] or $c_1/d_1 = c_2/d_2$ [which implies $a_1/b_1 = a_2/b_2$ in view of (5.4) again]. To show sufficiency we reverse the argument or we can easily show directly that each of (5.2) and (5.3) imply (5.1). □

THEOREM 5.2. *For weight of evidence W_C,* (5.1) *holds if and only if either* (5.2) *holds or*

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} \quad \textit{and} \quad \frac{a_1 + c_1}{b_1 + d_1} = \frac{a_2 + c_2}{b_2 + d_2}, \tag{5.5}$$

where the second condition asserts column-uniformity. Similarly, for W_R, (5.1) *holds if and only if either* (5.3) *holds or*

$$\frac{a_1}{c_1} = \frac{a_2}{c_2} \quad \textit{and} \quad \frac{a_1 + b_1}{c_1 + d_1} = \frac{a_2 + b_2}{c_2 + d_2}. \tag{5.6}$$

PROOF. It is enough to prove the theorem for W_C because the result for W_R follows by interchanging rows with columns, that is, by interchanging b with c. Let us write $\exp(W_C(\mathbf{a})) = \xi/\eta$, where $\xi = 1 + d/b$ and $\eta = 1 + c/a$. If we let $\gamma = b_1/(b_1 + b_2)$ and $\delta = a_1/(a_1 + a_2)$, then (5.1) becomes

$$\frac{\xi_1}{\eta_1} = \frac{\xi_2}{\eta_2} = \frac{\gamma\xi_1 + (1 - \gamma)\xi_2}{\delta\eta_1 + (1 - \delta)\eta_2}. \tag{5.7}$$

We now have

$$\xi_1\eta_2 = \xi_2\eta_1 \tag{5.8}$$

and

$$\delta\xi_1\eta_1 + (1 - \delta)\xi_1\eta_2 = \gamma\xi_1\eta_1 + (1 - \gamma)\xi_2\eta_1. \tag{5.9}$$

Substituting in (5.9) from (5.8) we have

$$(\gamma - \delta)\xi_2\eta_1 = (\gamma - \delta)\xi_1\eta_1 \quad \text{and} \quad (\delta - \gamma)\xi_1\eta_1 = (\delta - \gamma)\xi_1\eta_2. \tag{5.10}$$

Since $\xi_1 \neq 0$ and $\eta_1 \neq 0$, (5.10) will be true if either $\delta = \gamma$ [i.e., $a_1/b_1 = a_2/b_2$ and this implies $(a_1 + c_1)/(b_1 + d_1) = (a_2 + c_2)/(b_2 + d_2)$ in view of (5.1)] or $\eta_1 = \eta_2$ and $\xi_1 = \xi_2$ (i.e., $a_1/c_1 = a_2/c_2$ and $b_1/d_1 = b_2/d_2$). The proof of sufficiency is trivial. □

THEOREM 5.3. *For the causal propensity* Q_R, (5.1) *holds if and only if either* (5.3) *holds or*

$$\frac{b_1}{d_1} = \frac{b_2}{d_2} \quad \text{and} \quad \frac{a_1 + b_1}{c_1 + d_1} = \frac{a_2 + b_2}{c_2 + d_2}. \tag{5.11}$$

Similarly for Q_C, (5.1) *holds if and only if either* (5.2) *holds or*

$$\frac{c_1}{d_1} = \frac{c_2}{d_2} \quad \text{and} \quad \frac{a_1 + c_1}{b_1 + d_1} = \frac{a_2 + c_2}{b_2 + d_2}. \tag{5.12}$$

We omit the proof since it is very similar to that of Theorem 5.2.

THEOREM 5.4. *For Peirce's measure* π_R, (5.1) *holds if and only if either* (5.3) *holds or*

$$\frac{a_1 + b_1}{c_1 + d_1} = \frac{a_2 + b_2}{c_2 + d_2} = \frac{a_1 - \mu a_2}{c_1 - \mu c_2}, \quad \text{where} \quad \mu = \frac{a_1 + b_1}{a_2 + b_2} = \frac{c_1 + d_1}{c_2 + d_2}. \tag{5.13}$$

Similarly for π_C, *by interchanging* b_1 *with* c_1 *and* b_2 *with* c_2, *we see that* (5.1) *holds if and only if either* (5.2) *holds or*

$$\frac{a_1 + c_1}{b_1 + d_1} = \frac{a_2 + c_2}{b_2 + d_2} = \frac{a_1 - \mu' a_2}{b_1 - \mu' b_2}, \quad \text{where} \quad \mu' = \frac{a_1 + c_1}{a_2 + c_2} = \frac{b_1 + d_1}{b_2 + d_2}. \tag{5.14}$$

PROOF. We can write (5.1) as

$$\frac{a_1}{a_1 + b_1} - \frac{c_1}{c_1 + d_1} = \frac{a_2}{a_2 + b_2} - \frac{c_2}{c_2 + d_2} \tag{5.15}$$

$$= \frac{a_1 + a_2}{a_1 + a_2 + b_1 + b_2} - \frac{c_1 + d_1}{c_1 + c_2 + d_1 + d_2}. \tag{5.16}$$

Splitting both terms of (5.16) as convex combinations, we see that the expression equals

$$\gamma \frac{a_1}{a_1 + b_1} + (1 - \gamma) \frac{a_2}{a_2 + b_2} - \delta \frac{c_1}{c_1 + d_1} - (1 - \delta) \frac{c_2}{c_2 + d_2}, \tag{5.17}$$

where

$$\gamma = \frac{a_1 + b_1}{a_1 + b_1 + a_2 + b_2}$$

and

$$\delta = \frac{c_1 + d_1}{c_1 + d_1 + c_2 + d_2}.$$

We also know that any convex combination of $\pi_R(\mathbf{a}_1) = \pi_R(\mathbf{a}_2)$ is equal to $\pi_R(\mathbf{A})$, since all three are equal. Thus equating the right-hand side of (5.17) to $\gamma \pi_R(\mathbf{a}_1) + (1 - \gamma)\pi_R(\mathbf{a}_2)$, we get

$$\gamma \frac{c_1}{c_1 + d_1} + (1 - \gamma) \frac{c_2}{c_2 + d_2} = \delta \frac{c_1}{c_1 + d_1} + (1 - \delta) \frac{c_2}{c_2 + d_2}. \tag{5.18}$$

That is,

$$(\delta - \gamma)\frac{c_1}{c_1 + d_1} = (\delta - \gamma)\frac{c_2}{c_2 + d_2}.$$

Thus we must have *either* (i) $\delta = \gamma$, which is equivalent to

$$\frac{a_1 + b_1}{c_1 + d_1} = \frac{a_2 + b_2}{c_2 + d_2} = \lambda, \quad \text{say,}$$

in which case, from (5.15) we see that

$$\frac{a_1}{\lambda(c_1 + d_1)} - \frac{c_1}{c_1 + d_1} = \frac{a_2}{\lambda(c_2 + d_2)} - \frac{c_2}{c_2 + d_2},$$

so that

$$\frac{a_1 - \lambda c_1}{a_2 - \lambda c_2} = \frac{c_1 + d_1}{c_2 + d_2}$$

and hence $\lambda = (a_1 - \mu a_2)/(c_1 - \mu c_2)$, where μ is defined in (5.13) *or*

(ii) $$\frac{c_1}{c_1 + d_1} = \frac{c_2}{c_2 + d_2},$$

which is equivalent to $c_1/d_1 = c_2/d_2$ and implies $a_1/b_1 = a_2/b_2$ in view of (5.15), and this gives (5.3). □

The proof of sufficiency for Theorem 5.4 follows by reversing the argument. The following theorem follows easily.

THEOREM 5.5. *If* (5.2) *holds, then the two subpopulations are homogeneous with respect to* κ, W_C, Q_C *and* π_C. *If* (5.3) *holds then the two subpopulations are homogeneous with respect to* κ, W_R, Q_R *and* π_R.

THEOREM 5.6. *Under either row-uniform or column-uniform design, if the two subpopulations are homogeneous with respect to any three of our seven measures, then either*

(i) $a_1/a_2 = b_1/b_2 = c_1/c_2 = d_1/d_2$ (*i.e., the three tables are "essentially the same," that is, are proportional to one another*); *or*

(ii) $a_1/c_1 = b_1/d_1 = a_2/c_2 = b_2/d_2$ (*in particular, the rows and columns are "independent" within all three tables*).

PROOF. First we show that if the two subpopulations are homogeneous with respect to any three of the above measures, then either (5.2) or (5.3) must hold. For convenience we first abbreviate Theorems 5.1 to 5.4 in a self-explanatory way: $\kappa \Rightarrow$ (5.2) or (5.3); $W_C \Rightarrow$ (5.2) *or* (5.5); $Q_C \Rightarrow$ (5.2) or (5.12); $\pi_C \Rightarrow$ (5.2) or (5.14); $W_R \Rightarrow$ (5.3) or (5.6); $Q_R \Rightarrow$ (5.3) or (5.11); and $\pi_R \Rightarrow$ (5.3) or (5.13). We see at once that when three sets of these conditions hold simultaneously then either (5.2) or (5.3) or three of the six conditions (5.5), (5.6), (5.11), (5.12), (5.13) and

(5.14) must hold. If the three are from (5.5), (5.6), (5.11) and (5.12) then it is easy to see that (5.2) or (5.3) must hold. A little reflection shows that the only other possibility, for other combinations of three, is that both

$$\frac{a_1 + b_1}{a_2 + b_2} = \frac{c_1 + d_1}{c_2 + d_2} \tag{5.19}$$

and

$$\frac{a_1 + c_1}{a_2 + c_2} = \frac{b_1 + d_1}{b_2 + d_2} \tag{5.20}$$

must hold and at least one of the four equalities

$$x_1 = x_3, \quad x_2 = x_4, \quad x_1 = x_2, \quad x_3 = x_4, \tag{5.21}$$

holds, where, as before,

$$x_1 = a_1/a_2, \quad x_2 = b_1/b_2, \quad x_3 = c_1/c_2, \quad x_4 = d_1/d_2.$$

But (5.19) and (5.20) can be written in the forms

$$\delta x_1 + (1 - \delta)x_2 = \gamma x_3 + (1 - \gamma)x_4 \tag{5.22}$$

and

$$\delta' x_1 + (1 - \delta')x_3 = \gamma' x_2 + (1 - \gamma')x_4, \tag{5.23}$$

where δ, δ', γ and γ' all lie between 0 and 1.

If $x_1 = x_3$ then the second equation shows that x_3 lies between x_2 and x_4, but the first equation will then be violated unless $x_1 = x_2 = x_3 = x_4$ [that is, (i) holds]. We reach the same conclusion from each of the remaining three equalities in (5.21).

We finish the proof by showing that if one of (5.2) and (5.3) and one of (5.19) and (5.20) holds then (i) or (ii) in the statement of Theorem 5.6 must be true.

If (5.2) (that is, $x_1 = x_3$ and $x_2 = x_4$) and (5.22) hold then $\delta = \gamma$, that is, $a_2/(a_2 + b_2) = c_2/(c_2 + d_2)$ and hence $a_2/c_2 = (a_2 + b_2)/(c_2 + d_2) = b_2/d_2$, but (5.2) implies $a_1/c_1 = a_2/c_2$ and $b_1/d_1 = b_2/d_2$. Thus (ii) must hold. If (5.2) and (5.23) hold then $x_1 = x_3 = x_2 = x_4$ and (i) must hold. Similar arguments show that if (5.3) and (5.22) or (5.3) and (5.23) hold then (i) or (ii) hold in each case. This completes the proof of Theorem 5.6. □

APPENDIX[3]

The effect of small variations on the amalgamation paradox. Here we discuss the conditions of row-fair and column-fair designs further and look at the effect of small variations on the amalgamation paradox.

It is possible to design an experiment in which (2.1) holds [the discussion for condition (2.2) is similar] if the sampling procedure employed is Π_R, that is, the row totals are held fixed. For example, the statistician can choose the sample size N_i for the ith subpopulation as the minimum acceptable ($\geq n_0$, say) number for which pN_i is an integer, $0 < p < 1$ being fixed. Then the statistician can choose

[3]Written by the second-named author.

pN_i individuals at random from N_i and assign them to the treatment T. Such a design will obviously satisfy (2.1) for $\lambda = p/(1-p)$.

However, in many other practical situations, we may find that the equality in (2.1) is only approximately correct. For instance, in the preceding example, if the treatment were given to each one of the individuals from the ith subpopulation with a fixed probability p (i.e., the treatment is randomized), then we would find that (2.1) holds approximately if N_i is large for each i. Such a situation can occur in some retrospective studies as well. Suppose an equal proportion p of people were inoculated (T) against the Asian flu in both the U.S. and Canada. Later random samples were drawn from each country to study the adverse effects (S) of the inoculation. If the sample size from each country is large enough then we will find this design to be "approximately row-fair" [i.e., the equality in (2.1) is replaced by $\simeq$].

It is intuitively clear that whenever a row-fair design is a sufficient condition for (4.2), the approximate row-fair design should be sufficient as well for large enough sample sizes. In the following we make this idea explicit.

Let us first consider only two subpopulations with sample sizes N_1 and N_2. We define the design (which yielded the two contingency tables $\mathbf{a}_1$ and $\mathbf{a}_2$) to be *approximately row-fair* if

$$\text{(A1)} \qquad \frac{a_1 + b_1}{c_1 + d_1} = \lambda_1 \simeq \lambda_2 = \frac{a_2 + b_2}{c_2 + d_2}.$$

(We define an *approximate column-fair* design in a similar fashion. The analog of the following result is obvious and hence excluded.)

THEOREM A.1. *If the design is approximately row-fair, then*

$$\text{(A2)} \qquad \min(\alpha(\mathbf{a}_1); \alpha(\mathbf{a}_2)) \le \alpha(\mathbf{A}) \le \max(\alpha(\mathbf{a}_1); \alpha(\mathbf{a}_2)),$$

where $\mathbf{A} = [A, B; C, D]$ *is the amalgamated table and* α *is any one of the measures* π_R, W_R, Q_R *or* y.

PROOF. We first reformulate the problem in terms of the quantities of the amalgamated table **A**. Define $p = (A + B)/N$, $N = N_1 + N_2$, and $\lambda = p/(1-p)$. *All* tables (with N_1 and N_2 fixed) whose row-ratios are close to λ *and* whose amalgamated table is **A** must be of the form

Subpopulation 1			Subpopulation 2		
	S	$\bar{S}$		S	$\bar{S}$
T	a_1	$(p-\delta)N_1 - a_1$	T	a_2	$\left(p + \delta\frac{N_1}{N_2}\right)N_2 - a_2$
$\bar{T}$	c_1	$(1-p+\delta)N_1 - c_1$	$\bar{T}$	c_2	$\left(1-p-\delta\frac{N_1}{N_2}\right)N_2 - c_2$

where δ could be positive or negative. Let us call the two tables $\mathbf{a}_1^\delta$ and $\mathbf{a}_2^\delta$, respectively. Note that the values of a_1, c_1, a_2, c_2 are *not* fixed. They may vary

for different subpopulations as well as with N_1 and N_2 for the same subpopulations (see the example in the Introduction). We prove the result for *all* such tables. We introduce some more notation here which is in direct conflict with (A2). This should be allowed, however, since we are reformulating the problem. Let $\mathbf{a}_1 = [a_1, b_1;\ c_1, d_1]$ and $\mathbf{a}_2 = [a_2, b_2;\ c_2, d_2]$, where $b_i = pN_i - a_i$ and $d_i = (1-p)N_i - c_i$, $i = 1,2$. We will call $\mathbf{a}_1$ and $\mathbf{a}_2$ the "ideal" tables corresponding to $\mathbf{a}_1^\delta$ and $\mathbf{a}_2^\delta$. (The word "ideal" is used just for convenience and by no means implies a connection with any specific populations *or* with $\mathbf{a}_i^\delta$ for large N_i, $i = 1,2$. Also there is one such table $\mathbf{a}_i$ for each $\mathbf{a}_i^\delta$ but many such for the given amalgamation $\mathbf{A}$.) The amalgamation of the ideal tables is $\mathbf{A}$ as well and their row ratios are constant, equal to λ. By Section 4 we have

$$\alpha(\mathbf{a}_1) \le \alpha(\mathbf{A}) \le \alpha(\mathbf{a}_2) \tag{A3}$$

for all such ideal tables. [Without loss of generality, we have chosen the subscripts so that $\alpha(\mathbf{a}_1) \le \alpha(\mathbf{a}_2)$.] Theorem A.1 will be proved if we show that for fixed N_1, N_2 and p,

$$\min\big(\alpha(\mathbf{a}_1^\delta), \alpha(\mathbf{a}_2^\delta)\big) \le \alpha(\mathbf{A}) \le \max\big(\alpha(\mathbf{a}_1^\delta), \alpha(\mathbf{a}_2^\delta)\big) \tag{A4}$$

for δ small enough.

In case $\alpha(\mathbf{a}_1) < \alpha(\mathbf{A}) < \alpha(\mathbf{a}_2)$, the proof is obvious from the following observation. For $\alpha = \pi_R$, W_R, Q_R or y, we have

$$\frac{d}{d\delta}\alpha(\mathbf{a}_1^\delta) > 0 \quad \text{and} \quad \frac{d}{d\delta}\alpha(\mathbf{a}_2^\delta) < 0$$

for $\delta \neq 0$ small. Also for $\delta = 0$, $\alpha(\mathbf{a}_i^\delta) = \alpha(\mathbf{a}_i)$, $i = 1,2$, and (A4) is true in view of (A3). On the other hand if it were the case that $\alpha(\mathbf{a}_1) = \alpha(\mathbf{A}) = \alpha(\mathbf{a}_2)$ then $\alpha(\mathbf{a}_1^\delta) < \alpha(\mathbf{A}) < \alpha(\mathbf{a}_2^\delta)$ for $\delta < 0$ small and the reverse inequalities would be true for $\delta > 0$ small. [Note that it cannot be the case that $\alpha(\mathbf{a}_1) < \alpha(\mathbf{A}) = \alpha(\mathbf{a}_2)$ or $\alpha(\mathbf{a}_1) = \alpha(\mathbf{A}) < \alpha(\mathbf{a}_2)$ for row-fair designs and $\alpha = \pi_R$, W_R, Q_R or y.] The extension of the case when there are k subpopulations is obvious.

It is also illuminating to write $\alpha(\mathbf{a}_i^\delta)$, $i = 1,2$, as the sum of $\alpha(\mathbf{a}_i)$ and a function of δ since it would provide some guidance for choosing the values of N_1 and N_2. We write this below for $\alpha = \mathrm{II}_R$ to illustrate the point. We have

$$\mathrm{II}_R(\mathbf{a}_1^\delta) = \mathrm{II}_R(\mathbf{a}_1) + \frac{\delta}{(1-p)}\left[\frac{a_1}{\lambda(a_1 + b_1 - \delta N_1)} + \frac{c_1}{c_1 + d_1 + \delta N_1}\right]. \tag{A5}$$

We note that the term in the brackets of the right-hand side of (A5) can be bounded above for *all* tables $\mathbf{a}_1$ if δ is small enough. For example if $|\delta| < \frac{1}{2}\min(p, 1-p)$, then the term in the brackets is at most

$$\frac{a_1 + b_1}{\lambda(a_1 + b_1 - \delta_1)} + \frac{c_1 + d_1}{c_1 + d_1 + \delta N_1} < \frac{2}{\lambda} + 2.$$

Similarly, we can write

$$\pi_R(\mathbf{a}_2^\delta) = \pi_R(\mathbf{a}_2) - \frac{\delta N_1}{(1-p)N_2}\left[\frac{a_2}{\lambda(a_2 + b_2 + \delta N_1)} + \frac{c_2}{c_2 + d_2 - \delta N_1}\right]. \tag{A6}$$

Again the term in the brackets is at most

$$\frac{pN_2}{\lambda(pN_2 + \delta N_1)} + \frac{(1-p)N_2}{(1-p)(N_2 - \delta N_1)}.$$

If we were in one of the situations described in the Introduction, then δ would be a function of the standard error of the estimates of p in the two subpopulations and this would converge to zero as N_1 and N_2 tend to infinity. □

Finally we prove the following theorem.

THEOREM A.2. *If the design is both approximately row-fair and approximately column-fair, then* (A2) *holds for* $\alpha = \kappa$, *the odds ratio.*

PROOF. We reformulate the problem in the manner similar to the above and define the contingency tables corresponding to the two subpopulations as

Subpopulation 1 Subpopulation 2

$$\begin{array}{c|c} a_1 & b_1 - \delta N_1 \\ \hline c_1 - \varepsilon N_1 & d_1 + (\varepsilon\delta)N_1 \end{array} \qquad \begin{array}{c|c} a_2 & b_2 + \delta N_1 \\ \hline c_2 + \varepsilon N_2 & d_2 - (\varepsilon + \delta)N_1 \end{array},$$

where $b_i = pN_i - a_i$, $c_i = qN_i - a_i$ and $d_i = (1-p-q)N_i$ for $i = 1, 2$. Again the corresponding ideal tables are $\mathbf{a}_1$ and $\mathbf{a}_2$ for which $\kappa(\mathbf{a}_1) \leq \kappa(\mathbf{A}) \leq \kappa(\mathbf{a}_2)$. We note here that $(\partial/\partial\delta)\kappa(\mathbf{a}_1^{\varepsilon,\delta}) > 0$, $(\partial/\partial\varepsilon)\kappa(\mathbf{a}_1^{\varepsilon,\delta}) > 0$, $(\partial/\partial\delta)\kappa(\mathbf{a}_2^{\varepsilon,\delta}) < 0$ and $(\partial/\partial\varepsilon)\kappa(\mathbf{a}_2^{\varepsilon,\delta}) < 0$, the inequalities for the partial derivatives wrt δ being true for all ε small and for all $\delta \neq 0$ small; similarly for partial derivatives wrt ε. The proof now follows along similar lines. Again $\kappa(\mathbf{a}_i^{\varepsilon,\delta})$ can be written as the sum of $\kappa(\mathbf{a}_i)$ and a function of δ and ε for $i = 1, 2$. We omit the details. □

REFERENCES

BLYTH, C. (1972). On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* **67** 364–366.

CARTWRIGHT, N. (1979). Causal laws and effective strategies. *Nous* **13** 419–437.

COHEN, M. R. and NAGEL, E. (1934). *An Introduction to Logic and Scientific Method.* Harcourt, New York.

EDWARDS, A. W. F. (1963). The measure of association in a 2 × 2 table. *J. Roy. Statist. Soc. Ser. A* **126** 109–114.

FISHER, R. A. (1949). *The Design of Experiments*, 5th ed. Oliver and Boyd, Edinburgh.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* Griffin, London.

GOOD, I. J. (1961). A causal calculus. *British J. Philos. Sci.* **11** 305–318; **12** 43–51; **13** 88 (1962).

GOOD, I. J. (1984). The best explicatum for weight of evidence. *J. Statist. Comput. Simulation* **19** 294–299.

GOOD, I. J. (1985a). Who discovered the God–Koheleth law? *J. Statist. Comput. Simulation* **21** 87–88.

GOOD, I. J. (1985b). Weight of evidence: A brief survey (with discussion). In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 249–269. North-Holland, Amsterdam.

GOOD, I. J. (1985c). Mathematically natural generalizations of some measures of association and dependence for contingency tables. *J. Statist. Comput. Simulation* **22** 93–97.

GOOD, I. J. (1986a). A geometrical approach to the amalgamation paradox. *J. Statist. Comput. Simulation* **26** 129–132.

GOOD, I. J. (1986b). Recognizability, permutability, and homogeneity, *J. Statist. Comput. Simulation* **26** 133–134.

GOOD, I. J. (1986c). Causal propensity: A review. *Proc. Conference of the Philosophy of Science Association* (P. D. Askwith and P. Kitcher, eds.) 829–850.

GOODMAN, L. A. and KRUSKAL, W. H. (1954). Measures of association for cross classifications. *J. Amer. Statist. Assoc.* **49** 723–764.

GOODMAN, L. A. and KRUSKAL, W. H. (1959). Measures of association for cross classifications. II: Further discussion and references. *J. Amer. Statist. Assoc.* **54** 123–163.

GOODMAN, L. A. and KRUSKAL, W. H. (1963). Measures of association for cross classifications. III: Approximate sampling theory. *J. Amer. Statist. Assoc.* **58** 310–364.

GOODMAN, L. A. and KRUSKAL, W. H. (1972). Measures of association for cross classification. IV: Simplification of asymptotic variances. *J. Amer. Statist. Assoc.* **67** 132–146.

HERBST, A. L., ULFELDER, H. and POSKANZER, D. C. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England J. Medicine* **284** 878–881.

HERBST, A. L., POSKANZER, D. C., ROBBOY, S. J., FRIEDLANDER L. and SCULLY, R. E. (1975). Prenatal exposure to stilbestrol: A prospective comparison of exposed female offspring with unexposed controls. *New England J. Medicine* **292** 334–339.

KENDALL, M. G. (1945). *The Advanced Theory of Statistics* **1**. Griffin, London.

MESSICK, D. M. and VAN DE GEER, J. P. (1981). A reversal paradox. *Psychol. Bull.* **90** 582–593.

PEARSON, K. (1899). Theory of genetic (reproductive) selection. *Philos. Trans. Roy. Soc. London Ser. A* **192** 260–278 (especially pages 277–278, On the spurious correlation produced by forming a mixture of heterogeneous but uncorrelated materials).

PEIRCE, C. S. (1878). The probability of induction. *Popular Science Monthly*. Reprinted in *The World of Mathematics* (J. R. Newman, ed.) **2** 1341–1354. Simon and Schuster, New York (1956).

PEIRCE, C. S. (1884). The numerical measure of success of predictions. *Science* **4** 453–454.

SHEPS, M. C. (1958). Shall we count the living or the dead? *New England J. Medicine* **259** 1210–1214.

SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.

SOMERS, R. H. (1978). Statistics, descriptive: Association. In *International Encyclopedia of Statistics* **2** 1109–1118. Wiley, New York.

STIGLER, S. M. (1980). Stigler's law of eponymy. *Trans. New York Acad. Sci. Ser. 2* **39** 147–158.

YULE, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika* **2** 121–134. Reprinted in *Statistical Papers of George Udny Yule*. 71–84. Griffin, London.

YULE, G. U. (1911). *An Introduction to the Theory of Statistics*. Griffin, London.

YULE, G. U. and KENDALL, M. G. (1950). *An Introduction to the Theory of Statistics*, 14th ed. Griffin, London.

DEPARTMENT OF STATISTICS
VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
BLACKSBURG, VIRGINIA 24061