Use of the Log Odds Ratio to Assess the Reliability of Dichotomous Questionnaire Data

D. A. Sprott and M. D. Vogel-Sprott
University of Waterloo, Canada

The use of the log odds ratio to measure test-retest reliability of dichotomous questionnaire response data is discussed. Its application is illustrated using questionnaire data on family history of problem drinking. The superiority of the log odds ratio as a measure of reliability of such data is discussed. Uninformative datasets are characterized.

Investigators often examine family histories to evaluate the potential contribution of genetic or other factors to such disorders as psychoses or alcoholism (Goodwin, 1981; Thompson, Ortasche, Prusoll, & Kidd, 1982; Volicer, Volicer, & D'Angelo, 1983). It is usually most practical to obtain family history information from the persons concerned, rather than from their relatives or from clinical records. This requires some standardized procedure which ensures that an individual remembers to consider each family member, and that the classification of each with respect to the presence or absence of the disorder is consistent over time. A standard questionnaire, which presents a schematic tree to guide a systematic consideration of all possible relatives, has been devised by some investigators in the field of alcoholism (Mann, Sobell, Sobell, & Pavan, 1985). Some research has also attempted to assess the test-retest reliability of this Family Tree Questionnaire (Mann et al., 1985; Vogel-Sprott, Chipperfield, & Hart, 1985).

However, the statistical analysis of test-retest reliability is problematic because family history questionnaire data have an unusual characteristic: The number of items (relatives) judged by each respondent is uncontrolled and varies with family size. In most personality or other questionnaires, each respondent considers the same number of items.

Because family history questionnaires are administered on two separate occasions (trials), responses from an individual (e.g., on the presence or absence of a disorder among family members) can be ordered in a 2 × 2 contingency table. Although the chi-square test of independence could be applied to the data in such a table, this test only detects the existence of a relationship between a person's judgments on the two occasions, and such a relationship is already known to exist. An assessment of the magnitude of the relationship is required in order to evaluate the degree of reliability of the data. The evaluation of the test-retest reliability of a family history questionnaire further requires the assessment of data from a group of persons, and a 2 × 2 chi-square test of such data is also inappropriate. The problem, therefore, is to

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 307–316
© Copyright 1987 Applied Psychological Measurement Inc.
0146-6216/87/030307-10$1.75

obtain a useful and convenient measure of the degree to which the probability of agreement between Trials 1 and 2 in a person's classification (p_1) exceeds the probability of disagreement (p_2).

The purpose of the present paper is to demonstrate the relevance of the standard analysis of 2 × 2 contingency tables to this problem. This analysis customarily uses the log odds ratio to measure the degree to which p_1 differs from p_2. This measure expresses test-retest reliability, defined as the reproducibility on Trial 2 of the results on Trial 1, in terms of odds rather than in terms of probabilities. It has been applied to similar repeated-trial, dichotomous-response data in the areas of perception, biostatistics, and clinical trials (Breslow & Day, 1980; Bryden & Sprott, 1981).

The next section sets out the form and notation for the data from an individual in a 2 × 2 contingency table. This leads to a statement of the assumptions which produce the statistical model of the 2 × 2 contingency table. The log odds ratio, θ, arising from this statistical model is defined. The following sections summarize the application of the general theory of maximum likelihood estimation of θ to (1) the data arising from an individual, (2) the combination of data from a group of persons to estimate a common θ value, and (3) a test of the homogeneity of the reliability estimates from different persons.

These methods require that the 2 × 2 cell frequencies from a person are not too small. In many cases where the reliability is high, some cell frequencies will be very small or 0. Family history data are likely to have this characteristic. Such data require the use of the exact analysis of Fisher (1935, p. 50), which is outlined below and applied to the data of Vogel-Sprott et al. (1985). Observations of a certain type are uninformative with respect to reliability and must be discarded in the above analysis. The final section discusses other measures which might be considered (e.g., the kappa index) and the reasons for selecting θ over these other measures.

Assessing Reliability With the Log Odds Ratio

Form and Notation

When a person reports the presence or absence of a specific disorder among family members on two separate occasions (trials), these data can be put into the form shown in Table 1. Here x is the number of relatives classified by the person on both Trials 1 and 2 as having the disorder (D); m is the total number of relatives classified on Trial 1 as D; y is the number classified on Trial 1 as not having the disorder (ND), but reclassified on Trial 2 as D; and n is the total number classified as ND on Trial 1. Test-retest reliability measures the extent to which these classifications are reproduced from one trial to the next. Thus the diagonal cells x and $(n - y)$ represent agreement between Trials 1 and 2, and contribute to test-retest reliability, while the diagonal cells y and $(m - x)$ represent disagreement between Trials 1 and 2, and detract from test-retest reliability.

Table 1
2 x 2 Classification
of Response Frequencies

		Trial 2		
		D	ND	Total
Trial 1	D	x	m-x	m
	ND	y	n-y	n
	Total	t	m+n-t	m+n

Assumptions and Statistical Model; The Log Odds Ratio

These considerations lead to the statistical model for the data from a person, shown in Table 2. Here p_1 is the *conditional* probability that a given relative is classified as D on Trial 2 after having been classified as D on Trial 1, and p_2 is the corresponding conditional probability that a given relative is classified as D on Trial 2 after having been classified as ND on Trial 1.

Table 2
Contingency Table of
Conditional Probabilities

		Trial 2	
		D	ND
Trial 1	D	p_1	$1-p_1$
	ND	p_2	$1-p_2$

It is assumed that p_1 and p_2 are constant for all relatives classified by a given person, but are not necessarily the same for all persons. Note that since p_1 and p_2 are conditional probabilities of Trial 2 given Trial 1, this does not imply that a given person necessarily classifies all relatives as D or ND with the same probability. For example, a person might be more disposed to classify his father than his mother as D. The assumption is only that *if* they have both been classified as D on Trial 1, then the probability of both being classified separately as D on Trial 2 is p_1, the same, independently, for each (so that the probability of *both* being classified as D on Trial 2 would be p_1^2).

It is also assumed that a person's classification of one relative does not influence that person's classification of another. The assumption that relatives are classified independently by a person means that the classification of a relative as D or ND does not affect, and is not affected by, how another relative is classified by that person.

Under these conditions, the probability of a set of observations (x,y) given (m,n) is the product of two binomial distributions:

$$P(x,y;\, p_1,p_2) = \binom{m}{x} p_1^x \,(1-p_1)^{m-x} \binom{n}{y} p_2^y \,(1-p_2)^{n-y} \quad . \qquad (1)$$

The symbols $\binom{m}{x}$, $\binom{n}{y}$ are the "binomial coefficients" defined as $m!/x!(m-x)!$, $n!/y!(n-y)!$, discussed in standard books on probability (e.g., Kalbfleisch, 1985).

This statistical model is that of the 2 × 2 contingency table, which occurs in all fields of science in which data can occur in the form of cross-classified frequencies. It has a sizable literature, dating back to Pearson (1900). The generally accepted analysis of 2 × 2 tables is that of Fisher (1922, 1935, 1970).

In this model, test-retest reliability is defined as the repeatability of the results of Trial 1 to Trial 2. It is thus a function of the extent to which p_1 exceeds p_2. If $p_1 = p_2$, test-retest reliability is 0; the probability of being classified as D on Trial 2 is independent of the result of Trial 1. The standard chi-square test for independence in a 2 × 2 table, as well as Fisher's "exact test" (Fisher, 1970, pp. 96–97), are both tests of the hypothesis $p_1 = p_2$. However, in the present context, the hypothesis $p_1 = p_2$ is not the issue. It is the magnitude of the difference between p_1 and p_2 which is of interest.

In the classical analysis of the 2 × 2 contingency table, this difference is usually measured by the log odds ratio, defined by

$$\theta = \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) \quad . \tag{2}$$

The quantities $p_1/(1 - p_1)$ and $p_2/(1 - p_2)$ are the odds that a given relative is classified as D on Trial 2, after having been classified as either D or ND on Trial 1. The analysis is more conveniently performed on the odds than on the probabilities p_1 and p_2 themselves; this analysis is outlined in the next section.

It is important to note that the log odds is essentially independent of whether the probabilities are conditioned on Trial 1 (as above) or on Trial 2, or whether p_1 and p_2 are the probabilities of D or ND. That is, if θ in Equation 2 is denoted by $\theta_1(\mathrm{D}_2)$, then $\theta_1(\mathrm{D}_2) = \theta_2(\mathrm{D}_1) = -\theta_1(\mathrm{ND}_2) = -\theta_2(\mathrm{ND}_1)$, so that the log odds ratio is invariant, except for a change in sign, for the four obvious ways of viewing the data.

Maximum Likelihood Estimation of θ Based on a Single Person

The maximum likelihood estimate (MLE) of θ, based on the data in Table 1 from a single person, is denoted by $\hat{\theta}$, and its estimated variance by $\hat{\sigma}^2$. The numerical values of $\hat{\theta}$ and $\hat{\sigma}^2$ are given by

$$\hat{\theta} = \log\left[\frac{x(n-y)}{y(m-x)}\right] \tag{3}$$

$$\hat{\sigma}^2 = \frac{1}{x} + \frac{1}{m-x} + \frac{1}{y} + \frac{1}{n-y} \tag{4}$$

(see, e.g., Breslow & Day, 1980, Equation 4.18).

Let u be the standardized difference between θ and its MLE, $\hat{\theta}$. This is given by

$$u = (\hat{\theta} - \theta)/\hat{\sigma} \quad . \tag{5}$$

If the cell frequencies $(x, m-x, y, n-y)$ are not too small (i.e., 5 or less), then u has an approximate standard normal distribution. Assuming u has an approximate standard normal distribution,

$$\theta = \hat{\theta} \pm \hat{\sigma}u \tag{6}$$

are the approximate confidence intervals for θ, where u is chosen from the standard normal distribution to give the desired confidence level (e.g., $u = 1.96$ for a .95 confidence interval). These approximate intervals are fairly accurate, even when the cell frequencies are rather small.

Maximum Likelihood Estimation of θ Based on Several Persons

Although the test-retest reliability of such questionnaire data from a single person may be of interest, assessment of the test-retest reliability of the questionnaire itself requires the combination of data from a group of persons. This can be obtained by an extension of the procedures described above.

Suppose there are r persons, each yielding the data $(x_i, m_i - x_i, y_i, n_i - y_i)$, where $i = 1, 2, \ldots, r$. Then, using Equations 3 and 4, the MLE and its corresponding estimated variance can be calculated for each person, yielding values $\hat{\theta}_i$ and $\hat{\sigma}_i$ $(i = 1, 2, \ldots, r)$. If none of the cell frequencies is too small, then the approximate overall MLE and its estimated variance, based on all r persons, are calculated by the formulas

$$\hat{\theta} = \frac{\sum(\hat{\theta}_i/\hat{\sigma}_i^2)}{\sum(1/\hat{\sigma}_i^2)} \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{\sum(1/\hat{\sigma}_i^2)} \tag{8}$$

These can be used in Equation 5 to calculate confidence intervals for θ based on a group of persons.

Homogeneity of the $\hat{\theta}_i$s

The general utility of a questionnaire is enhanced if its retest reliability is consistent over a broad range of persons with different demographic or personal characteristics. The method of maximum likelihood also can be used to test the consistency of the estimates of test-retest reliability obtained from a group of r persons.

If the actual test-retest reliability for a given person i were known, it would be equal to the numerical value of θ_i. However, because the numerical value of θ_i is unknown, its MLE, $\hat{\theta}_i$, can be used, and is obtained from Equation 3. The standardized difference between θ_i and $\hat{\theta}_i$, obtained from Equation 5, is expressed for each of the r persons as

$$u_i = \frac{(\hat{\theta}_i - \theta_i)}{\hat{\sigma}_i} \qquad (i = 1, 2, \ldots, r) \quad . \tag{9}$$

With adequate sizes of the cell frequencies, these u_is are all approximate independent standard normal variates, so that their sum of squares, $\sum u_i^2$, is an approximate $\chi^2_{(r)}$ variate, that is, a chi-square with r degrees of freedom.

The hypothesis that the actual test-retest reliability does not differ among persons is expressed as H: $\theta_1 = \theta_2 = \ldots = \theta_r = \theta$. The numerical value of the common θ is also unknown, but its MLE, $\hat{\theta}$, is obtained from Equation 7. The hypothesis H may be tested by substituting $\hat{\theta}$ for the unknown θ in Equation 9, to give the estimated standardized difference, $\hat{u}_i$, for each person:

$$\hat{u}_i = \frac{(\hat{\theta}_i - \hat{\theta})}{\hat{\sigma}_i} \qquad (i = 1, 2, \ldots, r) \quad . \tag{10}$$

Then $\sum \hat{u}_i^2$ will have an approximate $\chi^2_{(r-1)}$ distribution, which can be used to test H, the homogeneity of the θ_is. A non-significant χ^2 indicates that the test-retest reliability, as measured by θ, is consistent among persons on the test. (More precisely, there is no evidence of inconsistency.)

Small Frequencies

One important limitation to the above use of maximum likelihood methods occurs when some of the cell frequencies for a person are small (i.e., less than 5). In such cases, the results will be inaccurate and their combination for a group of persons could produce an appreciable bias in the overall estimate. Thus, although these methods have many important applications, they may not be specifically appropriate to family history data, especially if the reliability is high. In such cases the observations will be characterized by zero cell frequencies. For example, the data of Vogel-Sprott et al. (1985) on problem drinkers in the immediate family (parents and siblings) indicate that respondents seldom had more than five first-degree relatives, and when their reports were consistent on both administrations of the Family Tree Questionnaire, two cells of the resulting 2 × 2 table contained zero frequencies.

For such data, an exact procedure is available in terms of the log odds ratio, which is similar to the method described above. This exact procedure is based on Fisher's (1935) demonstration that the conditional distribution of x and y, with the row and column totals held constant at their observed values (as in Table 1), depends on θ only. This distribution is given by the formula

$$P(x;\theta|m,n,t) = \frac{c(x,\theta)}{\sum\limits_{j\geqslant 0} c(j,\theta)} \quad , \tag{11}$$

where the quantities $c(j,\theta)$ are given by

$$c(j,\theta) = \binom{m}{j}\binom{n}{t-j}\exp(j\theta) \quad . \tag{12}$$

The symbols $\binom{m}{j}$ and $\binom{n}{t-j}$ are the binomial coefficients described in Equation 1. The sum in the denominator of Equation 11 is over all j such that $c(j,\theta) > 0$. For example, consider the table of data shown in the fourth line from the bottom of Table 3. The data are $x = 2$, $m - x = 0$, $y = 0$, $n - y = 3$, so that $m = 2$, $n = 3$, and $t = 2$. Equation 12 is applied iteratively with the value of j taken as 0, 1, and 2 on successive iterations. The sum of the terms thus calculated forms the denominator of Equation 11, and the term using $j = 2$ (the observed value of x) is the numerator. The ratio thus obtained is the probability of the observed table of data given its marginal totals, that is, $P(x = 2;\theta|m = 2,\ n = 3,\ t = 2)$.

The distribution given by Equation 11 is exact no matter how small the cell frequencies. Thus Equations 11 and 12 can, in principle, be used to set up confidence intervals for θ, even when some of the cell frequencies are 0. An illustration is presented below.

Application to Family History Data

The Vogel-Sprott et al. (1985) study administered the Family Tree Questionnaire of problem drinking to 60 male persons, and retested them four months later. Of this group, 24 persons reported at least one problem drinker in their immediate family on both tests (trials). For reasons to be explained in the next section, the data on the remaining 36 persons do not provide information pertinent to the assessment of test-retest reliability. Thus only the data from the subset of 24 persons are used in the exact procedure described below.

For each of these 24 persons, the data in the four cells of Table 1 may be more succinctly represented in a horizontal sequence $(x_i, m_i - x_i, y_i, n_i - y_i)$. Here x is the number of relatives classified by a person as having the disorder (D) on both Trials 1 and 2; $m - x$ is the number classified as D on Trial 1 and ND on Trial 2; y is the number classified as ND on Trial 1 and D on Trial 2; $n - y$ is the number classified as ND on both trials. Table 3 presents these data. Because many of the persons produced 2 × 2 tables with the same cell frequencies, the number of persons contributing to each is given in the first column.

Because every table has zero frequencies in one or more cells, the exact method must be used. The third column shows the probabilities of the tables listed in the second column, as calculated from Equations 11 and 12. The probability of the entire observed sample shown above is the product of the expressions in the third column raised to the powers N given in the first column:

$$\begin{aligned} P = {} & \left[\frac{\exp(\theta)}{1+\exp(\theta)}\right] \times \left[\frac{\exp(\theta)}{2+\exp(\theta)}\right]^5 \times \left[\frac{\exp(\theta)}{3+\exp(\theta)}\right]^5 \times \left[\frac{\exp(\theta)}{4+\exp(\theta)}\right]^4 \times \left[\frac{\exp(\theta)}{5+\exp(\theta)}\right]^2 \times \left[\frac{\exp(\theta)}{6+\exp(\theta)}\right] \\ & \times \left[\frac{\exp(\theta)}{7+\exp(\theta)}\right]^2 \times \left[\frac{\exp(2\theta)}{3+6\exp(\theta)+\exp(2\theta)}\right] \times \left[\frac{2\exp(\theta)}{3+2\exp(\theta)}\right]^2 \times \left[\frac{\exp(\theta)}{1+\exp(\theta)}\right] \ . \end{aligned} \tag{13}$$

For these data, the probability of the observed sample, as well as the probability of all the more extreme samples (i.e., the tail area of the distribution), is simply the probability P, given by Equation

Table 3
Observed Cell Frequencies in 2 x 2 Tables and Their Probabilities Obtained from 24 Individuals Reporting At Least One Problem Drinking Relative On Both Trials

Number (N) of Persons	Cell Frequencies x	n-x	y	n-y	Probability of the Table
1	1	0	0	1	$\exp(\theta)/[1+\exp(\theta)]$
5	1	0	0	2	$\exp(\theta)/[2+\exp(\theta)]$
4	1	0	0	3	$\exp(\theta)/[3+\exp(\theta)]$
1	3	0	0	1	
4	1	0	0	4	$\exp(\theta)/[4+\exp(\theta)]$
2	1	0	0	5	$\exp(\theta)/[5+\exp(\theta)]$
1	1	0	0	6	$\exp(\theta)/[6+\exp(\theta)]$
2	1	0	0	7	$\exp(\theta)/[7+\exp(\theta)]$
1	2	0	0	3	$\exp(2\theta)/[3+6\exp(\theta)+\exp(2\theta)]$
1	1	0	1	3	$2\exp(\theta)/[3+2\exp(\theta)]$
1	1	1	0	3	
1	1	0	1	2	$\exp(\theta)/[1+\exp(\theta)]$

13, of the observed sample itself. Because the numerical value of P depends on the numerical value of θ, it is easy to determine the values of θ that make the observed sample relatively probable. Table 4 shows some of the values of θ substituted in Equation 13, and the resulting values of P.

Table 4 shows that smaller values of θ make the observed sample less probable, and that values of θ less than 0 make it almost impossible. Similarly, the probability of obtaining the observed sample increases as θ increases, and is most probable (i.e., $P = 1$) when $\theta \rightarrow \infty$. Thus, in the context of these family history data, the evidence indicates that the log odds ratio of classifying a problem drinker the same way on two trials is almost certainly larger than 0. This means that the odds of a relative being classified as a problem drinker on Trial 2 are greater if he/she was also classified as a problem drinker on Trial 1.

From Table 4, $\theta = 3.23$ makes the probability of the observed sample .05. Thus the hypothesis H: $\theta = 3.23$ is significant at exactly $\alpha = .05$, and $3.27 \leqslant \theta \leqslant \infty$ is a 95% confidence interval. Because $\theta = 3.23$ is the log odds ratio, the actual odds ratio is exp(3.23) = 25.28. This means that at the 95% confidence level, the results of Trial 2 repeat those of Trial 1, in the sense that the odds of classifying a given relative as a problem drinker on Trial 2 are at least 25 times greater if he was also classified as a problem drinker on Trial 1.

Table 4 also shows that the 99% and 99.9% confidence intervals are $2.76 \leqslant \theta \leqslant \infty$ and $2.30 \leqslant \theta \leqslant \infty$, respectively. These serve to put lower bounds on θ at the 99% and 99.9% confidence levels respectively, and have interpretations similar to that of the 95% confidence interval.

In general, the use of Equations 11 and 12 to combine data from several persons is complicated. The above numerical example is simple because the homogeneity is obvious—most of the tables are close to being diagonal (m, 0, 0, n) and thus exhibit maximal test-retest reliability with MLE $\hat{\theta} = \infty$—and only the probability given by Equation 13 of the observed sample needs to be calculated. In other cases where the reliability is not as high, homogeneity may not be obvious, and confidence intervals will require the calculation of "tail area" probabilities, thus requiring the calculation of the probabilities of samples

Table 4
Some Numerical Values of the Log Odds Ratio θ and the Corresponding Numerical Values of the Probability P of the Sample Given by Equation 9 for the Data in Table 3

θ	0	2.30	2.76	3.23	∞
P	1.59×10^{-15}	0.001	0.10	0.05	1.00

additional to the one observed. In these cases the calculations could be simplified, and indeed reduced to those of Table 3, by assessing θ using the likelihood function, yielding relative likelihoods and likelihood intervals rather than significance levels and confidence intervals. This approach is discussed and exemplified by Sprott and Kalbfleisch (1965). But the computational complexities of Equations 11 and 12 are beyond the scope of this paper. These have been detailed by Thomas (1975).

Uninformative Results

The preceding analysis of the test-retest reliability of the Family Tree Questionnaire of problem drinking was based on the subset of 24 persons who reported at least *one* problem-drinking relative on *each* trial. The remaining 36 persons in the study reported *no* problem drinker on *at least one* of the trials. Of these, 31 actually classified no relative as D on either trial, hence their 2 × 2 tables had the form (0, 0, 0, *n*). The data of these 36 persons were not included in the above analysis because the statistical and logical aspects of these data make them uninformative about test-retest reliability as previously defined.

On statistical grounds, the MLE from these data is undefined (∞/∞), and the estimated standard errors are meaningless. The exact distribution given by Equations 11 and 12 is degenerate, as it assigns probability values of 1 to the observed sample and 0 to all other samples, irrespective of θ.

The logical issue is particularly important, and perhaps more subtle, because all persons with tables of the form (0, 0, 0, *n*) have given exactly the same responses twice. Thus such data may be thought to present strong evidence in favor of test-retest reliability, and the above analysis may be thought to be defective in its inability to use them. Further consideration, however, shows that the above analysis is correct in not using such tables, because they imply no evidence either for or against reliability as previously defined. This may be understood by noting that the analysis compares p_1 (the probability that a relative is classified as D on Trial 2 after having been classified as D on Trial 1) with p_2 (the probability that a relative is classified as D on Trial 2 after having been classified as ND on Trial 1). But when the result is of the form (0, 0, 0, *n*), *no* relative has been classified as D on Trial 1. Thus the data contain no evidence whatever about the numerical value of p_1, and in particular, whether p_1 differs from p_2.

Even if there is no test-retest reliability, the probability of observing (0, 0, 0, *n*) can be large. For example, if a person classifies 56 relatives to yield the 2 × 2 table with frequencies (1, 6, 7, 42), then the estimated values of p_1 and p_2 are equal (1/7 and 7/49 respectively); such a result thus provides no evidence of reliability. But it is important to note that the probability of classifying a relative as D on Trial 1 is only $7/56 = 1/8$. Suppose that these are in fact the true probabilities; that is, there is no reliability ($p_1 = p_2 = 1/7$) and a low probability (1/8) of classifying a relative as D. Then, if the total number of relatives classified by a person is small, there is a reasonably large probability of observing *no* relatives classified as D. In fact, if a total of only 8 relatives are classified by the person (instead of 56), the probability of having no relatives classified as D is $(7/8)^8 = .3436$. That is, the probability of observing a table (0, 0, y, $8 - y$), where y can be any number between 0 and 8, is .3436. Then the probability of observing $y = 0$ is $(1 - p_2)^8 = (6/7)^8 = .2914$. Thus the probability of observing the table (0, 0, 0, 8) is the product (.3436)(.2914) = .1001. Hence, in this example where there is no test-retest reliability, the probability of classifying a small number of ND relatives identically on two different occasions—observing (0, 0, 0, n)—is nonetheless reasonably large.

This example shows that 2 × 2 tables of the form (0, 0, 0, n) can be observed quite often in the absence of any retest reliability. Hence such an observation cannot be taken as evidence for test-retest reliability. Naturally, because such an observation could also occur when there is reliability, it also cannot be taken as evidence against reliability. The same argument applies to tables of the form (n, 0, 0, 0). In fact, the same comment applies to all tables that contain a row or column of 0s. This point has been recognized in other fields which generate cross-classified dichotomous data with similar characteristics. For example, in studies using identical twins, it is well known that only the *discordant* pairs provide information about θ; the data from all concordant pairs must be discarded.

Discussion

The analysis presented in the preceding sections can also be used to assess validity. In this case, the two trials are replaced by two judges, for example, the person and a physician. The resulting analysis then measures the degree to which the judges agree on their classification of the same family members.

Many different measures of association in the 2 × 2 table have been suggested (Kendall & Stuart, 1961). In the absence of a quantitative scientific theory sufficiently precise to predict a particular measure, such as a measure of linkage in Mendelian genetics, the selection of a measure is to some extent arbitrary, and must be based upon its logical, mathematical, and statistical properties. Although the analysis of 2 × 2 tables usually employs the log odds ratio θ, some other measures that have been used are the simple difference of probabilities $(p_1 - p_2)$, the ratio of probabilities (p_1/p_2), the log of p_1/p_2, z scores, and the kappa index (κ) defined by Cohen (1960).

The measure κ was designed specifically to assess the agreement between judges in the classification of the same set of items, and has been applied to data of the type considered here (Mann et al., 1985; Vogel-Sprott et al., 1985). The κ index was developed to measure the agreement between two judges when assigning the same set of N items to one of r categories. The observations arising from such an experiment form an $r \times r$ contingency table, for which the measure κ of agreement between the two judges is defined by

$$\kappa = (p_0 - p_c)/(1 - p_c) \quad , \tag{14}$$

where p_0 is the total probability of agreement between the two judges, and p_c is total probability of agreement expected by chance, that is, assuming the judges assigned persons to categories at random. This definition ensures that $\kappa = 0$ if the judges assign randomly, and $\kappa = 1$ when there is total agreement, that is, when the table is diagonal (m, 0, 0, n).

Although the κ index has been applied to assess the reliability of the data discussed in this paper, and might also be applied to assess validity, the properties of the log odds ratio make it a more convenient and appropriate measure for analyzing 2 × 2 tables. These properties are:

1. The log odds ratio is unbounded. It can vary unrestrictedly between $-\infty$ and $+\infty$. Specifically, $\theta < 0$ when $p_1 < p_2$; $\theta = 0$ when $p_1 = p_2$; and $\theta > 0$ when $p_1 > p_2$. This property facilitates its use in statistical analyses, such as regression, where the relationship of θ to other variables, such as age, is of interest. The unbounded variation of θ allows it in principle to be linearly related to other variables. This allows the application of standard statistical techniques to analyze linear regression models involving θ.

Bounded measures, such as κ or $p_1 - p_2$ (which varies between -1 and 1) cannot, even in principle, be linearly related over a wide range to other variables. Thus statistical analyses involving such measures are more difficult. In this regard it should be noted that the actual odds ratio $[p_1/(1 - p_1)]/[p_2/(1 - p_2)]$, which is $\exp(\theta)$, only varies from 0 to ∞. It is for this reason that its logarithm, θ, is used.

2. The conditional distribution, given by Equation 11, is a function of θ only. This allows an exact analysis of θ to be performed when cell frequencies are small or 0. Such data cannot easily be analyzed using other measures.

Contemporary research in statistics might be applied to generalize θ from the 2×2 table to the $r \times r$ table, as discussed by Cohen (1960) and later authors, to produce a measure preferable to κ. The current theory of log-linear models is specifically designed for such a problem.

References

Breslow, N., & Day, N. E. (1980). *Statistical methods in cancer research: Vol. 1. The analysis of case control studies*. Lyon, France: International Agency for Research on Cancer.

Bryden, M. P., & Sprott, D. A. (1981). Statistical determination of degree of laterality. *Neuropsychologia, 19*, 571–581.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Fisher, R. A. (1922). On the interpretation of chi square from contingency tables, and the calculation of *P*. *Journal of the Royal Statistical Society, 85*, 87–94.

Fisher, R. A. (1935). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society, 98*, 39–54.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed., pp. 96–97). Edinburgh: Oliver and Boyd.

Goodwin, D. (1981). Family studies of alcoholism. *Journal of Studies on Alcohol, 42*, 156–162.

Kalbfleisch, J. G. (1985). *Probability and statistical inference. Vol. 1: Probability*. New York: Springer Verlag.

Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics, Vol. 2* (pp. 536–591). London: Charles Griffin and Co.

Mann, R., Sobell, L., Sobell, M., & Pavan, D. (1985). Reliability of a family tree questionnaire for assessing family history of alcohol problems. *Drug and Alcohol Dependence, 15*, 61–68.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 1*, 157.

Sprott, D. A., & Kalbfleisch, J. G. (1965). Use of the likelihood function in inference. *Psychological Bulletin, 64*, 15–22.

Thomas, D. G. (1975). Exact and asymptotic methods for the combination of 2 × 2 tables. *Computers and Biomedical Research, 8*, 423–446.

Thompson, W., Ortasche, H., Prusoll, B., & Kidd, K. (1982). An evaluation of the family history method for ascertaining psychiatric disorders. *Archives of General Psychiatry, 39*, 53–58.

Vogel-Sprott, M., Chipperfield, B., & Hart, D. (1985). Family history of problem drinking among young male social drinkers: Reliability of the Family Tree Questionnaire. *Drug and Alcohol Dependence, 16*, 251–256.

Volicer, B., Volicer, L., & D'Angelo, N. (1983). Variation in length of time to development of alcoholism by family history of problem drinking. *Drug and Alcohol Dependence, 12*, 69–83.

Acknowledgments

This research was supported in part by National Science and Engineering Research Council grants #A3099 to the first author and #A8321 to the second author. The authors thank M. P. Bryden, S. Cronshaw, H. Lei, R. Mann, and R. V. Thysell for helpful comments on earlier drafts of the paper.

Author's Address

Send requests for reprints or further information to D. A. Sprott, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.