Discrete Mathematics 92 (1991) 39–47
North-Holland

On a discrete distribution associated with the statistical assessment of nominal scale agreement*

S.R. Chamberlin and D.A. Sprott

Department of Statistics and Actuarial Science, University of Waterloo, Ont., Canada N2L 3G1

Received 19 September 1989

Dedicated to Professor R.G. Stanton on the occasion of his 68th birthday.

Abstract

Chamberlin, S.R. and D.A. Sprott, On a discrete distribution associated with the statistical assessment of nominal scale agreement, Discrete Mathematics 92 (1991) 39–47.

A new measure of agreement is introduced to measure and assess the extent of agreement between two judges or raters in assigning n subjects to L unordered qualitative categories. The statistical analysis entails an interesting discrete conditional distribution, and provides an example of 'conditional' inference and 'exact' tests of significance. In fact, the measure is related to the log odds ratio in a 2×2 contingency table and its analysis is related to Fisher's (1935) conditional exact analysis of the 2×2 table.

1. Introduction

Consider two raters or judges who assign n subjects independently to L qualitative unordered categories $C_1, \ldots, C_L$. The resulting observations form an $L \times L$ contingency table $X = \{x_{ij}\}$, $i, j = 1, \ldots, L$, where x_{ij} is the number of subjects assigned to categories C_i, C_j simultaneously by Raters 1 and 2 respectively. The problem is to assess the extent to which the raters agree in their assignment of subjects to categories. The diagonal cell frequencies $\{x_{ii}\}$ represent agreements and the remaining cell frequencies represent disagreements. All disagreement cells are assumed to have the same importance and are hence weighted equally. The same is true of the agreement cells. For example, the raters may be diagnosticians assigning subjects to diagnostic classes such as normal, neurotic and psychotic. The extent to which the diagnosticians agree is an

* Research supported in part by NSERC.

indication of the merit or usefulness of the diagnostic classes. A diagnostic classification is not of much use if no two doctors can agree upon the diagnosis of any given subject.

Cohen [1] introduced kappa (κ) to assess the amount of agreement between two judges or raters in assigning subjects to L unordered qualitative categories like the above. He defined κ to be

$$\kappa = \frac{[\sum \text{pr}(i, i) - \sum \text{pr}(i.)\text{pr}(.i)]}{[1 - \sum \text{pr}(i.)\text{pr}(.i)]},$$

where pr(i, i) is the *joint* probability that a given subject is classified in category C_i by both raters, pr$(i.)$ is the marginal probability of being classified in category C_i by Rater 1, and pr$(.i)$ is the corresponding marginal probability for Rater 2. The definition of kappa ensures that if the agreements are totally fortuitous, so that

$$\text{pr}(i, i) = \text{pr}(i.)\text{pr}(.i),$$

then $\kappa = 0$, and if the agreement is perfect, $\sum \text{pr}(i, i) = 1$, then $\kappa = 1$. If the cell frequencies x_{ij} are not small, then the estimate $\hat{\kappa}$ of κ, obtained by replacing p_{ij} by x_{ij}/n, tends to be normally distributed, so that statistical tests can be applied to assess the evidence. Kappa has been widely used and modified. Indeed, it now seems to be the principal measure of agreement advocated with nominal scales such as the L categories described above.

However, a peculiar feature of this problem is that the more successful are the categories $\{C_i\}$ at discriminating among the subjects, the more problematic will be a statistical analysis based on normal approximations. For, in the favourable case of strong agreement between the raters, the cell frequencies X will tend to concentrate along the diagonal. Then $\{x_{ii}\}$ will be large, and $\{x_{ij},\ i \neq j\}$, will be small. Statistical analyses based on asymptotic normality will then be of questionable accuracy.

This difficulty can be avoided in the special case of $L = 2$ categories. Fisher [3] developed an 'exact' analysis for the 2×2 contingency table, based on the *conditional* probability of x_{11} *given* the row and column totals. This conditional probability function depends only on the log odds ratio. Hence if the amount of agreement between the two judges is defined to be the log odds ratio, this exact analysis can be applied, and the difficulties associated with small cell frequencies avoided. This was exemplified in [4] in the context of test–retest reliability, where the two raters are the same person on two different occasions.

The purpose here is to develop a measure, having the same properties as the log odds ratio, to measure agreement in the general case of an $L \times L$ table. In addition, a conditional probability function will be obtained which will serve as the basis for an exact analysis, valid when the cell frequencies are small or zero. When $L = 2$, the log odds ratio is obtained, and the conditional probability function will be equivalent to that of [3] cited above.

2. Derivation of the measure of agreement

In the $L \times L$ contingency table formed by the observations X, let p_{ij} be the theoretical *conditional* probabilities for the assignment of a given subject to a category,

$$p_{ij} = \text{pr}(C_j \text{ by Rater 2} \mid C_i \text{ by Rater 1}),$$
$$\sum_{j=1}^{L} p_{ij} = 1, \quad i = 1, \ldots, L. \tag{1}$$

Under the assumption that the probabilities (1) are constant across subjects, and that the raters classify subjects independently, for a fixed i the frequencies $x_{i1}, x_{i2}, \ldots, x_{iL}$ have an L-variate multinomial distribution with probability parameters (1) and index

$$r_i = \sum_{j=1}^{L} x_{ij}. \tag{2a}$$

The logarithm of this distribution is

$$\log P_i = \log P_i(x_{i1}, \ldots, x_{iL}; p_{i1}, \ldots, p_{iL})$$
$$= \log K_i + \sum_{j=1}^{L} x_{ij} \log(p_{ij}), \qquad K_i = \left(\frac{r_i!}{\prod_{j=1}^{L} x_{ij}!} \right). \tag{2b}$$

The logarithm of the distribution of all the frequencies X is the sum of the logarithms of these L L-variate multinomial distributions

$$\log P(\{x_{ij}\}; \{p_{ij}\}) = \sum_{i=1}^{L} \log P_i. \tag{3}$$

Notice that, under this model, the probability of the observations X is conditioned on the row totals $\{r_i\}$ at the outset.

Let

$$\alpha_{ij} = \log p_{ii}/p_{ij}. \tag{4a}$$

Note that

$$-\infty < \alpha_{ij} < \infty, \quad i \neq j, \quad \alpha_{ii} = 0, \tag{4b}$$

so that, subject to the restrictions in (1), the mapping $\{p_{ij}\} \leftrightarrow \{\alpha_{ij}\}$, $i \neq j$, is one-to-one. The quantity $\exp \alpha_{ij}$ is the *odds* that a given subject is assigned to category C_i versus C_j by Rater 2 *conditional* on being assigned to category C_i by Rater 1. Hence α_{ij} is the conditional log odds of agreement versus disagreement on categories C_i versus C_j, given C_i by Rater 1. Thus α_{ij} measures the difference between the agreement cell (i, i) and the disagreement cell (i, j). Also the functional form (4a) of each α_{ij} $(i \neq j)$ weights equally the agreement and disagreement cells. The fact that the categories $\{C_i\}$ are qualitative and

unordered implies that all such comparisons given by the $\{\alpha_{ij}\}$ are to be weighted equally. This yields the $L(L-1)/2$ log odds ratios

$$\nu_{ij} = \alpha_{ij} + \alpha_{ji}, \quad i<j, \tag{5a}$$

as the measures of agreement between the raters on categories C_i and C_j, $i, j = 1, 2, \ldots, L$, and

$$\bar{\nu} = \frac{2\nu}{L(L-1)}, \quad -\infty < \bar{\nu} < \infty, \tag{5b}$$

where

$$\nu = \sum\sum_{i<j} \nu_{ij} = \sum\sum_{i\neq j} \alpha_{ij} = \sum_i \sum_j \alpha_{ij}, \tag{5c}$$

as the measure of the overall state of agreement between the two raters.

The interpretation of these measures is that the odds of one rater classifying a subject in category C_i versus C_j are $\exp \nu_{ij}$ times greater if the subject was classified in category C_i rather than C_j by the other rater. The quantity $\exp(\bar{\nu})$ is the geometric mean of these odds ratios over all $L(L-1)/2$ pairs of categories. It thus may be thought of as the extent to which, on the average, the odds of a subject being classified in a given category by one rater are increased by being classified in the same category by the other rater.

From (4), (5c) can be written

$$\nu = \left(L \sum_{i=1}^{L} \log p_{ii}\right) - \sum_{i=1}^{L}\sum_{j=1}^{L} \log p_{ij}. \tag{6}$$

The quantity ν is zero if the rows and columns are statistically independent, so that the agreement is fortuitous. The results of one rater cannot be predicted from the results of the other.

3. Statistical inference for $\bar{\nu}$

It is mathematically more convenient to deal with ν than with $\bar{\nu}$. Inferences about ν can be directly converted into inferences about $\bar{\nu}$. From (4a), (4b), and (1),

$$p_{ij} = \frac{[\exp(-\alpha_{ij})]}{D_i}, \quad D_i = \sum_{j=1}^{L} \exp(-\alpha_{ij}). \tag{7}$$

Thus the probability (3) of the observations X expressed as a function of the $\{\alpha_{ij}\}$, and of ν given by (5c), is

$$\begin{aligned} P &= K \exp\left[\left(-\sum\sum \alpha_{ij}x_{ij}\right) - \sum r_i \log D_i\right] \\ &= K(\exp -x_{12}\nu)\left\{\exp\left[-\left(\sum\sum_{i\neq j} a_{ij}\alpha_{ij}\right) - \sum r_i \log D_i\right]\right\}, \end{aligned} \tag{8}$$

since $\alpha_{ii} = 0$ for all i, where

$$a_{ij} = x_{ij} - x_{12}, \quad (i \neq j), \qquad K = \prod K_i. \tag{9}$$

Since v is the parameter of interest about which inferences are to be made, it is necessary to eliminate the $\{\alpha_{ij}\}$ from (8). This can be done by conditioning (8) on the $\{a_{ij}\}$. The *conditional* probability of x_{12}, given $\{a_{ij}\}$ of (9) and the row totals $\{r_i\}$ of (2a), is proportional to (8) with $\{a_{ij}\}$ and $\{r_i\}$ held constant. Renormalizing (8) so that the total probability is unity with $\{a_{ij}\}$, and $\{r_i\}$ held constant yields the conditional probability of x_{12} to be

$$\begin{aligned} g(x_{12}; v) &= \mathrm{pr}(x_{12}; v \mid \{a_{ij}\}, \{r_i\}) \\ &= \frac{K(x_{12})(\exp - x_{12}v)}{\sum_h K(h)(\exp - hv)} \end{aligned} \tag{10}$$

where

$$K(h) = \frac{(\prod r_i!)}{\prod d_{ij}(h)!},$$

$$d_{ij}(h) = a_{ij} + h, \quad (i \neq j), \tag{11a}$$

$$d_{ii}(h) = r_i - \sum_{j \neq i} d_{ij}(h). \tag{11b}$$

The sum in (10) is over all h for which the quantities (11) are nonnegative. Holding $\{a_{ij}\}$ and $\{r_i\}$ constant also holds the column totals of the $L \times L$ table constant.

The probability (10) depends only on v. The parameter v has been separated from the $\{\alpha_{ij}\}$, and isolated in the probability function (10). Thus the probability function (10) provides information about v free from possible misinterpretations through ignorance of the other parameters $\{\alpha_{ij}\}$. Furthermore, (10) is exact. Its use does not require asymptotic normal approximations that depend for their validity on the cell frequencies being sufficiently large. Statistical tests of significance for specific values of $\bar{v}$, e.g. $\bar{v} = 0$, and, more importantly, confidence intervals for $\bar{v}$, can be obtained using (10). This is illustrated in Section 5.

4. The 2 × 2 table

For the 2×2 table, from (5b) and (4a)

$$\begin{aligned} \bar{v} = v = \alpha_{12} + \alpha_{21} &= \log\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right) \\ &= \log\left[\frac{p_{11}}{(1-p_{11})}\right] - \log\left[\frac{p_{21}}{(1-p_{21})}\right], \end{aligned}$$

which is the log odds ratio for the 2×2 table. The restrictions (11) produce the set of 2×2 tables $\{(r_1 - h, h); (a_{21} + h, r_2 - a_{21} - h)\}$, where the first ordered pair represents the first row of the table and the second, the second row. Setting $i = r_1 - h$ produces the equivalent set

$$\{(i, r_1 - i); (c_1 - i, r_2 - c_1 + i)\}.$$

The conditional probability $g(x_{12}; v)$ of (10) is thus equivalent to $\text{pr}(x_{11}; v \mid r_1, r_2, c_1, c_2)$, that is, the conditional probability of x_{11} given the marginal totals. This conditional probability is proportional to

$$\frac{\exp(x_{11}v)}{x_{11}!(r_1 - x_{11})!(c_1 - x_{11})!(r_2 - c_1 + x_{11})!}.$$

This result was obtained in [3], and forms the basis of the standard 'exact' analysis of the 2×2 table.

b25. Examples

In Example 1 the frequencies are very small, and so the exact procedure is required. In Example 2 the frequencies are large, and so simplifying approximations can be made.

Example 1. The data of Table 1 are taken from [2]. They are the frequencies arising from the classification of 46 trees and shrubs on two occasions one week apart by one observer according to four health categories.

Since $x_{12} = 0$, from (9) $\{a_{ij}\} = \{x_{ij}\}$. Only for $h = 0$ or 1 are the quantities (11) nonnegative. The value $h = x_{12} = 0$ yields the observed Table 1. The probabilities of $h = 0$, $h = 1$ can easily be calculated from (10) and (11) for any specified value of v.

Using $v = 0$ in (10), the probability of $h = 0$ is 1.74×10^{-6}. Therefore, either $v > 0$, that is, $\bar{v} > 0$, or the observed table $h = x_{12} = 0$ has probability less than 1.74×10^{-6}. Thus there is strong evidence against $\bar{v} = 0$.

Table 1
4×4 classification of trees and shrubs according to health by one observer on two occasions

		Occasion 2				
	Health Category	C_1	C_2	C_3	C_4	Total
Occasion 1	C_1	6	0	0	0	6
	C_2	1	4	1	0	6
	C_3	0	1	3	5	9
	C_4	0	0	4	21	25
	Total	7	5	8	26	46

Table 2
4 × 4 classification of trees and shrubs according to health by two observers

		Observer 2				
	Health Category	C_1	C_2	C_3	C_4	Total
Observer 1	C_1	239	18	9	11	277
	C_2	24	38	41	11	114
	C_3	15	49	113	94	271
	C_4	6	22	109	193	330
	Total	284	127	272	309	992

Using $v = 10.32$ in (10), the probability of $h = 0$ is 0.05, and of $h = 1$ is 0.95. Thus either $v > 10.32$ or an event of probability $P < 0.05$ has been observed. From (5a), $\bar{v} > 1.72$. This puts a lower 0.95 confidence bound on $\bar{v}$. At this level of confidence, the odds of the observer classifying a plant in a given category on Occasion 2 are on the average at least $\exp(1.72) = 5.6$ times greater if the plant was similarly classified by the observer on Occasion 1.

Example 2. The data in Table 2 are also taken from [2]. They are the frequencies arising from the classification of trees and shrubs by two different observers into the same health categories as in Example 1. From (9),

$$\{a_{ij}\} = \{-9, -7, 6, 23, -7, -3, 31, 76, -12, 4, 91\}.$$

There are nineteen values of h that make the quantities (11) nonnegative: $h = 12, 13, \ldots, 30$. Using (10), their respective probabilities can be obtained as in Example 1, although the calculations are much more extensive. For $v = 17.057$ the probability of the observed value $h = x_{12} = 18$ is 0.0238, and the probability $h \leq 18$ is 0.025. Thus a lower 0.975 confidence bound is $v = 17.057$. Similarly, when $v = 22.101$, the probability $h \geq 18$ is 0.025, giving an upper 0.975 confidence bound for v, and a 0.95 confidence interval $v = (17.1, 22.1)$. The corresponding confidence interval for $\bar{v}$ is (2.85, 3.68), or (17.3, 39.8) for $\exp \bar{v}$.

6. Maximum likelihood approximations

If the cell frequencies X are not too small, maximum likelihood (ML) approximations can be applied with the gain of considerable computational simplicity.

The likelihood function of $\{\alpha_{i1}, \ldots, \alpha_{iL}\}$ is proportional to the probability function P_i of (2a) expressed as a function of $\{\alpha_{i1}, \ldots, \alpha_{iL}\}$. From (2b) and (7) it can easily be seen that the logarithm of this is

$$\log P_i = -\sum_{j=1}^{L} \alpha_{ij} x_{ij} - r_i \log D_i + \log K_i. \tag{12}$$

The maximum likelihood estimate of α_{ij} is the value $\hat{\alpha}_{ij} = \log(x_{ii}/x_{ij})$ that maximizes (12). Thus from (6), the ML estimate of v is

$$\hat{v} = \sum \log \frac{x_{ii}}{x_{ij}} = \left(L \sum \log x_{ii}\right) - \sum\sum \log x_{ij}. \tag{13}$$

The application of ML estimation requires the calculation of the inverse of the matrix of negative second derivatives of (12) with respect to the $\{\alpha_{ij}\}$ evaluated at the ML estimate $\{\hat{\alpha}_{ij}\}$. A straightforward calculation shows these to be

$$\begin{aligned} I_i^{ij} &= \frac{1}{x_{ii}} + \frac{1}{x_{ij}}, \quad j \neq i, \\ I_i^{jk} &= \frac{1}{x_{ii}}, \qquad j \neq k, \quad j, k \neq i. \end{aligned} \tag{14}$$

The matrix of elements (14) can be loosely thought of as an estimate of the covariance matrix (cov $\hat{\alpha}_{ij}$, $\hat{\alpha}_{ik}$). In particular, they combine like variances and covariances.

Let

$$v_{i.} = \sum_j \alpha_{ij}, \quad i = 1, \ldots, L.$$

Since the quantities (14) combine like variances and covariances, the quantity I^{ii} applicable to $\hat{v}_{i.}$ is

$$\begin{aligned} I^{ii} &= \sum_j{}' I_i^{ij} + \sum_{j \neq k}{}' \sum{}' I_i^{jk} \\ &= \left\{\sum_j{}' \left[\frac{1}{x_{ii}} + \frac{1}{x_{ij}}\right]\right\} + \sum_{j \neq k}{}' \sum{}' \frac{1}{x_{ii}} \\ &= \left[\frac{L-1}{x_{ii}}\right] + \left[\sum_j{}' \frac{1}{x_{ij}}\right] + \frac{[(L-1)^2 - (L-1)]}{x_{ii}} \\ &= \left[\sum_j{}' \frac{1}{x_{ij}}\right] + \frac{(L-1)^2}{x_{ii}}, \end{aligned}$$

where $\sum'$ means the sum over j (and k) $\neq i$. Since from (5) $v = \sum v_{i.}$, and the rows i are statistically independent, the corresponding quantity for $\hat{v}$, denoted by I^v, is

$$\begin{aligned} I^v &= \sum I^{ii} = \left[\sum_i \sum_j{}' \frac{1}{x_{ij}}\right] + (L-1)^2 \sum_i \frac{1}{x_{ii}} \\ &= \left[\sum_i \sum_j \frac{1}{x_{ij}}\right] + L(L-2) \sum_i \frac{1}{x_{ii}}. \end{aligned} \tag{15}$$

From ML theory, the quantity

$$\mu(X, v) = \frac{(\hat{v} - v)}{(I^v)^{\frac{1}{2}}} \tag{16}$$

is an approximate standard normal (N(0, 1)) variate. Approximate confidence intervals and tests of significance for ν, and hence $\bar{\nu}$, can thereby be obtained.

Because of the marked discreteness of the probability function (10), a continuity correction will appreciably improve the accuracy of the confidence intervals obtained from (16). This correction is essentially the same as that required for the 2×2 table. It consists in replacing x_{12} by $x_{12} + 0.5$ and by $x_{12} - 0.5$ in calculating the lower and upper confidence bounds respectively. Because the analysis holds the marginal totals and the a_{ij}'s of (9) constant, this correction will result in adding and subtracting respectively 0.5 to all the off-diagonal cells, and subtracting and adding respectively $(L-1)/2$ to the diagonal cells.

This can be illustrated by Example 2. To obtain lower confidence bounds, add 0.5 to all the off-diagonal observations and subtract 1.5 from all the diagonal observations in Table 2. From (13) and (15) with $L = 4$, resulting ML estimate is $\hat{\nu} = 19.258$, and $I^{\nu} = 1.113$. Using (16) as an approximate N(0, 1) variate, the lower 0.975 confidence bound is $\nu = \hat{\nu} - 1.96(I^{\nu})^{\frac{1}{2}} = 17.19$, and $\bar{\nu} = 2.865$. To obtain upper confidence bounds, subtract 0.5 and add 1.5 to all off-diagonal and diagonal observations respectively. This results in $\hat{\nu} = 20.394$ and $I^{\nu} = 1.160$, so that $\nu = \hat{\nu} + 1.96(I^{\nu})^{\frac{1}{2}} = 22.50$, and $\bar{\nu} = 3.399$. This results in the approximate 0.95 confidence interval (2.865, 3.399) for $\bar{\nu}$. Using (10), the exact confidence level of this interval is 0.956. Hence this interval compares favourably with the exact interval $\bar{\nu} = (2.85, 3.68)$ obtained in Example 2.

The classifications C_i in Section 5 are in order of increasing health, and so are actually ordinal categories. They are treated in Section 5 as nominal categories for illustrative purposes, as was also done in [2]. Hoever, this raises the question, as suggested by a referee, whether these methods can be extended to cover ordinal categories. It would appear that regression techniques and weighted sums could replace (5c). This is an interesting area of future research.

Acknowledgements

The authors would like to thank V.T. Farewell for helpful comments on earlier drafts of the paper.

References

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

[2] J.N. Darroch and P.I. McCloud, Category distinguishability and observer agreement, Austral. Statist. 28 (1986) 371–388.

[3] R.A. Fisher, The logic of inductive inference (with discussion), J. Roy. Statist. Soc. 98 (1935) 39–54.

[4] D.A. Sprott and M.D. Vogel-Sprott, Use of the log odds ratio to assess the reliability of dichotomous questionnaire data, Appl. Psychological Measurement 11 (1987) 307–316.