PSYCHOMETRIKA—VOL. 62, NO. 1, 133–161
MARCH 1997

THE KNOWLEDGE CONTENT OF STATISTICAL DATA

LUCIEN PREUSS

FELDEGGSTRASSE 74, CH 8008 ZÜRICH, SWITZERLAND

HELMUT VORKAUF

UNIVERSITY OF FRIBOURG AND SWISS FEDERAL OFFICE OF PUBLIC HEALTH

An information-theoretic framework is used to analyze the knowledge content in multivariate cross classified data. Several related measures based directly on the information concept are proposed: the knowledge content (S) of a cross classification, its terseness (Zeta), and the separability (Gamma$_X$) of one variable, given all others. Exemplary applications are presented which illustrate the solutions obtained where classical analysis is unsatisfactory, such as optimal grouping, the analysis of very skew tables, or the interpretation of well-known paradoxes. Further, the separability suggests a solution for the classic problem of inductive inference which is independent of sample size.

Key words: cross classification, association, information, entropy, inference, paradox, significance.

1. Introduction

In a fundamental paper Lindley and Novick (1981) consider the basic problem of inference and analyze Simpson's well-known but still disquieting paradox about inference (cf. Exemplary Applications, section 7.1), stating that "The problem addressed (in this paper) is that of providing a formal framework within which related problems can systematically be resolved."

Consider the fundamental nature of Simpson's paradox, and of several other simple, yet seemingly untractable problems of inference, such as Lindley's (1957) paradox, the search for an optimal number of control cases (Fleiss, 1981), the search for an optimal formation of classes (Goodman & Kruskal, 1954), and Finley's tornado prediction (Goodman & Kruskal, 1959). Noting that most of them involve the use of a measure of association, the following declaration of intent remains as valid today as it was when Goodman and Kruskal stated it 40 years ago in their classic review "Measures of Association for Cross Classifications" or when it was reprinted 25 years later in 1979: "Our major theme is that the measures of association used by an empirical investigator should not be blindly chosen . . . , but should be constructed in a manner having operational meaning within the context of the particular problem" (preface).

The present paper revisits the fundamental inferential problems in this spirit and proposes an approach prompted by the statement: "To obtain a measure of association one must sharpen the definition of association, and this means that of the many vague intuitive notions of the concept some must be dropped." (Goodman & Kruskal, 1954, p. 742)

The statistical dependances found in a cross classification represent knowledge de-

Lucien Preuss gratefully acknowledges the support of the Swiss National Science Foundation (grant 21-25'757.88) which provided the initial impetus for this work. Also, we owe more than the usual thanks to the Editor, Shizuhiko Nishisato, whose relentless demands for clarity resulted in quite a few improvements.

Requests for reprints should be sent to Helmut Vorkauf, Swiss Federal Office of Public Health, Hessstrasse 27, CH 3097 Liebefeld/Bern, SWITZERLAND.

0033-3123/97/0300-94237$00.75/0

rived from past experience and can be used for a transfer of information between the variables. One must maintain a sharp distinction, however, between the description of dependance, generally expressed in terms of probabilities, and an evaluation of the dependance's strength, expressed as an amount of information about a dependent variable that can be derived from an independent one. Consider a 2 × 2 table with four identical frequencies: the probability to guess the value of one variable given the other has the sizeable value $\frac{1}{2}$, although there is clearly no association. Measures of association should reflect the amount of information that the knowledge of one variable provides about the other. Note that the probability to get certain results is not such a measure of association, but that Shannon's (1948) concept of information is. It defines information as the reduction of uncertainty or entropy from a distribution before an event to a distribution after it, irrespective of whether the event reveals the result of a single observation or of a whole set. This concept can be directly applied to a cross classification to evaluate the average information "w" obtained when passing from a mere knowledge of the margins to the unambiguous identification of a single event, and, perhaps more importantly, to distinguish two distinct parts of this information:

- The global restriction imposed on the distribution when the disclosure of a cross classification replaces the expected distribution with the observed frequencies, called the static content "S" of the cross classification, and
- The remainder ($w - S$), which represents the contribution of an actual observation, given the cross classification. This remainder is equal to the joint entropy of the cross classification.

Further, Shannon's concept can be used to derive a directed and normalized measure of dissimilarity for comparing any number of distributions for inductive inference. This dissimilarity measure, called "separability" and defined later, satisfies the following conditions:

- It is invariant when all frequencies are multiplied by the same factor. This is trivially necessary for a comparison between theoretical distributions for which only relative frequencies are defined, but it must also hold for observed distributions, because such distributions will not become more dissimilar when their frequencies are doubled or tripled.
- It is applicable to distributions which are not mutually absolutely continuous; this is necessary both for reasons of continuity and for generality.
- Its interpretation does not depend on a discontinuous parameter such as the degrees of freedom of a contingency table. Karl Pearson was apparently never quite satisfied with the accepted method of determining the number of degrees of freedom of a contingency table (Kendall, 1943, p. 305), possibly because an arbitrarily small alteration of the data allows one to inflate this number at will through the addition of cells with near-vanishing expected frequencies outside the original frame of the table. It is unsatisfactory in theory and poses problems in practice to avoid this problem by imposing some arbitrary lower bound for all expected frequencies.

This dissimilarity measure has been suggested in the past as a measure of dependency (Särndal, 1974; Press, Flannery, Teukolsky & Vetterling, 1989). In order to capitalize on the power and simplicity of the information approach, the separability of distributions must be analyzed in some depth and complemented with related parameters. The use of the following three measures then greatly simplifies the analysis of dependence between any number of variables and yields satisfactory and coherent solutions for several vexing problems and seemingly untractable paradoxes:

- The static content S of a multidimensional table to evaluate the information provided when the expected frequencies are replaced by observed frequencies. Monitoring the static content is essential when there is an abundance of variables, some of which must be removed to get a better overview (sec. 7.2). It also allows to compare the useful content of very skewed or otherwise unusual cross classifications with that of more familiar ones (sec. 7.4).
- The "terseness" of a multidimensional frequency table to measure the efficiency of the grid used to represent the observations. It is defined as the ratio of the maximal information that can be extracted from the cross classification through an appropriate input, to the amount of information that must be invested for the purpose. The terseness is needed to assess the desirability of deliberate modifications of the representation, as in searching for an optimal grouping (sec. 7.5), or when reducing the number of variables (sec. 7.1, 7.2).
- The separability to measure the dependence of any subset of variables on the complementary subset. It allows a systematic analysis of dependence in multidimensional cross classifications (sec. 7.2) and a comparison between widely different ones (sec. 7.4). Further, it is an absolute measure for the inhomogeneity of an ensemble of distributions (sec. 7.5) and, last not least, for the adequacy of a hypothesis (sec. 6.1).

The first two measures pertain to the cross classification as a whole and are symmetric in all its variables, whilst the last is a directed measure which characterizes a single variable, or a subset of variables. None depends on the number of observations, but their dependability (hence that of any conclusion based on them) clearly does. A theoretical derivation of their asymptotic standard errors seems desirable, but was not attempted here.

2. Entropy

The concept of entropy used in information theory (Khinchin, 1957; Schmitt, 1969; Shannon, 1948) will be sketched only very briefly. The entropy of a distribution is a quantitative measure for the uncertainty of its outcome. This uncertainty does not depend on any single probability, but on the homogeneity of the entire distribution. If the distribution encompasses only a few and widely different probabilities, the outcomes with larger probabilities will occur overwhelmingly often, and the uncertainty of the distribution becomes correspondingly low. Conversely, a distribution of events with nearly equal probabilities has a highly uncertain outcome; therefore, much information is delivered when an outcome becomes known.

The entropy H of a distribution $p_1, p_2, \ldots, p_i, \ldots, p_k$ with k distinct elements is:

$$\mathrm{H}[p_1, p_2, p_3, \ldots, p_k] = \sum \varphi(p_i), \tag{1a}$$

or, when using frequencies a_i instead of probabilities p_i:

$$\mathrm{H}[a_1, a_2, a_3, \ldots, a_k] = \frac{1}{N} \sum \varphi(a_i) + \ln N \tag{1b}$$

where

$$\varphi(a) = \begin{bmatrix} -a \cdot \ln(a) & \text{if } a > 0 \\ 0 & \text{if } a = 0 \end{bmatrix} \quad \text{and} \quad N = \sum a_i.$$

For a cross classification represented by a three-dimensional table with marginal sums $a_{i..}$, $a_{.j.}$, $a_{..k}$, the expected distribution e_{ijk} under mutual independence is

$$e_{ijk} = \frac{c}{N^2} \prod a_{i..} a_{.j.} a_{..k}. \tag{2}$$

As entropies depend on relative frequencies only (see (1a)), the factor c can be chosen arbitrarily, so that we can, without loss of generality, normalize the sum of all frequencies to 1 through a division by N. The entropy w of the expected or a priori distribution is the sum of the entropies of its marginal distributions, and this is the maximal uncertainty obtainable:

$$\begin{aligned} w &= \mathrm{H}(X) + \mathrm{H}(Y) + \mathrm{H}(Z) \\ &= \frac{1}{N}\left[\sum \varphi(a_{i..}) + \sum \varphi(a_{.j.}) + \sum \varphi(a_{..k})\right] + 3 \ln N. \end{aligned} \tag{3}$$

Once the actual or a posteriori frequencies a_{ijk} of a cross classification are known, its entropy is given by

$$u = \frac{1}{N} \sum \varphi(a_{ijk}) + \ln N. \tag{4}$$

This is called the joint entropy $\mathrm{H}(X, Y, Z)$ of the cross classification. Although three dimensions were used for convenience, the above equations extend to more dimensions in a straightforward manner.

3. Information

Information is a reduction of uncertainty caused by an event that replaces a prior distribution by a posterior distribution. Thus, if the result of an event is either not unique or cannot be determined with certainty, a residual posterior entropy remains and the information delivered by the event will be less than the full entropy of the a priori distribution. This includes the limiting case when nothing happens: the original distribution remains in force, and no information is generated. In communication theory, the existence of a residual entropy in any data after an event (e.g., the transmission of a message) nearly always implies a loss, for instance due to the noise in the channel. In contrast, statistics is the art of willfully introducing uncertainties, grouping similar but unequal results into classes with sometimes sizeable entropies, in order to gain a terser and more easily interpreted image of the observed phenomena. The extent of the grouping is a fundamental, if often neglected, question. Grouping is usually performed casually, without guidance from statistical theory.

A cross classification is a static representation of knowledge, derived from past observations and stored as more or less reliable linkages of the type: "If something is living, grey, and huge, then it is probably an elephant" or "A person inoculated against cholera is probably not befallen by cholera". Knowledge reflects the awareness of such stochastic links, and the ensemble of all links between the variables of a cross classification will be considered as its "knowledge content", a concept distinct from that of information because it is static while information exists only in status nascendi, when passing from one state to another.

The analysis of cross classifications then raises the following questions:

- How much information is contained in a cross classification, or, more precisely, how much information is conveyed when one is informed of its content?
- How thoroughly can the knowledge in a cross classification be exploited?
- Last not least, how strongly does any arbitrarily chosen variable or set of variables

depend on the ensemble of all others? How much can be inferred about the dependent variables without looking at them, merely from inspecting the independent variables?

All three questions are readily answered in terms of information, through the definition of three fundamental coefficients that characterize a cross classification:

- S, its knowledge content
- ζ, its terseness (Zeta)
- $\gamma_{XY\ldots}$, the separability (Gamma) of variables X, Y, . . . , given all others.

These coefficients will now be presented in some detail, and the solutions they provide for a number of unsolved problems will be examined.

4. Static Content of a Cross Classification

Given a cross classification with any number of variables, it is natural to ask for an evaluation of the amount of information generated by its disclosure, when passing from a mere knowledge of its margins to that of joint frequencies.

When only the marginal distributions are known, one must assume the distribution that exhibits the greatest possible uncertainty, or entropy, given the constraints exerted by the margins. This choice of the most non-committal prior distribution can be justified by Jaynes' (1978) principle of maximum entropy (Maxent). In other words, the entropy before the disclosure of the cross classification equals the joint entropy w of the expected distribution.

By definition, the entropy after the disclosure of a cross classification is equal to its joint entropy u, hence the information gain or uncertainty reduction, obtained through a knowledge of the cross classification, is:

$$S = w - u. \tag{5a}$$

This is the amount of information delivered when an "expected" distribution is replaced by a set of observed frequencies which reflect the linkages between the variables. As this information is delivered once and for all by the cross classification and does not depend on further input, it will be called the *static content* of the cross classification. According to (3) and (4) it can be calculated as follows for $D = 3$ variables:

$$S = w - u = \frac{1}{N}\left[\sum \varphi(a_{i..}) + \sum \varphi(a_{.j.}) + \sum \varphi(a_{..k}) - \sum \varphi(a_{ijk})\right] + (D - 1)\ln N. \tag{5b}$$

It is easy to generalize this to any number D of dimensions.

Like any measure of information, S depends on the base of the logarithms; natural logarithms will be used in this paper. Because S measures the difference in information between expected and observed distribution, it is closely related to what Fisher (1922) called "the departure of the sample from expectation". To evaluate this departure, Fisher (1922, 1956) and others (Edwards, 1972, 190–197; Wilks, 1935; Woolf, 1957) advocated "L", the logarithm of the likelihood ratio, which in two dimensions is identical to the product of the static content with the number N of observations. However, this inclusion of N into the measure obscures the essential distinction between:

- the dissimilarity between the expected and the observed distribution, and
- the dependability with which the dissimilarity can be inferred from available data.

Table 4.0

A

9	2
1	7

B

900	745
680	700

Clearly, dissimilarity does not depend on the number of observations, because multiplying all frequencies by a factor, even a large factor, does not change the distributions and will not make them more dissimilar. Hence a sharp distinction must be maintained between the magnitude of a dissimilarity and the dependability with which this magnitude is determined.

This is illustrated by Tables 4.0 A and B. Table 4.0 A displays a large dissimilarity between the expected and the observed frequency distribution, but the magnitude of this dissimilarity is not known precisely because few events have been observed and the knowledge they provide is correspondingly uncertain. Conversely, Table 4.0 B exhibits only a small dissimilarity which can, however, be assessed quite dependably. Yet the χ^2 of both tables is 8.9. The necessary distinction between the magnitude of a dissimilarity and the dependability with which it can be established requires the use of the static content as a measure of the deviation of a sample from expectation instead of the logarithm of the likelihood ratio or χ^2. Using the static content also avoids the thorny subject of determining the degrees of freedom.

It may be worthwhile to remember that Pearson's χ^2 is an approximation to twice the log likelihood ratio L that holds for sufficiently small deviations from expectation and sufficiently large cell frequencies: "in those cases, therefore, when χ^2 is a valid measure of the departure of the sample from expectation, it is equal to $2L$; in other cases the approximation fails and L itself must be used" (Fisher, 1922, p. 358). Fisher then notes that when χ^2 approximates $2L$ closely enough to be valid, its practical value lies in the availability of a general formula for its distribution, an essential asset before the advent of computers.

4.1 Additivity Theorem

An essential feature of the static content is its additivity when a multidimensional cross classification $C_{1..Q}$ (spanned by the variables $X_1, X_2, \ldots X_L, \ldots X_Q$) is broken down into two complementary tables:

- a cross classification $C_{1..L}$ of L dimensions ($1 < L < Q$), for which the remaining dimensions are suppressed by summation, plus
- a cross classification $C_{(1;L),L+1,\ldots Q}$ obtained by removing all mutual dependences between dimensions $1 \ldots L$. The removal is effected by replacing variables $1 \ldots L$ by their direct product, i.e. by a single variable each value of which represents a combination of values of $X_1, X_2 \ldots X_L$. (this will be called "uncoupling").

Let $u_{1..Q}$ be the joint entropy of $C_{1..Q}$, and $w_{1..Q}$ its expected entropy, the entropy of the distribution expected from its margins, and $S_{1..Q}$ its static content. Similarly, u, w, and S with subscripts "$1 \ldots L$" and "$(1; L), L + 1, \ldots Q$" indicate the corresponding measures for the two other cross classifications. The entropies of the margins of the original cross classification $C_{1..Q}$ are designated by $\mathrm{H}(X_1)$, $\mathrm{H}(X_2) \ldots \mathrm{H}(X_Q)$.

The expected entropy of the cross classification $C_{1..L}$ is

Table 4.2.1: Data calculated from Fleiss (1981)
(Fleiss' table reports proportions only)

	New York		London	
Age Group	Normal	Schizophrenic	Normal	Schizophrenic
I	24	81	71	34
II	74	118	105	69
III	63	82	93	52

$$w_{1..L} = w_{1..Q} - [\mathrm{H}(X_{L+1}) + \mathrm{H}(X_{L+2}) + \cdots + \mathrm{H}(X_Q)]. \tag{6a}$$

The expected entropy of the cross classification $C_{(1;L),L+1,\ldots Q}$ is

$$w_{(1;L),L+1,\ldots Q} = u_{1..L} + [\mathrm{H}(X_{L+1}) + \mathrm{H}(X_{L+2}) + \cdots + \mathrm{H}(X_Q)]. \tag{6b}$$

Because uncoupling any number of variables does not change the joint entropy of a cross classification, one has

$$u_{(1;L),L+1,\ldots Q} = u_{1..Q}.$$

This, together with (6a), (6b), and the definition $S_k = w_k - u_k$ for any index k (simple or compound), leads to:

$$\begin{aligned} S_{1..L} + S_{(1;L),L+1,\ldots Q} &= w_{(1..L)} - u_{(1..L)} + w_{(1;L),L+1,..Q} - u_{(1;L),L+1,\ldots Q} \\ &= w_{1..Q} - [\mathrm{H}(X_{L+1}) + \mathrm{H}(X_{L+2}) + \cdots + \mathrm{H}(X_Q)] - u_{1..L} + u_{1..L} \\ &\quad + [\mathrm{H}(X_{L+1}) + \mathrm{H}(X_{L+2}) + \cdots + \mathrm{H}(X_Q)] - u_{1..Q} \\ &= w_{1..Q} - u_{1..Q} \\ &= S_{1..Q}, \end{aligned}$$

which proves the decomposition of the static content of $C_{1..Q}$ into the sum of the static contents of $C_{1,2,\ldots L}$ and of $C_{(1;L),L+1,\ldots Q}$, q.e.d.

McGill (1954) and Fano (1961, 57–58) suggested a measure of mutual information defined in a space with an arbitrary number of variables, which can be expanded into a sum of similar mutual informations defined in subspaces of that space. It can, however, be either positive or negative and is difficult to interpret.

4.2 Additivity Theorem: Numerical Illustration

The additivity theorem will be illustrated by a three-way classification shown in Table 4.2.1, which presents the number of patients diagnosed as schizophrenic by resident hospital psychiatrists in New York and London, broken down by age.

The χ^2 of this 2 × 2 × 3 table amounts to $\chi^2 = 75.98$, its static content is $S = 0.0453$.

It seems desirable to compare the static contents which accrue in two complementary situations:

- When considering only the dependence between the Diagnosis and the Town where it originated, without regard to the patients' *Age*. The corresponding table is obtained

Table 4.2.2: Age summed out

	New York	London
Normal	161	269
Schizophrenic	281	155

by summing out the Age, which results in the 2 × 2 Table 4.2.2. For this table one has $\chi^2_{\text{Town,Diagnosis}} = 63.19$, the corresponding $S_{\text{Town,Diagnosis}} = 0.0369$.

- When neglecting the dependence between the Diagnosis and the Town where it originated. To this end one must uncouple the variables Town and Diagnosis, which yields the 3 × 4 Table 4.2.3, (superficially identical to Table 4.2.1 which was 3 rows by 2 × 2 columns, whereas Table 4.2.3 is restructured as 3 rows by 4 columns) showing the dependence between Age and the two other variables when these are made independent of each other. Here, $\chi^2_{\text{(Town;Diagnosis),Age}}$ is reduced to 13.63, $S_{\text{(Town;Diagnosis),Age}}$ to 0.0084.

The listed values show that for a mere dependence between Town and Diagnosis the static content is about 4.4 times larger than for a dependence between Age and (Diagnosis + Town), neglecting any dependence between the latter two. This provides essential information about the relative importance of the different factors involved. Adding both calculated values of S gives the static content of the original three-dimensional table, as required by the additivity theorem:

$$S_{\text{Town,Diagnosis}} + S_{\text{(Town;Diagnosis),Age}} = S_{\text{Town,Diagnosis,Age}} = 0.0369 + 0.0084 = 0.0453.$$

The additivity guarantees that these values of S can be meaningfully compared, which shows that by disregarding the interaction between Town and Diagnosis one discards more than 80% of the static content of the survey. Conversely, an exclusive consideration of this interaction alone results in a loss of less than 20% of the original content. This comparison provides a vivid picture of the actual dependence.

The corresponding values of χ^2 have been added for completeness, and they show that in this well-behaved case χ^2 deviates only slightly from additivity:

$$\chi^2_{\text{Town,Diagnosis}} + \chi^2_{\text{(Town;Diagnosis),Age}} \approx \chi^2_{\text{Town,Diagnosis,Age}}, \text{ i.e., } 63.19 + 13.63 = 76.82 \approx 75.98.$$

This is not always true, however, and large deviations from additivity can occur.

Table 4.2.3: Town and Diagnosis uncoupled

Age Group	New York Normal	New York Schizophrenic	London Normal	London Schizophrenic
I	24	81	71	34
II	74	118	105	69
III	63	82	93	52

5. Efficiency, or Terseness, of Knowledge in a Cross Classification

Passing from an expected distribution to the observed cross classification reduces the entropy of the considered ensemble from w to u and yields the information $S = w - u$. The entropy u left over reflects the uncertainty that remains because the process specifies an entire distribution, without pointing at any cell in particular. The remnant is equal to the average amount of information needed for addressing a cell. Now, one often wishes to use the linkages defined by non-empty cells to transfer information from one or several variables which are viewed as independent to a variable which is viewed as dependent. For each such transfer one must access the appropriate cell, which requires an input of information at least equal to the joint entropy u. The economy of such a transfer is a direct measure of usefulness of the knowledge embodied in the cross classification.

Remarkably, neither the total input u nor the total output—which will be called the variety "v" of the cross classification—depends on which variable is chosen as input, provided the input is entered in a way to generate the maximum possible output of information, as shown below. Therefore the ratio v/u of output to input is a parameter of the cross classification, and indicates quantitatively how efficiently it can be used for the retrieval of a single variable value. This ratio shall be called the *terseness* of the cross classification and designated by the Greek letter ζ (Zeta). It is the efficiency of the cross classification as a repository of knowledge, namely the information that can be extracted from it (e.g., as an inference) relative to that which must be entered for the purpose (e.g., as a premise).

For the sake of simplicity, u and v will be determined for a cross classification with just four variables X, Y, Z, and T. The symmetry of the resulting expressions makes it easy to extend the formulation to an arbitrary number of variables. The minimal necessary input for the selection of a cell is identical to the joint entropy u of the cross classification, as shown in the following identification sequence, at each stage of which all residual entropy is removed from the current coordinate. Its steps are:

- An unconditional determination of X, leaving no residual uncertainty about X.
- The choice of Y conditional on the value of X already chosen, resulting in no residual uncertainty with regard to Y or X.
- The choice of Z conditional on the chosen values of X and Y, leaving no residual uncertainty with regard to Z, Y, or X.
- The choice of T conditional on the chosen values of all other variables. This final step leaves no residual uncertainty at all.

Summing up the entropy reductions at all four steps shows that the total information I necessary for the identification of a cell is equal to the joint entropy u (or $\mathrm{H}(XYZT)$) of the cross classification, as expected:

$$\mathrm{I} = \mathrm{H}(X) + \mathrm{H}(Y|X) + \mathrm{H}(Z|XY) + \mathrm{H}(T|XYZ) = u. \quad (7)$$

It remains to determine the amount v of information recuperated in the course of this sequence. To identify a cell, the variables X, Y, Z, and T must be assigned definite values. Given the constraints imposed by the cross classification, these assignments are generally not mutually independent. Suppose that X is determined first. This choice reduces the entropy by $\Delta\mathrm{H}_X$, where

$$\underline{\Delta\mathrm{H}_X} = \mathrm{H}(YZT) - \mathrm{H}(YZT|X) = \mathrm{I}(YZT; X). \quad (8a)$$

The identity on the right side of (8a) is but a particular form of a standard identity of communication theory, where $\mathrm{I}(YZT; X)$ is called "mutual information" between X and the three-dimensional variable YZT, or "information transmitted" between X and YZT.

Being an average, this amount does not depend on the actual value of X. Due to the symmetric nature of $\mathrm{I}(YZT; X)$, equation (8a) can be written in the following form which is more convenient here:

$$\Delta \mathrm{H}_X = \mathrm{I}(X; YZT) = \mathrm{H}(X) - \mathrm{H}(X|YZT). \tag{8b}$$

Thereafter, when a particular value of Y is given, the entropy reduction of Z and T can be expressed as

$$\Delta \mathrm{H}_{XY} = \mathrm{H}(ZT|X) - \mathrm{H}(ZT|XY) = \mathrm{I}(ZT; Y|X), \tag{9a}$$

or

$$\Delta \mathrm{H}_{XY} = \mathrm{I}(Y; ZT|X) = \mathrm{H}(Y|X) - \mathrm{H}(Y|ZTX). \tag{9b}$$

Finally, when a particular value for Z is chosen among those still allowed for it, this selection reduces the entropy of the last remaining variable T by the amount

$$\Delta \mathrm{H}_{XYZ} = \mathrm{H}(T|XY) - \mathrm{H}(T|XYZ) = \mathrm{I}(T; Z|XY), \tag{10a}$$

or

$$\Delta \mathrm{H}_{XYZ} = \mathrm{I}(Z; T|XY) = \mathrm{H}(Z|XY) - \mathrm{H}(Z|TXY). \tag{10b}$$

The addition of (8b), (9b) and (10b), yields the sum v of all entropy reductions that the cross classification imposes on the output, given the required input u:

$$\begin{aligned} v &= \Delta \mathrm{H}_X + \Delta \mathrm{H}_{XY} + \Delta \mathrm{H}_{XYZ} \\ &= \mathrm{H}(X) + \mathrm{H}(Y|X) + \mathrm{H}(Z|XY) - \mathrm{H}(X|YZT) \\ &\quad - \mathrm{H}(Y|ZTX) - \mathrm{H}(Z|TXY). \end{aligned} \tag{11a}$$

Equation (11a) shows the restrictions that the cross classification imposes on the choice of the coordinates, when the input values are entered in the chosen order (to wit: X, Y, Z). One can easily show, however, that v is independent of this order. Adding and subtracting the additional term $\mathrm{H}(T|XYZ)$ on the right-hand side of (11a), yields

$$v = u - \mathrm{H}(X|YZT) - \mathrm{H}(Y|ZTX) - \mathrm{H}(Z|TXY) - \mathrm{H}(T|XYZ), \tag{11b}$$

where

$$\begin{aligned} u &= \mathrm{H}(X) + \mathrm{H}(Y|X) + \mathrm{H}(Z|XY) + \mathrm{H}(T|XYZ) \\ &= \frac{1}{N} \sum \varphi(a_{xyzt}) + \ln N \end{aligned} \tag{12}$$

is the joint entropy $\mathrm{H}(X, Y, Z, T)$ of the cross classification, and symmetric in all variables. As the remaining four terms on the right-hand side of (11b) taken together are symmetric too, v is symmetric in all variables, and therefore independent of the chosen sequence.

Extending (11b) and (12) to D dimensions $X_1, X_2, X_3, \ldots, X_D$ yields

$$v = \mathrm{H}(X_1, X_2, X_3, \ldots, X_D) - \sum_{i=1}^{D} \mathrm{H}(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_D) \tag{13}$$

with

$$u = \mathrm{H}(X_1, X_2, X_3, \ldots, X_D). \tag{14}$$

Thus, the terseness of a cross classification with an arbitrary number D of entries becomes

$$\zeta = \frac{v}{u} = 1 - \frac{\sum \mathrm{H}(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_D)}{\mathrm{H}(X_1, X_2, X_3, \ldots, X_D)}. \tag{15}$$

One easily sees that

$$0 \leq \sum_{i=1}^{D} \mathrm{H}(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_D) \leq \mathrm{H}(X_1, X_2, X_3, \ldots, X_D), \tag{15a}$$

with equality on the left if and only if any $D - 1$ variables uniquely determine the remaining one, and with equality on the right if and only if all variables are statistically independent, from which it follows that $0 \leq \zeta \leq 1$. Since the terseness is a quotient of two logarithms, it provides an absolute measure, independent of the logarithms' base. Rajski (1963) suggested the quotient of transinformation and joint entropy as a replacement for the correlation coefficient in a two-way table; this equals Zeta in two dimensions, but differs in more than two dimensions, where it may even become negative (Rajski, 1961).

Thus a cross classification yields two distinct pieces of information, namely:

- the static content S determined by (5b), and
- the variety $v = \zeta \cdot u$ determined by (13).

Adding S and v, and indicating by $\Sigma\, \mathrm{H}_{\text{average}}$ the sum of the average entropies of all rows, columns, etcetera in a multidimensional cross classification (each average being conditional on all variables save one, as in (11b)), one gets the entropy reduction:

$$K = S + \zeta \cdot u = (w - u) + \left(u - \sum \mathrm{H}_{\text{average}}\right) = w - \sum \mathrm{H}_{\text{average}}. \tag{16}$$

This is the total amount of information gained from a cross classification in the course of a transition from total ignorance to the unambiguous identification of one particular event. Overall, this transition delivers the amount of information equal to w, less the useless entropy reductions due to a distinction between events that are mutually "aligned" along rows, columns, lines, etc. of the cross classification. Events which are mutually aligned differ only by a single feature, and the selection of one such event cannot yield any other output than the very information that was entered in order to select it; abandoning this distinction without a difference entails a loss equal to the sum subtracted from w on the right-hand side of (16). The static content S plus the product ζu of terseness and joint entropy represent the total K of all information that can be obtained from a cross classification, given an appropriate input. K will be called the overall knowledge content of a cross classification.

In two dimensions both parentheses in equation (16) are numerically equal. In other words, the amount of information S obtained when one is informed of the content of a cross classification is equal to the largest that can be obtained from it thereafter through an appropriate input. However, this equality holds only in two dimensions, and created some confusion in the past. It led to the definition of a symmetric uncertainty coefficient $2S/w$ occasionally cited in the literature (Brown, 1975; Press et al., 1989) and implemented in statistics packages (SAS, SPSS, SYSTAT).[1] This coefficient is unsatisfactory because it lacks a straightforward interpretation. Also, it does not readily extend to more than two

[1] For information regarding SAS Version 6 contact SAS Institute, Inc., SAS Campus Drive, Cary NC, 27513-2414 (Phone: 1-919-677-8000. Web site: www.sas.com). For information regarding SPSS/PC + Advanced Statistics or SYSTAT contact SPSS, Inc., 444 North Michigan Avenue, Chicago, IL 60611 (Phone: 1-800-543-2185. Web site: www.spss.com).

variables and must hence be normalized artificially by including the number of variables in its definition; this cannot be achieved, however, without introducing inconsistency when the frequencies in some cells tend to zero.

6. Separability of a Variable

When analyzing a cross classification, the primary interest often lies neither in S nor in ζ, but in the amount of information that the cross classification can usefully transfer from one variable, or from one set of variables, to another. Thus, it may be desirable to determine, for example, how much information about the future health state of a person can be inferred from his or her observable present vaccination status, using the statistical knowledge contained in a table of past results of this type of vaccination.

The term "dependence" is traditionally used for the strength of such inferences (also called "associations") and will be retained here. It should be understood that it need not imply any actual relation from cause to effect (Jaynes, 1989), although the well-entrenched term "dependence" will often be applied here, even if common usage suggests a causal interpretation.

As a rule one is not so much interested in the absolute amount of transferred information than in its relative extent compared to what might have been obtained through a direct observation of the variable of interest (see Salk vaccine example, Table 7.4.1). Given a cross classification, a fraction of the information that would be gained by directly observing the independent variable can also be gathered indirectly through an observation of the dependent variable. This fraction provides a readily interpretable measure of the extent to which the dependent variable determines the independent one. It is a dimensionless, normalized measure of the efficiency of the directed inference from a variable, or a set of variables, to the complementary one. Because it evaluates the ease with which values of a variable can be distinguished through an exclusive observation of the complementary variable, it will be called the *separability* γ_X of X, when X is viewed as the independent variable.

The efficiency of an indirect observation will now be derived in more detail. As this involves only two variables, independent and dependent (observed), the following considerations can be restricted to a two dimensional cross classification spanned by these two (possibly multidimensional) variables. One may further suppose that the independent variable represents entities, whilst the actually observed variable identifies features attached to the same. The ratio of the indirect information to the direct one then measures how efficiently entities can be distinguished from each other through the mere observation of their features.

Let the frequency of a simultaneous occurrence of x_i and y_j be a_{ij}, and call the sum of all frequencies N, that is, $N = \Sigma a_{ij}$. If an event x_i occurs and is observed directly, no residual uncertainty is left over after the event; hence the generated information equals the original entropy of X, that is, the entropy of the marginal row:

$$\mathrm{H}(X) = \frac{1}{N} \sum \varphi(a_{i.}) + \ln N. \tag{17}$$

Suppose now that variable X cannot be observed directly, but only through an observation of variable Y which has the value y_j. As a result of this observation, the distribution of X shrinks from that of the marginal row to that of row j. The former has the entropy $\mathrm{H}(X)$, and the latter the smaller entropy $\mathrm{H}(X|y_j)$. It follows that, on average, each observation of Y reduces the original entropy $\mathrm{H}(X)$ to a posterior value $\mathrm{H}(X|Y)$ which is the average entropy of X given Y. For the calculation of this average, the entropy of each row must be weighted according to its frequency $a_{.j}$. Subtracting the average residual entropy of X after an observation of Y from the entropy of X before this observation yields the information $\mathrm{I}(X; Y)$ about X returned—on average—by an observation of Y:

$$\mathrm{I} = \mathrm{H}(X) - \mathrm{H}(X|Y). \tag{18}$$

This is the "transinformation" of communication theory. If one divides it by the amount of information $\mathrm{H}(X)$ that could in principle have been gathered through a direct observation of X, one obtains the fraction of the total information about X obtainable through an observation of Y, which will be called the separability γ_X of X given Y:

$$\gamma_X = \frac{\mathrm{H}(X) - \mathrm{H}(X|Y)}{\mathrm{H}(X)}. \tag{19a}$$

The separability γ_X is normalized to one and remains constant when all frequencies are multiplied by the same factor. The separability γ_X applies to an arbitrarily large set of distributions Y_i, one distribution per each particular value of X, and measures the ease of distinguishing (or separating) the distributions through the mere observation of Y. In general the separability thus determines the heterogeneity of a set of distributions. If X is a dichotomous variable, a difference between two distributions is measured as in the Kolmogorov-Smirnov test or with Rajski's (1961) $d_{X,Y}$. The latter evaluates a distance between the margins of a two-entry table and satisfies the metric axioms. Formally, it is closely related to the terseness in two dimensions and to the separability, through the equations $d_{X,Y} = 1 - \zeta$ and $(2 - d_{X,Y})/(1 - d_{X,Y}) = 1/\gamma_X + 1/\gamma_Y$, but its rationale is different. In particular, the distance $d_{X,Y}$ is limited to two distributions and does not evaluate their dissimilarity because it remains constant under all permutations of the frequencies in either distribution.

In analogy to definition (19a), the separability γ_Y of Y given X equals the fraction of the total information about Y obtainable through an observation of X:

$$\gamma_Y = \frac{\mathrm{H}(Y) - \mathrm{H}(Y|X)}{\mathrm{H}(Y)} \tag{19b}$$

The separability ranges from zero to one, inclusive, and is independent of the base chosen for the logarithms used when calculating the entropies. Note that the range from 0 to 1 of both terseness and separability should not be used as in the interpretation of correlation coefficients. Whereas a correlation of 0.10 may express less than remarkable association, a terseness or a separability of that magnitude should be considered striking, even 0.01 is nothing to be discarded easily. Values of γ or ζ above 0.10 indicate an exceptionally strong association.

The separability has been repeatedly suggested—under various names like "asymmetric uncertainty coefficient"—as a measure of association (e.g. Press et al, 1989), often on the ground of its belonging to a class of measures which evaluate "the relative reduction in uncertainty about Y from getting to know X" (Särndal 1974). Actually, it is the only measure which fully satisfies this definition when one interprets uncertainty as lack of information. Further grounds for its importance lie in its close relationship with the information S obtained "from getting to know the cross classification" (to paraphrase the above definition) and also with the terseness ζ.

6.1 Model for Inductive Inference based on the Separability

Consider a two-column table where the column widths represent time intervals. The table contains events generated by a hypothetical parent distribution of the marginal

Table 6.1.1: Minimum necessary change for some artificial data

			Columns Q_N								
	H	D	N=25	=50	=75	=88	=100	=125	=150	=175	=200
Y	0.64	75	5	21	37	45.3	53	69	85	101	117
Yy	0.32	23	17	25	33	37.2	41	49	57	65	73
y	0.04	2	3	4	5	5.5	6	7	8	9	10
$J = \gamma(D,Q_N)\cdot(N_Q/N_D)$			.05279	.04188	.03983	.03970	.03976	.04049	.04160	.04292	.04438

column. Each row represents a possible outcome. If the boundary between the two columns corresponds to the present time, the table can be viewed as a yet incomplete cross classification that compares the left column of N_D observed data with hypothetical future data that wait to be collected in the right column. Because empirical observations may never be disavowed, any conceivable future data set that one wishes to compare with the hypothesis must include the observations already made. The essential consequence of a divergence between the assumed parent distribution and the data already sampled can be verified empirically and is to be found in the occurrence of future outcomes not consistent with the most probable distribution to date (that of the sample in the left column), but needed to compensate the divergence between the observed data and the assumed hypothesis. This set Q, composed of N_Q virtual events, is not uniquely determined: there may be either few events strongly separated from the observed set, or many events weakly separated from it. By definition, past and future sets stem from the same parent universe, hence their separability γ_{DQ} should be small if the hypothesis is to be plausible. Because the ratio N_Q/N_D evaluates the relative weight of the unwarranted virtual events, the product $J = (N_Q/N_D) \cdot \gamma_{DQ}$ of the separability with this ratio is a measure of the distortion introduced in order to turn the observed set into part and parcel of a most probable set under the hypothesis. The minimum of J will be called the *twist* and designated by τ. It is a lower bound for the arbitrariness required to bring the full set of events into line with the hypothesis and evaluates the smallest departure from the distribution of the sample that must be imposed on future events to satisfy the hypothesis. The twist is a natural measure, and hopefully a valid quantitative estimate, for the distortion which the assumption of the hypothesis imposes on the future events, given the data.

These considerations suggest the following procedure, illustrated by the artificial data of Table 6.1.1, where column D represents the observed set of data, column H the proportions expected under a hypothesis (a binomial distribution with $p = 0.8$ in this example), and the columns Q_N represent sets of virtual observations that are so calculated that each column Q_N combined with column D forms a two-column table with the marginal proportions listed as in column H. (The lengthy computation is eased by a little program for the PC, available from the second author upon request). Now, N_Q can be made to vary continuously, and each value N_Q will then determine a set Q_N of virtual frequencies having a well-defined separability with respect to the observed set. One can then view J as a function of N_Q alone, for a given hypothesis and an observed set; its minimum is the twist. The virtual weights N_Q can vary from the lowest possible value (i.e., one not requiring negative frequencies) to one beyond which no further minimum of J exists, and can include fractional frequencies for continuity.

In Table 6.1.1 the twist amounts to 0.03967 and occurs for the set Q_{88} = (45.3, 37.2, 5.5). Because the number of virtual events differs only moderately from the sample size, the denominator H(100, 88) of $\gamma(D, Q_{88})$ can be approximated by ln(2), so that the relation due to Fisher (1922, p. 357–58), $\chi^2 \approx 2Nv$, leads to the following approximation for the twist

$$\tau \equiv \left(\frac{N_Q}{N_D}\right)\gamma(D, Q_{88}) \approx \left(\frac{N_Q}{N_D}\right)\frac{\chi^2_{D,Q}}{N}\cdot(2 \ln 2) = 0.03916.$$

This shows that, when the size of the virtual complement does not differ too much from that of the sample, the twist tends to be proportional to χ^2/N. The twist, however, addresses the question, "How strongly must one distort the distribution of the data to satisfy the hypothesis?"; while a significance test addresses a different question, namely, "Might the sample have been generated by the hypothesis, given some reasonable variance?". (Note that the answer depends on the sample size.) The twist provides a measure of the disagreement between hypothesis and sampled data which, unlike χ^2, depends only on the distribution in the sample, not on the number of observed events. This distinction heeds Fisher's (1951, p. 195) dictum, "The tests appropriate for discriminating among a group of hypothetical populations having different variance are thus quite distinct from those appropriate to a discrimination among distributions having different means". Using the twist for a comparison between several samples eliminates the difficulties due to an exclusive use of significance tests because the latter evaluate the variance that must be forced upon a hypothesis which is assumed to be absolutely exact, and which will therefore become less and less likely when a sample gets larger without change in its distribution (except in the trivial case of complete concordance between sample and hypothesis). This not only precludes meaningful comparisons between samples of different sizes (see the example in section 6.2 which illustrates the point), it is also a hindrance when the hypothesis to be verified is not derived from some theory (as sometime available in genetics), but empirically (as is the rule in the social sciences): even if it is a very good approximation, as far as empirical formulas go, the unavoidable deviation from its (unknown) exact form will be endlessly magnified when the number of observations becomes large. Thus, due to the unavoidable asymmetry of a real die, throwing it sufficiently often will always generate an arbitrarily large significance to reject the null hypothesis of equal probability of the six sides; this, however, reveals nothing about the degree of the die's asymmetry.

The use of the twist for inductive inference can be illustrated by the data from which Bernstein (1924) concluded that the ABO blood-groups in man are determined by three alleles (A, B, and O) at a single locus (Hypothesis 1) rather than by two alleles at each of two loci (A, a; B, b) as formerly believed (Hypothesis 2). The observed frequencies and the proportions expected on each hypothesis are shown in Table 6.1.2. As the source states proportions only, the listed frequencies were reconstructed to the nearest integer from these proportions and the known sample size.

The twist necessary to compensate for the departure from the hypothesis is not only much larger for the two loci hypothesis than for the one locus hypothesis, the twist for one locus is also very small in absolute terms. The evidence in favor of Bernstein's judgment is thus overwhelming.

6.2 *Lindley's Paradox, an Example*

Let it be known that a certain race of rabbits has different fur patterns a, b, c, d in the proportion 4:3:2:1. One would like to know if the rabbits in A-valley belong to that race or not.

Table 6.1.2: The twist for AB0 phenotypes and one- vs two-loci hypotheses

Blood group phenotype	Observed frequencies	Proportions assuming	
		one locus	two loci
A	212	0.4112	0.358
B	103	0.1943	0.142
AB	39	0.0911	0.142
0	148	0.3034	0.358
τ, required twist:		0.00226	0.05922

A first survey of 100 rabbits resulted in the frequencies 46, 28, 18, 8 of the fur patterns (column A of Table 6.2.1). Then, without authorization, gamekeeper Fred Busybody surveyed another 100 rabbits, with frequencies 43, 29, 19, 9 (column B), producing counts 89, 57, 37, 17 for the enlarged sample of 200 rabbits (column C). As the proportions moved closer to the theoretical expectation for all fur patterns, the distribution has come closer to the expectation.

Does the enlarged survey (column C) give more support to the hypothesis than the smaller first survey (column A), in spite of its larger significance for rejection? Suppose further that survey B is followed by an endless sequence of surveys $B_1, B_2, \ldots, B_k, \ldots$, each of which is twice the size of the preceding one and has pattern frequencies midway between the sum of those observed to date and the hypothetical ones (allowing fractional frequencies for convenience, e.g., $B_1 = \{84.5, 58.5, 38.5, 18.5\}$, $B_2 = \{166.75, 117.75, 77.75,$

Table 6.2.1: Two surveys and their combination

Fur pattern	Hypothesis	A First survey		B Busybody's additional survey		C First survey plus Busybody's data	
		N	p	N	p	N	p
a	0.40	46	0.46	43	0.43	89	0.445
b	0.30	28	0.28	29	0.29	57	0.285
c	0.20	18	0.18	19	0.19	37	0.185
d	0.10	8	0.08	9	0.09	17	0.085
N		100		100		200	
χ^2		1.633		0.408		1.838	
Twist τ		0.011828		0.002948		0.006642	

37.75}). Clearly, the observed overall distribution converges asymptotically toward the hypothesis, and one can easily show that χ^2 simultaneously tends to infinity with 1.125^k.

This is a simple numerical instance of Lindley's (1957) paradox, who concluded that (a significance level of) "5% in today's small sample does not mean the same as 5% in tomorrow's large one."

Survey C comes closer to the hypothesis than A, as documented by its smaller twist. In view of the history this should not be surprising: Busybody's additional sample (column B) is certainly closer by any criterion, so its addition can only move the combined sample C in the direction of the hypothesis. Yet χ^2 is increased for sample C, the twist behaves as it should. At the same time A, due to the smaller sample size, has a larger margin of error than C, and so has its twist. Thus the distribution underlying A could occasionally be closer to the hypothesis than the distribution underlying C, although this is hardly probable.

Consequently, to answer the question as to whether a distribution fits a hypothesis, we require two parameters:

- The magnitude of the disagreement between observation and hypothesis.
- The error connected with the estimate of that magnitude, or the variance about the estimated magnitude.

The disagreement is measured by the twist τ, but we can not yet give a formula for the twist's standard error. Brown (1975) published a formula for the asymptotic standard error of γ, based on methods described by Goodman and Kruskal (1972). This calculation is not directly applicable to the twist, however, and we felt we must postpone the problem of deriving its error estimate.

7. Exemplary Applications

7.1 Simpson's Paradox

This well-known paradox (Blyth, 1972a, 1972b; Lindley & Novick, 1981; Novick, 1980; Székely, 1990) is that an effect exists in both of two groups and is reversed when both groups are combined. Consider an artificial example in which the starting of a car depends on the grade of gasoline and the time of starting (see Table 7.11). High octane gas facilitates starting early in the morning as well as later in the morning, yet combining the two tables seems to indicate that low octane gas lets the car start more easily. Surely this contradiction is irritating and needs an explanation.

It is clear that cars have greater difficulty to start in the morning cold, but another aspect of the data is also clear, namely that early starters tend to use high octane gas. Thus, the choice of high octane gasoline appears inferior for starting the car, in striking contrast to the superiority of high octane gas within each time period, via the tendency of high octane gas users to start under the unfavorable condition of morning cold.

The statistician who is asked for advice might recommend high octane gas if he knows the time that a driver usually leaves home and that this time will not change; if the time of starting the car were undecided, he might recommend low octane gas, and his advice would not be as illogical as it seems at first sight. Indeed, if the strong relation between type of gas and starting time remains in force, a driver following the advice for high octane gas will more likely start in early morning, satisfying the empirical correlation between gas quality and preferred time to start the car, and as a consequence he will increase the overall frequency of unsuccessful starts. This indirect influence overpowers the superiority of high octane gas observed severally in the early morning and later. Thus, unless early starters remain early starters and late starters remain late starters—irrespective of the type of gas they use—one must indeed recommend low octane gas in order to prevent an increase of starting failures under the unfavorable conditions of early morning cold.

Table 7.1.1: Fictitious data, adapted from Novick (1980)

	Early in the morning		Later in the morning		Early and Later Combined	
Motor started	Low Octane	High Octane	Low Octane	High Octane	Low Octane	High Octane
Yes	2	9	18	7	20	16
No	8	21	12	3	20	24
Prop. Yes	0.20	0.30	0.60	0.70	0.50	0.40

This explains the paradox, but a simple quantitative criterion to signal the danger of the paradox occurring would be helpful. If ζ decreases when a variable is eliminated from a table, this is a warning that the usefulness of the table deteriorates through the loss of a dependence essential for the interpretation of the full table. As shown in Table 7.1.2, $\zeta = 0.1056$ for the full Table 7.1.1 with all three variables. It drops to 0.0037 (by a factor of 25) when one ignores *Time* and merely looks at the cross tabulation of *Gasoline* × *Success*, meaning that Time is of such paramount importance for the data set that its exclusion almost amounts to an "illegal" act. The only "legal" suppression would be that of the variable of primary interest, namely Success, as ζ for Time × Gasoline equals 0.1042, almost as high as for the complete three-dimensional table. Thus the main content of Table 7.1.1 is the strong association between Gasoline and Time, and ignoring this aspect of the data is punished by the baffling paradox. However, the correlation between the type of gas used and the time of starting the car—even if it is strong—is probably not a genuine dependence, as the driver who is advised to use high octane gas will scarcely change his habit of starting early or late when following the advice. Hence this advice can rest on the separate data sets for early and late starters, but a different context may lead to a different conclusion.

7.2 "The American Soldier" Revisited

A venerable data set first presented by Stouffer (1949) and used by Goodman (1978) to demonstrate the logit multiple regression approach to the analysis of dichotomous variables, and by Theil (1972) to examine the relation between logit and entropy, will be re-analyzed using the new approach.

Table 7.1.2: Terseness of all possible subtables

Variables in the cross tabulation	Terseness
Time, Gasoline, Success	0.1056
Time and Gasoline	0.1042
Time and Success	0.0480
Gasoline and Success	0.0037

Table 7.2.1: Preferred location by *race*, *origin* and *present camp*

	Black Race				White Race			
	Northern Origin		Southern Origin		Northern Origin		Southern Origin	
	Present Camp		Present Camp		Present Camp		Present Camp	
	North	South	North	South	North	South	North	South
Prefer North	387	876	383	381	955	874	104	91
Prefer South	36	250	270	1712	162	510	176	869
Prop. North	0.915	0.778	0.587	0.182	0.855	0.632	0.371	0.095

The data in Table 7.2.1 served to study the preferred location of a training camp (Northern or Southern US state) for American soldiers of Black or White racial origin, Northern or Southern origin, stationed in a Northern or Southern camp.

An analysis of these data will raise several questions:

Q1: Which factors are least informative, and may be discarded with tolerable loss of information?

A1: The least informative variables are those the removal of which causes the smallest loss of static content. They can be found by removing variables one after the other such that each removal causes the smallest loss of content achievable at that stage. Table 7.2.2 (left half) lists S and ζ obtained when removing one variable from the original cross classification. *Present camp* turns out to be most disposable. Among the three remaining variables *Race* is the most unnecessary, its removal reduces the static content least, as can be seen in Table 7.2.2 (right half). After the removal of Present camp and Race, only two variables are left, so no further step is possible.

Q2: Can one make the cross classification more efficient by neglecting less informative variables?

A2: The preceding tables reveal that the initial elimination of Present Camp marginally increased the terseness of the cross classification, and that the following

Table 7.2.2

	For selecting first variable to be removed		After removing *Present Camp*	
Removed Variable	Static Content of Remainder	Terseness of Remainder	Static Content of Remainder	Terseness of Remainder
None	0.2620	0.1078	0.1986	0.1081
Present Camp	0.1986	0.1081		
Race	0.1967	0.1031	0.1437	0.1157
Preference	0.0684	0.0349	0.0480	0.0359
Origin	0.0602	0.0301	0.0024	0.0017

Table 7.2.3

	Origin		Present Camp		Race	
	South	North	South	North	Black	White
Prefer North	959	3092	1829	2222	2027	2024
Prefer South	3027	958	644	3341	2268	1717
Sep. of Pref.	0.2074		0.0734		0.0034	
Terseness	0.1157		0.0404		0.0017	

elimination of Race increased the terseness noticeably. The process then ends for lack of variables, but in the general case it should be pursued until the terseness attains a maximum.

Q3: Which single factor influences the Preference most strongly?

A3: This question is closely related to the preceding one, the difference being that Question 2 relates to the efficiency of the knowledge as a whole, while the present question specifically concerns the dependence of the single variable *Preference* on each other variable severally. Table 7.2.3 shows the separability of Preference and the terseness for all three two-dimensional cross classifications spanned by Preference and one other variable. The greatest separability of Preference, by a large margin, occurs for the 2 × 2 table linking *Preference* with *Origin*, which is also the tersest one.

Q4: Which variable is most dependent on the ensemble of all others, disregarding all mutual dependences between the latter?

A4: This differs from Question 3 in that it relates to the dependence of each separate variable on all others, rather than on only one. To answer it, one must successively uncouple all possible combinations of three variables in order to generate all two-entry tables which connect a single variable with the conglomerate of all others. As an example, Table 7.2.4 shows the cross classification obtained through the uncoupling of Preference, Origin, and Present Camp. According to the addition theorem, the static content of this cross classification (0.0653), plus that of the cross classification obtained when removing Race by summation (0.1967), equals the static content of the full original data (0.2620); this can be easily checked (see Table 7.2.2).

Table 7.2.5 lists each variable's separability, given the ensemble of all others, as calculated for Table 7.2.4 and three analogous ones, in each of which a different set of three variables has been uncoupled. Note that Race is least

Table 7.2.4

Pref/Orig/Camp :	NNN	NNS	NSN	NSS	SNN	SNS	SSN	SSS
Black Race	876	387	381	383	250	36	1712	270
White Race	874	955	91	104	510	162	869	176

Table 7.2.5

Variable	Separability
Origin	0.2912
Preference	0.2793
Present Camp	0.1028
Race	0.0945

dependent on all others, although it is Present Camp that contributes least to the overall static content, as was seen in Table 7.2.2.

The above results can be summarized as follows:

1. The least informative variable is Present Camp, closely followed by Race.
2. The tersest representation of knowledge obtainable from the original data is the 2 × 2 table that connects Preference with Origin and neglects the two other variables.
3. Preference depends by far most strongly on the single variable Origin.
4. Origin is the variable most efficiently inferred from all others, closely followed by Preference. Race comes last in this respect.

The analysis provides a compact quantitative view of Stouffer's data by using parameters which allow meaningful comparisons between cross classifications of varying dimensionality. It strongly supports one conclusion: "Soldiers want to live near home", and a weak corollary: "The location of his present camp is a much more important consideration for a soldier's preference between North and South than his race" (see Table 7.2.3).

7.3 Finley's Tornado Prediction

A vexing problem, which to the best of the authors' knowledge has not been solved satisfactorily, was presented by Goodman and Kruskal (1959, p. 127–128): Finley, sergeant in the US Signal Corps, compared his predictions about tornados' occurrence with the actual occurrence. One of Finley's summary tables is given here as an example (Table 7.3.0). As he predicted correctly in 917 (11 + 906) of 934 cases, he gave himself a percentage score of 98.18%. Goodman and Kruskal point out that a completely ignorant person could always predict "No Tornado" and attain a score of 98.50% (920/934), and they conclude "of course, Finley did appreciably better than this 'prediction', the question is that of measuring his skill by a single number."

Consider two separate worlds, that of Finley and an ideal one which is the abode of an infallible forecaster. A sound measure of Finley's skill is the separability between both worlds, the ease with which one can distinguish them when told only that a certain event (a pair prediction/occurrence among the four possible ones: yes/yes, no/yes, yes/no, no/no)

Table 7.3.0

	Tornado	No Tornado	Total
Tornado predicted	11	14	25
No Tornado pred.	3	906	909
Totals	14	920	934

Table 7.3.1: Comparison of Finley's Prediction with different worlds

X = prediction	yes	yes	no	no		
Y = observation	yes	no	yes	no		
Q = compound	y/y	y/n	n/y	n/n	Proportion correct	Separability from ideal
Infallible forecaster	14	0	0	920	1.0000	0.0000
Finley's world	11	14	3	906	0.9818	0.0093
Always No	0	0	14	920	0.9850	0.0150
Random	0.2	13.8	13.8	906.2	0.9704	0.0216
Always erring	0	14	920	0	0.0000	1.0000

took place, but not in which world this happened. It is not enough here simply to record the success or failure of a forecast: an event must be identified by a prediction/outcome pair, because the incidence of tornados is so rare that any sensible bookmaker would clearly offer better odds to Finley for the actual prediction of a tornado than for the successful prediction of no tornado occurring. This asymmetry is an essential feature of the problem, disregarding it leads to the apparent success of the no-sayer. The more difficult it is to infer the location of an event from its type (e.g., that a tornado was predicted but did not occur), the smaller the difference between Finley's results and those of the ideal forecaster. Thus the separability of the two worlds, given an event that occurred in either, is a direct measure of Finley's lack of skill.

It is instructive to consider the two extremes of a forecaster's skill and the corresponding values of the separability:

- In an ideal case only events y/y and n/n occur, and their ratio is the same as the ratio of tornados to no tornados. The infallible forecaster will also produce only events y/y and n/n in the same ratio; therefore one cannot infer from any observed pair "prediction/reality" whether it occurred in the forecaster's world or in the ideal one. The separability of the two worlds is then nought.
- Conversely, if the forecaster is always wrong, if he produces only false positive and false negative predictions that never occur in the ideal world, then one can always infer with certainty in which world an observed pair took place: all false predictions occur in the forecaster's world, and all correct predictions in the ideal world. The separability of the two worlds is then equal to one, as it must be.

Table 7.3.1 shows the frequencies of Finley's predictions together with the mentioned extreme possibilities (rows of the table); the values of the compound variable Q built from X = prediction and Y = occurrence form columns of the table. This replacement of X and Y by the compound Q is an uncoupling which effectively removes any dependance between X and Y.

To evaluate Finley's skill one must calculate the separability of the row of his predictions from the row of the infallible forecaster. The average entropy over all four values of the compound variable Q is:

$$\begin{aligned}\bar{H} = \frac{1}{1868} \times [&\varphi(11) + \varphi(14) - \varphi(25)\\ &+ \varphi(14) + \varphi(0) - \varphi(14)\\ &+ \varphi(3) + \varphi(0) - \varphi(3)\\ &+ \varphi(906) + \varphi(920) - \varphi(1826)]\\ = 0.6867.&\end{aligned}$$

Because both rows have the same sum, the entropy of the margin along the binary variable worlds is equal to ln(2), so that the separability between Finley's world and that of the infallible forecaster becomes:

$$\gamma_{worlds} = \frac{\ln(2) - \bar{H}}{\ln(2)} = 0.0093.$$

The separability of the "No"-sayer's world from the ideal world is equal to 0.0150, i.e., about 1.6 times larger than that between Finley's world and the ideal one, which is appropriate and solves the problem raised by Goodman and Kruskal.

Incidentally, the comparison of the "No"-sayer with the infallible forecaster illustrates an interesting special case, where all possible events can be divided into two mutually exclusive classes:

- In 14/934 of all cases—when a tornado struck—one can unambiguously infer whether an event (predicted and occurred, or not predicted and occurred) stems from the "No"-sayer or from the infallible forecaster.
- In the remaining 920/934 of all cases—when no tornado occurred—the nonevent (forecast or not) has the same relative frequency in both worlds, therefore its occurrence gives no clue about the identity of the forecaster.

No intermediate case exists, and one can easily show that in such a situation the separability becomes identical to the fraction of events (here: 14/934) where an unambiguous inference is possible. In this particular case Finley's naive—and modest—evaluation of his skill through a percentage of successful predictions would actually be appropriate. This agreement between a straightforward estimation and the separability in a simple situation substantiates the interpretation of γ.

For the sake of completeness, one may also consider a forecaster who makes random predictions based solely on expected frequencies, in other words, who generates a table with zero content. The separability between the world of this forecaster and the ideal one amounts to 0.0216, which is 1.5 times larger (i.e., worse) than what was obtained by the systematic no-sayer who exploited the asymmetry due to the rarity of tornados.

Note that, although in the above examples γ was used to determine the dissimilarity between only two distributions, it is not a distance (which makes only sense for two entities), but a measure of heterogeneity defined for an ensemble comprising any number of distributions.

7.4 Miscellaneous Treatment-Outcome Studies

The measures introduced will now be illustrated by an analysis of four fairly dissimilar 2 × 2 tables from medical surveys that can to a large extent be compared by inspection (Table 7.4.1). In all four cases one is mainly interested in the dependence of health status on treatment, and accessorily in the static content generated on the passage from the

Table 7.4.1: Four medical 2 by 2 tables linking outcome to treatment

	Lung Cancer Treatment (Schmitt, 1969)		Salk Vaccine against Poliomyelitis (Schmitt, 1969)		Typhoid Vaccination (Vessereau, 1947)		Cholera Inoculation (Kendall, 1943)	
	alive	died	well	ill	well	ill	well	ill
Treated	56	252	200712	33	6759	56	276	3
Control	34	212	201114	115	11396	272	473	66
γ_X	0.0039		0.0182		0.0192		0.0831	
S	0.00174		0.00006		0.00171		0.02403	
χ^2	1.9		45.3		56.2		29.7	
Pearson C	0.0586		0.0106		0.0551		0.1872	

noncommittal "expected distribution" to the actually observed one. Clearly, the dependence of X on Y increases from Cancer to Typhoid, and again from there to Cholera, while the status of Salk is difficult to assess by inspection alone, due to the large frequencies involved and to the extreme skewness of this table. So far the values of the separability γ_X shown in Table 7.4.1 place all other tables in the proper order.

For comparison, Table 7.4.1 also lists two classical statistics, of which χ^2 heavily misplaces the Cholera, while Pearson's C puts Typhoid and Cancer in the wrong order. Others, such as Cramer's V, which are not listed, do not fare much better, and often lack a reasonably direct interpretation. Quite generally, no coefficient symmetric in the variables is satisfactory as a measure of inference, because usually the relative strength of an inference from X to Y differs from that in the reverse direction. This is particularly noticeable for the very skew Salk data, where γ_Y practically vanishes.

Turning now to the poliomyelitis-Salk data, a quandary that plagues most people when confronted with this table can be expressed as follows: "Let us see: by any criterion the 3:1 reduction of the incidence of illness through the vaccination is of the essence. On the other hand, a shift that concerns only 148 people among 401'826 cannot be that important." Actually, both statements are valid, because the Salk data combines a very small knowledge content with a modest, but real, dependence of X on Y. The shift in the relative frequency of illness due to vaccination is similar for Salk and Typhoid, hence the similar values of their separabilities. But—in view of the tiny fraction of all people befallen by poliomyelitis—this shift carries only a small information value in the Salk case because an overwhelming proportion of people stay healthy anyway, hence the minute static content of the Salk data. This tallies nicely with the common sense response of any individual who rightly feels that, in a sense, the Salk table is of little impact, because whatever effect it exhibits affects so comparatively few people. This is not a subjective view due to the particular situation, but remains true if the headings 'healthy' and 'ill' were reversed (and thereby the subjective appreciation too).

The existence of two different aspects, to wit the amount of information "contained" in a table, and the relative dependence of one of its variables upon another, stresses the necessity to use two distinct parameters. The dilemma is resolved when one complements the efficiency of the directed inference—as evaluated by the separability of X—with the static content S of the tables. This content is nearly the same for Typhoid and Cancer, about 14 times larger for Cholera, and nearly 30 times smaller for Salk, which adequately

Table 7.5.1: Letter frequencies in three languages, (H.F. Gaines, 1956)

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
English	78	13	29	41	131	29	14	58	68	2	4	36	26	73	82	22	1	66	65	90	28	10	15	3	15	1
German	50	25	15	50	185	15	40	40	80	0	10	30	25	115	35	5	0	70	70	50	50	10	15	0	0	15
French	94	10	26	34	159	10	10	8	84	9	0	53	32	72	51	29	11	65	79	73	62	22	0	3	2	3

reflects the large differences in the "usable content" of the considered tables, and vindicates the subjective opinion quoted above. Thus, the static content S complements the separability of X, which provides a consistent and readily interpretable evaluation of the extent to which one can infer a patient's state of health from a knowledge of the treatment he received.

7.5 Optimal Grouping of Categories

An illustrative application of the separability consists in the comparison of letter frequencies in the three languages English, French, and German. The frequencies per 1000 letters of text are shown in Table 7.5.1. Table 7.5.2 lists the values of γ for the three possible pairs. Some variant of the χ^2-test would have been the customary choice, were it not for the small frequencies of some letters. An arbitrary grouping of letters, customary as well, might help to alleviate χ^2's problem of small frequencies at a cost, namely the loss of possibly pertinent information contained in rare letters that have a high discrimination value. The separability γ remains applicable in spite of the small frequencies, and it can help in grouping the letters on a rational basis. A smaller and more efficient table than 7.5.1 is certainly desirable, beyond the necessity stemming from the χ^2-test's inability to deal with small frequencies. But which letter groupings should one choose? Consonants versus vowels? Frequent letters versus rare ones? Among the multitude of possible groupings we propose the one that maximizes the terseness of the table while disregarding distinctions between letters which are least efficient for a separation of the languages. This can be realized by successively merging into a single group the pair of letters (or previously

Table 7.5.2
Separabilities of language pairs

	Separability γ
English - French	0.0528
English - German	0.0549
German - French	0.0721

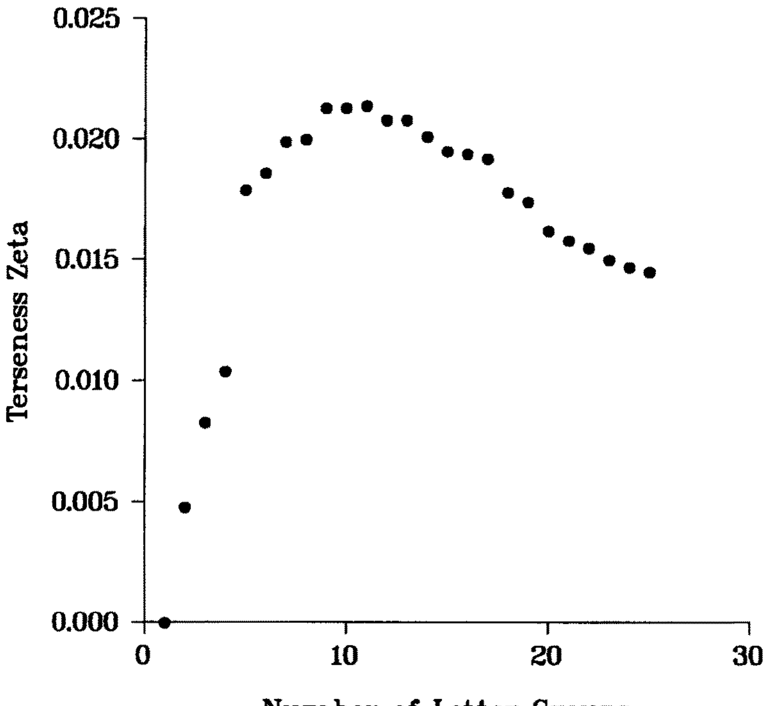

FIGURE 7.5

formed groups) which are least separable, given their frequencies in the three languages. The number of groups thus progressively shrinks from 26 to a single one, and it can be shown that the terseness reaches exactly one maximum in the course of this process, the results of which are shown in Figure 7.5. (A program "TAXIS" for the PC,[2] that automatically finds the maximum, is available).

It turns out that forming 11 groups produces the tersest table. The terseness increases from an initial value of 0.014 for all 26 letters to a maximum of 0.021 for 11 letter groups. Table 7.5.3 reveals that the majority of frequent but not very discriminating letters, including all the vowels, are grouped together, while letters that have discriminating power (because their use is typical for one language) retain their standing (*y* is typically English, *k* and *z* German, and *j* and *q* French). Note from Figure 7.5 that for a quick and dirty

[2] For information regarding TAXIS contact Software for Science, Chicago, IL 60614-3011.

Table 7.5.3: Optimal grouping of letters

	y	k	w	h	bg	aeiou cdlmn rstv	f	p	x	z	jq
English	15	4	15	58	27	823	29	22	3	1	3
German		10	15	40	65	835	15	5		15	
French	2			8	20	906	10	29	3	3	20

Table 7.6.1: Conditional survival in two hospitals

	Hospital I			Hospital II		
	X=Survival status			X=Survival status		
	Lived	Died	Total	Lived	Died	Total
Treated	0.84	0.04	0.88	0.42	0.02	0.44
Not treated	0.03	0.09	0.12	0.14	0.42	0.56
Total	0.87	0.13	1.00	0.56	0.44	1.00
Separability of x	0.404			0.422		
Overall content K	0.312			0.579		

overview the distinction of only five groups (heavy lines in 7.5.3) would also be quite sufficient.

7.6 *Choosing the Number of Control Cases*

An ever recurring problem is the choice of an appropriate number of control cases, especially for costly medical surveys. In this context, Goodman and Kruskal (1954) analyzed the artificial cross classifications shown in Table 7.6.1 which relate to the effects of a medical treatment on persons contracting an often fatal disease. Both samples are supposed to be very large (the tables list proportions only).

The conditional survival probabilities are the same for both hospitals, and this raises two questions:

Q1. Is the dependence of life expectation on treatment essentially the same in both cases, as one is wont to expect?

A1. The difference between the separabilities of life expectation is about 4%, which shows the relative insensitivity of the separability to large variations in the proportions of the considered samples. At the same time, the great difference in

Table 7.6.2: Optimal number of controls

	X=Survival status		
	Lived	Died	Total
Treated	0.840	0.040	0.880
Not treated	0.187	0.561	0.750
Total	1.027	0.601	1.630
Separability of x	0.456		
Overall content K	0.600		

overall content between both tables stresses the need of a clear distinction between "dependence" (in a given direction) and "content".

Q2. If choosing the proportion of treated persons at will were possible, how could one maximize the obtained information for a given total number of observations?

A2. As expected, the overall knowledge content K of the table for hospital I, where the numbers of treated and not treated patients differ excessively, is smaller. Goodman and Kruskal (1954) suggested that it may seem reasonable to choose samples of equal size to obtain some sort of standardization. A quantitatively more precise result can be obtained by choosing the relative number of controls (persons not treated) so that the overall content reaches a maximum. For the present data this occurs for 46% controls, as shown in Table 7.6.2. This proportion of control cases delivers a slightly larger overall content than the right-hand table in Table 7.6.1, and also a somewhat greater separability of life expectation.

References

Bernstein, F. (1966). *Selected contributions to the literature of blood groups and immunology (Dunsford Memorial): I. The ABO system.* Fort Knox: U.S. Army Medical Research Laboratory. (Originally published in 1924. Cited from Edwards, 1972.)

Blyth, C. R. (1972a). On Simpson's paradox and the sure thing principle. *Journal of the American Statistical Association, 67,* 364–366.

Blyth, C. R. (1972b). Some probability paradoxes in choice from among random alternatives (with comments by D. V. Lindley, I. J. Good, R. L. Winkler, and J. W. Pratt). *Journal of the American Statistical Association, 67,* 366–388. (Originally appeared in *Siam Review,* April, 1970.)

Brown, Morton B. (1975). The asymptotic standard errors of some estimates of uncertainty in the two-way contingency table. *Psychometrika, 40,* 291–296.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge: Cambridge University Press.

Fano, R. M. (1961). *Transmission of information.* New York: M.I.T. Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, Series A,* 222.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). Edinburgh: Oliver and Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Edinburgh: Oliver and Boyd.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Gaines, H. F. (1956). *Cryptoanalysis.* New York: Dover Publications.

Goodman, L. A. (1978). *Analyzing qualitative/categorical data, log-linear models and latent-structure analysis.* London: Addison-Wesley.

Goodman, L. A., & Kruskal, W. H. (1954). Measurement of association for cross classifications. *Journal of the American Statistical Society, 49,* 732–764. (Reprinted in L. A. Goodman & W. H. Kruskal, 1979, *Measures of association for cross classifications, Series in Statistics No. 1.* New York: Springer.)

Goodman, L. A., & Kruskal, W. H. (1959). Measurement of association for cross classifications. II: Further discussion and references. *Journal of the American Statistical Association, 54,* 123–163. (Reprinted in L. A. Goodman & W. H. Kruskal, 1979, *Measures of association for cross classifications, Series in Statistics No. 1.* New York: Springer.)

Goodman, L. A., & Kruskal, W. H. (1963). Measurement of association for cross classifications. III: Approximate sampling theory. *Journal of the American Statistical Association, 58,* 310–364. (Reprinted in L. A. Goodman & W. H. Kruskal, 1979, *Measures of association for cross classifications, Series in Statistics No. 1.* New York: Springer.)

Goodman, L. A., & Kruskal, W. H. (1972). Measurement of association for cross classifications. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association, 67,* 415–421. (Reprinted in L. A. Goodman & W. H. Kruskal, 1979, *Measures of association for cross classifications, Series in Statistics No. 1.* New York: Springer.)

Jaynes, E. T. (1978). Where do we stand on maximum entropy? In M. D. Levine & M. Tribus (Eds), *The maximum entropy formalism* (pp. 211–314). Cambridge, MA: M.I.T. Press. (Reprinted in *Papers on probability, statistics, and statistical physics* by R. D. Rosenkranz, Ed., 1983, Dordrecht/Boston: Reidel)

Jaynes, E. T. (1989). Clearing up mysteries—The original goal (pp. 13–14). In J. Skilling (Ed.), *Maximum entropy and Bayesian methods.* Dordrecht/Boston/London: Kluwer Academic Publications.

Kendall, M. G. (1943). *The advanced theory of statistics.* London: Ch. Griffin & Co.

Khinchin, A. I. (1957). *Mathematical foundations of information theory.* New York: Dover Publications.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44,* 187–192.

Lindley, D. V., & Novick, M. R. (1981). The Rôle of Exchangeability in Inference. *The Annals of Statistics*, *9*(1), 45–48.

McGill, W. J. (1954). Multivariate information Transmission. *Proceedings of Transactions PGIiT 1954 Symposium on information Theory*, *4*, 93–111.

Novick, M. R. (1980). Statistics as psychometrics. *Psychometrika*, *45*, 411–424.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal* (pp. 527–532). Cambridge: Cambridge University Press.

Rajski, C. (1961). A metric space of discrete probability distributions. *Information and Control*, *4*, 371–377.

Rajski, C. (1963). On the normed information rate of discrete random variables. *Zastosowania Mathematiki*, *6*, 459–461.

Särndal, Carl Erik (1974). A comparative study of association measures. *Psychometrika*, *39*, 165–187.

Schmitt, S. A. (1969). Measuring uncertainty, an elementary introduction to Bayesian statistics. Reading, MA: Addison Wesley.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Stouffer, S. A., et al. (1949). *The American Soldier: Adjustment during army life. Studies in social psychology in World War II, Vol. 1.* Princeton, NJ: Princeton University Press. (Cited from Goodman, 1978)

Székely, G. J. (1990). *Paradoxa*. Thun & Frankfurt: Verlag Harri Deutsch.

Theil, H. (1972). *Statistical decomposition analysis, with applications in the social and administrative sciences.* Amsterdam & London: North-Holland.

Vessereau, A. (1947). *La Statistique*. Paris: Presses Universitaires de France.

Wilks, S. S. (1935). The likelihood test of independence in contingency tables. *Annals of Mathematical Statistics*, *6*, 190–196.

Woolf, B. (1957). The log likelihood ratio test (the *G*-test). *Annals of Human Genetics*, *21*, 397–409.

Manuscript received 9/14/94
Final version received 10/2/95