

19338



Measuring Observer Agreement with Log Linear Models

Olga Towstopiat, Ph.D.

Institute for Program Evaluation

U.S. General Accounting Office

Washington, D.C. 20548

Paper presented at the American Psychological Association, Los Angeles,
California, August, 1981.

018381

Abstract

This paper presents a technology for hand calculating the degree of observer agreement using log linear models. A coefficient of agreement may be calculated which describes the magnitude of observer agreement as the estimated probability, under a quasi-independence model, that different observer responses will agree. Procedures and formulas for measuring the probability of agreement for one or more response categories within a larger set are also available. Finally, systematic disagreement among observers may also be examined with the available technology.

Observational studies of human behavior often require the recording of a number of behavioral categories. In addition, most observational studies require the assessment of agreement between observers. Measures such as the percentage of agreement, Cohen's kappa, and phi have been used to measure observer agreement, but these coefficients have limitations. Procedures that avoid the limitations of kappa and phi have been introduced in the literature (Bergan, 1980a). However, this technology, which utilizes log linear models, requires the use of a high speed digital computer. This limits the use of such procedures in applied behavioral research. The present paper introduces a new procedure for hand calculating the degree of observer agreement using log linear models. The application of log linear models for measuring observer agreement has the advantages of yielding a probability based coefficient of agreement with a directly interpretable meaning, correcting for the proportion of "chance" agreement, and providing an interpretable coefficient of "no agreement."

Although various log linear models may be applied for measuring observer agreement, this paper will focus on the use of the quasi-independence concept for assessing agreement. As Bergan (1980a) pointed out, the application of this procedure for measuring observer agreement has several advantages. Use of the quasi-independence concept yields a coefficient of observer agreement that varies between zero and one and measures agreement in terms of the probability that the observers' judgments will agree, as estimated under a quasi-independence model. This procedure may also be used to investigate if a single observational category or specific group of categories is a major contributor to the coefficient of agreement. Finally, systematic occurrences of disagreement between observers may be located and measured.

Assessment Procedures

To assess agreement, the judgments of the observers are organized into a contingency table. The quasi-independence model is recommended for measuring observer agreement when the table encompasses two or more observers recording three or more response categories. Quasi-independence among the variables comprising a contingency table is measured by testing the hypothesis that a subset of the contingency table cells are independent. Furthermore, by eliminating specific cells from the initial contingency table it is possible to segregate critical cells that account for association between the variables.

The procedure for measuring observer agreement requires the hand calculation of maximum likelihood estimates of expected cell frequencies, under the model of quasi-independence. To derive these expected cell frequencies the Deming-Stephens iterative fitting procedure must be applied. In the Deming-Stephens algorithm, preliminary estimates of the expected values are made, then successively adjusted until they meet the criterion that the row and column sums (i.e., marginals) for the estimated frequencies within the table equal the row and column sums for the observed values. Once the maximum likelihood expected cell frequencies are calculated, the chi-square statistic may be used to assess independence among the subset contingency table cells.

Since information regarding agreement by two observers is located within the diagonal cells in the contingency table, disagreement in the table may be assessed with a chi-square test of quasi-independence with diagonal cells deleted. By applying the chi-square test of independence, which measures agreement and disagreement, a baseline model can be formed. Statistical tests measuring the significance of the diagonal cells contribution to

agreement may be conducted by subtracting the chi-square values for assorted tests of quasi-independence from the chi-square value for the test of independence.

Following the attainment of a quasi-independence model that fits the observed contingency table frequencies, as indicated by a non-significant chi-square value, the magnitude of observer agreement under the model of quasi-independence may be hand calculated. More specifically, by using maximum likelihood probability estimates, that are based on the expected cell frequencies, the degree of agreement within the diagonal cells representing observer agreement and degree of observer disagreement within the off-diagonal cells may be assessed and examined.

The following sections will discuss in detail the procedures for testing the model-data fit, assessing the magnitude of observer agreement, investigating systematic observer disagreement, and hand calculating the expected cell frequencies with the Deming-Stephens iterative method. In addition, a quantitative example will be presented to facilitate the understanding of these procedures.

Independence and Quasi-Independence Models

The models of quasi-independence are best illustrated by associating them with the model of independence. If the judgments of two observers, A and B, are organized into a two-dimensional $I \times I$ contingency table, the rows in the table will represent the first observer's responses, 1 to I, and the columns will represent the second observer's responses, 1 to I. Cell frequencies within the contingency table are labeled with f's. For instance, f_{22} represents the frequency with which both observers coded the second response category. An inspection of the table will reveal that observer agreement frequencies are represented in the diagonal cells.

A test of independence of responses by two observers, depicted in a two-dimensional table, may be portrayed with the model $\pi_{ii} = \pi_i^A \times \pi_i^B$. The symbol π_{ii} represents the probability of occurrence of cell ii , π_i^A represents the probability of occurrence of variable A at level i and π_i^B represents the probability of occurrence of variable B at level i . The calculation of maximum likelihood estimates of expected cell frequencies for the test of independence are based on the aforementioned mathematical model. These estimates are computed by multiplying the cell probabilities by the total frequency of observations represented in the table (N). The model under investigation "fits" the data if the maximum likelihood estimates of expected cell frequencies conform closely to the observed cell frequencies. The likelihood-ratio statistic tests the fit of the data and model hypothesizing independence between observer responses.

Quasi-independence among variables comprising a contingency table is measured by testing the hypothesis that a subset of the contingency table cells are independent (Bishop, Fienberg, & Holland, 1975). By eliminating specific cells from the initial contingency table it is possible to segregate critical cells that account for association between the variables. The actual process of eliminating cells from the contingency table refers to placing structural zeros within the critical cells. Structural zeros are created by constraining expected cell frequencies to be equal to observed values. Setting estimates of expected frequencies equal to observed frequencies achieves this constraint and does not contribute to the value of the likelihood-ratio chi-square statistic. A demonstration of how structural zeros do not contribute to the chi-square value can be shown by applying the following likelihood-ratio statistic:

$$X^2_L = 2 \sum (\text{observed}) \log \frac{\text{Observed}}{\text{Expected}}$$

For example, if the diagonal cell's observed and expected values were set equal, the quantity for the portion of the formula $\log (\text{observed}/\text{expected})$ would be zero for all of the diagonal cells. Therefore, placing structural zeros in the diagonal cells would eliminate any contribution to the chi-square value by the diagonal cells.

To test the hypothesis of quasi-independence it is mandatory that an algorithm called iterative proportional fitting be used to estimate the maximum likelihood expected cell frequencies. This procedure, which will be discussed later in the paper, establishes preliminary estimates of the expected values and successively adjusts them until they meet the criterion that the marginal totals for the estimated frequencies is equal to the marginal totals for the observed values. The expected and observed marginal totals, in an incomplete table with structural zeros in the diagonal, will converge only if the following assumption is met:

$$X_{i+} + X_{+i} < N$$

where X_{i+} is the sum of the frequencies in non-structural-zero cells in row i , X_{+i} is the sum of the frequencies in non-structural-zero cells in column i , and N is the sum of the frequencies in all of the non-structural zero cells (Bishop, Feinberg, & Holland, 1975).

Once the maximum likelihood expected frequencies are calculated, the likelihood-ratio chi-square statistic may be used to assess independence among the non-structural zero cells. Degrees of freedom for the model of quasi-independence are determined by subtracting from the total number of contingency table cells the number of cells with structural zeros, one for the sample size constraint, and the number of independent parameters.

Assessing Agreement by Comparing Models of Independence and Quasi-Independence

Since information regarding agreement by two observers is located within the diagonal cells in the contingency table, disagreement in the table may be assessed with a chi-square test of quasi-independence with diagonal cells deleted (Bishop, Feinberg, & Holland, 1975). By applying the chi-square test of independence, which measures agreement and disagreement, a baseline model can be formed. Statistical tests measuring the significance of the diagonal cells contribution to agreement may be conducted by subtracting the chi-square values for assorted hierarchical tests of quasi-independence from the chi-square value for the test of independence. Goodman (1975) defined two models as hierarchically related if the subordinate model possessed all of the constraints of the superordinate model in addition to one or more further constraints. **For** instance, the model of independence is hierarchically related to a model of quasi-independence with the diagonal cells deleted. With hierarchical models the superordinate model implies the subordinate model. Therefore, the model of independence implies the model of quasi-independence. **If the model of** independence fits the data (i.e., has a statistically nonsignificant chi-square value), then the model of quasi-independence would also fit the data.

The advantage of the likelihood-ratio chi-square statistic lies in its ability to be partitioned exactly into independent component chi-squares and summed to achieve the overall contingency table chi-square and degrees of freedom (Cochran, 1954). This property allows the independence model and chi-square to be partitioned into component chi-squares such as the chi-square for the test of quasi-independence and the chi-square indicating the difference between the independence and quasi-independence values. **Subtracting** the chi-square value and related degrees of freedom for a test of quasi-independence

with all the diagonal cells eliminated from the chi-square value and related degrees of freedom for the test of independence would provide a chi-square value that would measure if the diagonal cells provide a significant contribution to the association in the contingency table.

By applying the aforementioned model comparison procedures, the specific chi-square contribution of a single agreement cell or subset of agreement cells can be assessed. An experimenter wishing to investigate the contribution of each agreement category represented in the diagonals of a 3 x 3 table may accomplish this by using several different quasi-independence models and compare them with the independence model. For example, the investigator could set up three quasi-independence models each ruling out one of the diagonal cells f_{11} , f_{22} , and f_{33} , respectively. Each of the chi-square values for these independence models could be subtracted from the chi-square value for the independence model to test if the specific cell provided a significant contribution to model-data fit. If a model of quasi-independence ruled out a single cell such as f_{11} , and the difference between the chi-square values for the quasi-independence and independence models had a value of 3.84 (critical value for 1 degree of freedom) or larger then the contribution of that cell to agreement would be statistically significant. If the difference chi-square value was less than 3.84 then the investigator could not conclude that the observers' judgments agreed for the first behavioral category, regardless of the number of agreement frequencies in the f_{11} cell.

Investigators may also find that an off-diagonal disagreement cell provides a significant contribution to the overall chi-square value. A case of systematic disagreement between observers may occur if observer A codes a specific behavior in the first category and observer B codes that behavior

in the second category. To test if a significant association between the observers' responses exists, a quasi-independence model may be developed which places structural zeros in the hypothesized cell or cells denoting systematic disagreement. The chi-square value for the quasi-independence model is subtracted from the chi-square value for the independence model to test the statistical significance of the association.

Estimating the Magnitude of Agreement under the Model of Quasi-Independence

A quasi-independence model and the maximum likelihood estimates of probabilities for agreement and disagreement may be used for the computation of the degree of agreement between observers. Goodman's (1975) work with response scaling and Bergan (1980a) have demonstrated that from models of quasi-independence, with structural zeros in the cells representing agreement, maximum likelihood probability estimates may be calculated for the agreement cells. In addition, the off-diagonal or disagreement cells may also have a probability estimate computed. The precision of the probability estimates is based on the model fitting the data. Therefore, a chi-square value for a quasi-independence model must be statistically non-significant to indicate an appropriate model-data fit. Given a 3 x 3 table with observations for three behavioral categories, the maximum likelihood estimates representing agreement and disagreement would be expressed in four classifications. The first three classifications would represent each of the diagonal cells, respectively. The fourth classification would represent the six cumulative off-diagonal disagreement cells.

Goodman (1975) assumed that the off-diagonal observer disagreement responses (i.e., cells without structural zeros) were independent. He also assumed the expected response pattern for each diagonal cell with a structural

zero had a probability of 1. Given these assumptions, the following formula computes the maximum likelihood estimate for the probability that observers' A and B responses would be represented within the disagreement category:

$$\hat{\pi}_0 = \hat{\pi}^{AB}_{ij} / \hat{\pi}^{\bar{A}}_{i0} \hat{\pi}^{\bar{B}}_{j0} \quad (1)$$

where $\hat{\pi}_0$ is the estimated probability of disagreement between observers, $\hat{\pi}^{AB}_{ij}$ is the estimated probability of a disagreement response F_{ij} ($i \neq j$) for both observers; $\hat{\pi}^A_{i0}$ is the conditional probability of observer A emitting response i , assuming the observers' ij response pattern denotes a disagreement between the observers; and $\hat{\pi}^B_{j0}$ is the conditional probability of observer B emitting response j . The probability of a specific agreement category \underline{t} is expressed with the following maximum likelihood estimate:

$$\hat{\pi}_t = p_{ij} - \hat{\pi}_0 \hat{\pi}^A_{i0} \hat{\pi}^B_{j0}$$

where p_{ij} is the observed proportion of a specific observer agreement category \underline{t} as designated in the ij cell, $\hat{\pi}^A_{i0}$ is the maximum likelihood estimate of observer A's response i given the disagreement category 0 and $\hat{\pi}^B_{j0}$ is the maximum likelihood estimate of observer B's response j given the disagreement category.

In formula (1) the $\hat{\pi}^{AB}_{ij}$ value is calculated by dividing the response pattern ij expected cell frequency (\hat{F}_{ij}) by the total number of observer responses (N). The computation of the $\hat{\pi}^A_{i0}$ and $\hat{\pi}^B_{j0}$ values require the expected cell frequencies and use of the following formula for polytomous variables:

$$\hat{\pi}^{\bar{A}}_{i0} = \Omega^A_i / i'0 \left(\sum_{i=1}^I \Omega^A_i / i'0 \right)$$

where $\Omega_i^A / i'o$ represents the odds of an i disagreement response to an i' disagreement response, by observer A. These odds are obtained from the following estimated expected cell frequencies:

$$\Omega_i^A / i'o = \hat{F}_{ij} / \hat{F}_{ij'}$$

where ij and ij' are disagreement response patterns.

Goodman's work with response scaling and models of quasi-independence demonstrated that the probability for the agreement (i.e., structural zero cells) and disagreement (i.e., non-structural zero cells) categories add to one. Therefore, the estimated proportion of the sum of the agreement cells equals one minus the probability for the disagreement cells. By placing structural zeros in the agreement/diagonal cells within a contingency table signifying the response distribution of two observers, an index of the magnitude of observer agreement can be developed. The following formula connotes the magnitude of observer agreement as the estimated probability that judgments from two observers will occur in one of the agreement categories ($\hat{\pi}_A$):

$$\hat{\pi}_A = 1 - \hat{\pi}_0$$

where $\hat{\pi}_0$ is the estimated probability that a pair of judgments from the observers will occur in the disagreement category.

Iterative Computations of Expected Frequencies

The estimated expected frequencies under the model of quasi-independence must be computed by an algorithm called iterative proportional fitting. In the Deming-Stephens (Feinberg, 1978) algorithm, preliminary estimates of the expected values are made, then successively adjusted until they meet the criterion that the marginal totals for the estimated frequencies equal the

marginal totals for the observed values. Therefore:

$$\hat{F}_{i+} = f_{i+} \text{ and } \hat{F}_{+j} = f_{+j}$$

for all i and j in the subset of off-diagonal cells. Since the diagonal cells contain structural zeros under the agreement model of quasi-independence, the frequency summations are only across non-structural-zero cells.

Let \hat{F}_{ij} equal the expected frequency of the (i,j) th cell, with X_{ij} equal to the observed frequency. Let us also assume that X_{i+} or \hat{F}_{i+} , etc....., refer to the summation across only non-structural-zero cells. The initial start values within the table are denoted as $\hat{F}_{ij}^{(0)}$ and the subsequent K th iteration as $\hat{F}_{ij}^{(K)}$.

To hasten the iterative process, rather than insert values of one within the table for start values it is recommended that a proportion of the cell frequencies be estimated from the marginal values and used as the start values ($\hat{F}_{ij}^{(0)}$). The procedure sequentially fixes one set of marginal values and allows the other set of marginals to vary. To calculate the proportional start values, consider a 3 x 3 table with the diagonal cells deleted. Begin by fixing the column marginals and allowing the row marginals to vary. Cell $\hat{F}_{21}^{(0)}$ is estimated with specific marginal frequencies from the original table:

$$\hat{F}_{21}^{(0)} = X_{+1} \left(\frac{X_{2+}}{X_{2+} + X_{3+}} \right)$$

Similarly cells $\hat{F}_{31}^{(0)}$ and $\hat{F}_{13}^{(0)}$ are estimated by:

$$\hat{F}_{31}^{(0)} = X_{+1} \left(\frac{X_{3+}}{X_{2+} + X_{3+}} \right)$$

$$\hat{F}_{13}^{(0)} = X_{+2} \left(\frac{X_{1+}}{X_{1+} + X_{3+}} \right)$$

Once the start values are calculated, the algorithm proceeds in a two-step manner, with:

$$\hat{F}_{ij}^{(K+1)} = \frac{\hat{F}_{ij}^{(K)} X_{i+}}{\hat{F}_{i+}^{(K)}}$$

and

$$\hat{F}_{ij}^{(k+2)} = \frac{\hat{F}_{ij}^{(K+1)} X_{+j}}{\hat{F}_{+j}^{(K+1)}}$$

The procedure alternates between fixing the row and column marginals. During the iterative process the improvement in the subsequent fits come from the estimated marginals getting nearer to the observed marginals. Following several large estimation leaps initially, the convergence process slows down with smaller estimation changes. Thus, once the investigator gets past the initial estimation leaps, a "reasonable" approximation may be expected without considerable iteration.

Summary

The present paper has briefly described how observational studies that assess agreement between observers may establish the reliability of the observations with a greater degree of accuracy. By applying log linear models, such as the model of quasi-independence, for the purpose of measuring observer agreement the investigator may hand calculate a probability-based coefficient of agreement with a directly interpretable meaning, correct for the proportion of "chance" agreement, and calculate a meaningful coefficient of "no agreement."

REFERENCES

- Bergan, J. Measuring observer agreement using the quasi-independence concept. Journal of Educational Measurement, 1980a, 17, 59-68.
- Bishop, Y., Fienberg, S., and P. Holland. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press, 1975.
- Cochran, W. Some methods for strengthening the common X^2 tests, Biometrics, 1954, 10, 417-451.
- Fienberg, S. The Analysis of Cross-Classified Categorical Data. Cambridge, Mass.: MIT Press, 1978.
- Goodman, L. A new model for scaling response patterns: An application of the quasi-independence concept. Journal of the American Statistical Association, 1975, 70, 755-768.

Full table

	1	2	3	
1	7	1	2	10
2	3	8	4	15
3	5	6	9	20
	15	15	15	45 Total

Quasi-Independence

		1	2	3
3			4	7
5		6		11
	8	7	6	21 Total

	1.500	1.80	3.3
3.111		4.2	7.311
4.889	5.500		10.389
	8	7	6

	1.364	1.636	3
2.979		4.021	7
5.177	5.823		11
	8.156	7.187	5.657

	1.329	1.735	3.064
2.922		4.265	7.187
5.078	5.671		10.749
	8	7	6

	1.301	1.699	3
2.846		4.154	7
5.197	5.803		11
	8.043	7.104	5.853

	1.282	1.742	3.024
2.831		4.258	7.089
5.169	5.718		10.887
	8	7	6

	1.272	1.728	3
2.795		4.205	7
5.223	5.777		11
	8.043	7.104	5.853

	1.263	1.748	3.011
2.789		4.252	7.041
5.211	5.737		10.948
8	7	6	

	1.258	1.742	3
2.773		4.227	7
5.236	5.764		11
8.009	7.022	5.969	

$$\Omega_{\frac{1}{10}}^A = \frac{1.258}{1.258} = 1.0$$

$$\Omega_{\frac{2}{10}}^A = \frac{4.227}{1.742} = 2.4265$$

$$\Omega_{\frac{3}{10}}^A = \frac{5.764}{1.258} = 4.5819$$

$$\Omega_{\frac{1}{10}}^B = 1.0$$

$$\Omega_{\frac{2}{10}}^B = 1.1008$$

$$\Omega_{\frac{3}{10}}^B = 1.5243$$

$$\textcircled{\text{II}} \pi_{10}^A = \frac{1}{1+2.43+4.58} = .1249$$

$$\pi_{20}^A = \frac{2.4265}{8.0084} = .3030$$

$$\pi_{30}^A = \frac{4.5819}{8.0084} = .5721$$

$$\pi_{10}^B = .2759$$

$$\pi_{20}^B = .3037$$

$$\pi_{30}^B = .4205$$

$$\textcircled{\text{IV}} \pi_1 = .15556 - .737(-.1249)(.2759) = .1302$$

$$\pi_2 = .1099$$

$$\pi_3 = .0227$$

$$\pi_0^A = \frac{\pi_{12}^{AB}}{\pi_{10}^A} \pi_{20}^B = \frac{1.258/45}{.1249(.3037)} = .737$$