ORDER STATISTICS, CONCOMITANTS OF *See* MULTIVARIATE ORDER STATISTICS

ORDINAL DATA

An ordinal variable is one that has a natural ordering of its possible values, but for which the distances between the values are undefined. Ordinal variables usually have categorical scales. Examples are social class, which is often measured as upper, middle, or lower, and political philosophy, which might be measured as liberal, moderate, or conservative. Continuous variables that are measured using ranks are also treated as ordinal.

In this article we describe methods for analyzing only ordinal categorical variables. In particular, we summarize some association measures and models that are appropriate for the analysis of contingency tables* having at least one ordered classification. Methods for analyzing continuous observation variables are summarized in DISTRIBUTION-FREE STATISTICS and RANK TESTS.

ORDINAL MEASURES OF ASSOCIATION

We present three types of ordinal measures of association: measures based on the notions of concordance* and discordance, which utilize ordinal information only; correlation* and mean measures that require a user-supplied or data-generated scoring of ordered categories; sets of odds-ratio measures that contain as much information regarding association as the original cell counts. For a discussion of the rationale of measures of association, *see* ASSOCIATION, MEASURES OF.

Concordance–Discordance Measures

We discuss most of the methodology in this article in the context of a two-way contingency table*. Denote the cell counts of an $r \times c$ table by $\{n_{ij}\}$ and let $\{p_{ij} = n_{ij}/n\}$ be the corresponding cell proportions. Let X denote the row variable and Y the column variable. For now, we suppose that the rows and columns are both ordered, with the first row and first column being the low ends of the two scales.

A pair of observations is *concordant* if the member that ranks higher on X also ranks higher on Y. A pair of observations is *discordant* if the member that ranks higher on X ranks lower on Y. The numbers of concordant and discordant pairs are

$$C = \sum_{i'>i} \sum_{j'>j} n_{ij} n_{i'j'} \quad \text{and}$$

$$D = \sum_{i'<i} \sum_{j'>j} n_{ij} n_{i'j'} .$$

Let $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$. We can express the total number of pairs of observations as

$$n(n-1)/2 = C + D + T_X + T_Y - T_{XY},$$

where $T_X = \sum_i n_{i+}(n_{i+} - 1)/2$ is the number of pairs tied on X, $T_Y = \sum_j n_{+j}(n_{+j} - 1)/2$ is the number of pairs tied on Y, and $T_{XY} = \sum n_{ij}(n_{ij} - 1)/2$ is the number of pairs from a common cell (tied on X and Y).

Several measures of association* are based on the difference $C - D$. They are discrete generalizations of the Kendall's tau* measure for continuous variables. For each, the greater the relative number of concordant pairs, the more evidence there is of a positive association.

Of the untied pairs, $C/(C+D)$ is the proportion of concordant pairs and $D/(C+D)$ is the proportion of discordant pairs. The measure *gamma*, proposed by Goodman and Kruskal [9], is the difference between these proportions,

$$\hat{\gamma} = (C - D)/(C + D).$$

For 2×2 tables $\hat{\gamma}$ is also referred to as Yule's Q. In 1945, Kendall [14] proposed the related measure *tau-b* given by

$$\hat{\tau}_b = \frac{C - D}{[\{\frac{1}{2}n(n-1) - T_X\}\{\frac{1}{2}n(n-1) - T_Y\}]^{1/2}} .$$

For 2×2 tables $\hat{\tau}_b$ simplifies to the Pearson correlation obtained by assigning any scores to the rows and to the columns that reflect their orderings.

Gamma and tau-b assume the same values regardless of whether X or Y (or neither) is regarded as a response variable. In 1962, Somers proposed the asymmetric measure

$$d_{YX} = (C - D)/[n(n-1)/2 - T_X],$$

the difference between the proportions of concordant and discordant pairs, out of those pairs that are untied on X. For 2×2 tables d_{YX} simplifies to the difference of proportions $n_{11}/n_{1+} - n_{21}/n_{2+}$. For $2 \times c$ tables, it estimates $P(Y_2 > Y_1) - P(Y_1 > Y_2)$, where Y_1 and Y_2 are independent observations on the column variable in rows one and two of the tables, respectively.

All three of these ordinal measures are restricted to the range $[-1, +1]$. Independence implies that their population values equal zero, but the converse is not true. Note that $|\hat{\tau}_b| \leqslant |\hat{\gamma}|$ and $|d_{YX}| \leqslant |\hat{\gamma}|$, and $\hat{\tau}_b^2 = d_{YX}d_{XY}$, where d_{XY} has T_Y instead of T_X in its denominator. Tau-b may be interpreted as a Pearson correlation and Somers' d may be interpreted as a least-squares* slope for a linear regression* model defined using sign scores for pairs of observations.

Measures Based on Scores

Many methods for analyzing ordinal data require assigning scores to the levels of ordinal variables. To compute the Pearson correlation* between the row and column variables, e.g., one must assign fixed scores to the rows and to the columns. The canonical correlation is the maximum correlation obtained out of all possible choices of scores. The scores needed to achieve the maximum need not be monotone, however. Alternatively, one can generate monotone scores from the data. For example, one could use average cumulative probability scores, which for the column (Y) marginal distribution are

$$r_j = \sum_{i=1}^{j-1} p_{+i} + p_{+j}/2, \qquad j = 1, \dots, c.$$

The correlation measure then obtained is a discrete analog of Spearman's rank correlation coefficient, referred to as $\hat{\rho}_b$ (see Kendall [14, p. 38]). Or one could use scores at which a distribution function (such as the normal or logistic) takes on the $\{r_j\}$ values.

If X is nominal (unordered levels) and Y is ordinal, often it is useful to compute a mean score on Y within each level of X. The scores $\{r_j\}$ just defined are called *ridits* for the marginal distribution of Y. The measure $\bar{R}_i = \sum_j r_j(n_{ij}/n_{i+})$ is the sample mean ridit in row i. It estimates

$$P(Y_i > Y^*) + \tfrac{1}{2}P(Y_i = Y^*),$$

where Y_i and Y^* are categories of Y for observations randomly selected from row i and from the marginal distribution of Y, respectively. It is necessary that $\sum_j p_{+j} r_j = \sum_i p_{i+} \bar{R}_i = 0.5$. See Bross [4] for a discussion of ridit analysis*. For an example of a scaling method that assumes a particular form for an underlying continuous distribution, see Snell [21].

Odds-Ratio* Measures

The measures discussed summarize association by a single number. To avoid the loss of information we get by this condensation, we can describe the table through a set of $(r-1)(c-1)$ odds ratios. For ordinal variables, it is natural to form the *local odds ratios*

$$\hat{\theta}_{ij} = n_{ij}n_{i+j,j+1}/(n_{i,j+1}n_{i+1,j}),$$
$$i = 1, \dots, r-1, \quad j = 1, \dots, c-1.$$

Each $\hat{\theta}_{ij}$ describes the sample association in a restricted region of the table, with $\log \hat{\theta}_{ij}$ indicating whether the association is positive or negative in that region. Goodman [7] suggested log-linear models for analyzing the $\{\hat{\theta}_{ij}\}$. An alternative set of odds ratios is based on the $(r-1)(c-1)$ ways of collapsing the table into a 2×2 table.

ORDINAL MODELS

In recent years, much work has been devoted to formulating models for cross-classifications of ordinal variables. The models discussed here are directly related to standard log-linear and logit models (*see*

CONTINGENCY TABLES and MULTIDIMENSIONAL CONTINGENCY TABLES).

Log-linear Models

Suppose that $\{\rho_{ij}\}$ denotes the true cell proportions in an $r \times c$ contingency table, where $\sum \rho_{ij} = 1$. For a random sample of size n, the expected number of observations in a cell is $m_{ij} = n\rho_{ij}$. If the variables are independent, then $m_{ij} = m_{i+}m_{+j}$ for all i and j. There is a corresponding additive relationship for $\log m_{ij}$. That is, we can describe independence by the log-linear model $\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y$, where μ is the mean of the $\{\log m_{ij}\}$ and $\sum \lambda_i^X = \sum \lambda_j^Y = 0$. Haberman [12], Simon [20], and Goodman [7] have formulated more complex log-linear models for situations where at least one variable is ordinal and there is some association.

The log-linear models can be described in terms of properties of the local odds ratios $\{\theta_{ij} = (m_{ij}m_{i+1,j+1})/(m_{i,j+1}m_{i+1,j})\}$. A simple model has the form $\log \theta_{ij} = \beta$ for all i and j, whereby the local association is uniform throughout the table. A more general model is obtained by assigning monotone scores $\{u_i\}$ to the rows and $\{v_j\}$ to the columns and assuming that

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j .$$

In this model

$$\log \theta_{ij} = \beta(u_{i+1} - u_i)(v_{j+1} - v_j).$$

When the $\{u_i\}$ are equally spaced and the $\{v_j\}$ are equally spaced, we obtain the uniform association model. When $\beta = 0$, we obtain the independence model. The goodness of fit of the uniform association model can be tested with a chi-squared statistic* having $rc - r - c$ degrees of freedom.

Goodman [7] discussed several other models that include the uniform association model as a special case. A *row effects* model has the property $\log \theta_{ij} = \alpha_i$ for all i and j. The row variable may be nominal for this model, which can be tested with a chi-squared statistic having $(r-1)(c-2)$ degrees of freedom. This model is itself a special case of two *row and column effects* models, one for which $\log \theta_{ij} = \alpha_i + \beta_j$ and the other of which has the multiplicative form $\log \theta_{ij} = \alpha_i \beta_j$. These models have $(r-2)(c-2)$ residual degrees of freedom. Analogous models can be formulated for multidimensional tables. See Clogg [5] for details.

These log-linear models treat the variables alike in the sense that no variable is identified as a response. Iterative methods are necessary to obtain maximum likelihood estimates of parameters and goodness-of-fit statistics for these models. See the sections on Estimation and Computer Packages.

Logit Models*

Suppose now that an ordinal variable Y is a response variable and let $\mathbf{X}$ denote explanatory variables. Let $\rho_i(\mathbf{x})$ denote the probability that Y falls in category i when $\mathbf{X} = \mathbf{x}$, where $\sum_{i=1}^{c} \rho_i(\mathbf{x}) = 1$. When $c = 2$, the *logit* transformation is $\log[\rho_2(\mathbf{x})/\rho_1(\mathbf{x})]$. The linear logit regression model

$$\log\left[\rho_2(\mathbf{x})/\rho_1(\mathbf{x})\right] = \alpha + \boldsymbol{\beta}'\mathbf{x}$$

is one that yields predicted values of $\rho_i(\mathbf{x})$ between 0 and 1, the relationship being S-shaped between $\rho_i(\mathbf{x})$ and each x_i (*see* LOGIT).

When there are $c > 2$ responses, there are several ways of forming logits that take the ordering of the categories into account. The *cumulative logits*

$$L_j = \log\left[\sum_{i>j} \rho_i(\mathbf{x}) \Big/ \sum_{i \leqslant j} \rho_i(\mathbf{x})\right], \qquad j = 1, \ldots, c-1,$$

are logits of distribution function values and lend themselves nicely to interpretation. Williams and Grizzle [22] and McCullagh [17] have suggested models for them.

We illustrate with a logit model for a two-way table having column variable Y as a response. The jth cumulative logit in row i is

$$L_{ij} = \log\left(\frac{\rho_{i,j+1} + \cdots + \rho_{ic}}{\rho_{i1} + \cdots + \rho_{ij}}\right),$$

$i = 1, \ldots, r, j = 1, \ldots, c-1$. Suppose that X is also ordinal and that we assign scores

$\{u_i\}$ to its levels. A simple linear model is

$$L_{ij} = \alpha_j + \beta u_i,$$
$$i = 1, \ldots, r, \quad j = 1, \ldots, c - 1.$$

This model implies that the effect β of X on the logit for Y is the same for all cut points $j = 1, \ldots, c - 1$ for forming the logit. For the integer scores $\{u_i = i\}$, $L_{i+1,j} - L_{ij} = \beta$ for all i, j. Thus this logit model can also be regarded as a type of uniform association model. In this case, β is a log odds ratio formed using adjacent rows when the response is collapsed into two categories. Like the log-linear uniform association model, it has $rc - r - c$ residual degrees of freedom for testing goodness of fit*.

Logit models for multidimensional tables can be constructed like multiple regression models by including terms for qualitative and quantitative explanatory variables. Iterative methods are needed for maximum likelihood estimation of the models, as described in the sections on Estimation and Computer Packages.

Models for Square Tables

In some applications, each classification in a table has the same categories. This happens, for example, for matched-pairs data such as occur in social mobility tables. Cell probabilities in square tables often exhibit a type of symmetry relative to the main diagonal. Also, when the categories are ordered, it is often of interest to study whether one marginal distribution tends to have larger responses, in some sense, than the other.

An example of the type of model that has been proposed for $r \times r$ ordinal tables is Goodman's [8] diagonals-parameter symmetry model,

$$m_{ij} = m_{ji}\delta_{j-i}, \qquad i < j.$$

The parameter δ_k, $k = 1, \ldots, r - 1$ is the odds that an observation falls in a cell k diagonals above the main one instead of in a corresponding cell k diagonals below the main one. For the special case $\delta_1 = \cdots = \delta_{r-1} = \delta$, this model exhibits the conditional symmetry $P(X = i, \ Y = j \mid X < Y) = P(X = j, Y = i \mid X > Y)$. The further special case, in which all $\delta_k = 1$, gives the symmetry model $m_{ij} = m_{ji}$, $i \neq j$. Each of these models can be expressed as a log-linear model and tested using standard chi-squared statistics. Whether the delta parameters in these models exceed one or are less than one determine how the marginal distributions are stochastically ordered.

There are several other log-linear models in which the effect of a cell on the association depends on its distance from the main diagonal. Also, standard log-linear models for ordinal variables (e.g., uniform association model) often fit square tables well when the main diagonal is deleted. See Haberman [13, pp. 500–503] and Goodman [8] for examples; *see also* MARGINAL SYMMETRY.

Other Models

Several alternative ways have been proposed for modeling ordinal variables. Some of these assume an underlying continuous distribution of a certain form. McCullagh [17] discussed a "proportional hazards" model that utilizes the log$(-\log)$ transformation of the complement of the distribution function of the response variable. He argued that it would be appropriate for underlying distributions of the types used in survival analysis.

If one feels justified in assigning scores to the levels of an ordinal response variable, then one can construct simple models for the mean response that are similar to analysis of variance and regression models for continuous variables. This approach is especially appealing if the categorical nature of the ordinal response is due to crude measurement of an inherently continuous variable. Grizzle et al. [11] gave a general weighted least-squares approach for fitting models of this type. Similar models have been constructed for mean ridits* (see Semenya and Koch [19]).

INFERENCE FOR ORDINAL VARIABLES

In this section we discuss estimation of ordinal measures of association and models and describe ways of using the estimates to test certain basic hypotheses. We assume that the sample was obtained by full multinomial sampling or else by independent multinomial sampling within combinations of levels of explanatory variables.

Estimation

Under these sampling models, the measures of association discussed in the first section are asymptotically normally distributed. Goodman and Kruskal [10] applied the delta method (*see* STATISTICAL DIFFERENTIALS) to obtain approximate standard errors for these measures. Hence one can form confidence intervals for them.

The ordinal log-linear and logit models can be fit using weighted least squares* (WLS) or maximum likelihood* (ML). The WLS estimate has a simple closed-form expression. See Williams and Grizzle [22], e.g., for WLS estimation of the cumulative logit model.

The ordinal log-linear models discussed in the Log-linear Models section are special cases of generalized linear models* proposed by Nelder and Wedderburn [18]. The ML estimates may be obtained using the iterative Newton–Raphson method described in their paper, which corresponds to an iterative use of WLS. ML estimates can also be obtained using an iterative scaling approach given by Darroch and Ratcliff [6] or by using a Newton unidimensional iterative procedure suggested by Goodman [7]. The latter approaches are simpler than Newton–Raphson*, but convergence is much slower. McCullagh [17] showed how to use the Newton–Raphson method to obtain ML estimates for a class of models that includes the cumulative logit models.

Testing Hypotheses

Basic hypotheses concerning independence, conditional independence, and higher-order interactions can be tested using estimates of measures of association or estimates of certain model parameters. For example, consider the null hypothesis of independence for the ordinal–ordinal table. Goodman and Kruskal [10] showed that a broad class of measures of association have asymptotic normal distributions for multinomial sampling. In particular, an ordinal measure such as gamma or tau-b divided by its standard error has an asymptotic standard normal null distribution. This statistic will (as $n \to \infty$) detect associations where the true value of the measure is nonzero. If the logit or log-linear uniform association model holds, then independence is equivalent to $\beta = 0$. The estimate of β divided by its standard error also has an asymptotic standard normal null distribution. Alternatively, the difference in values of the likelihood-ratio statistics for testing goodness of fit of the independence model and the uniform association model has an asymptotic chi-squared distribution* with a single degree of freedom.

Similar remarks apply to the two-way table with r unordered rows and c ordered columns. Independence can be tested using a discrete version of the Kruskal–Wallis test, which detects differences in true mean ridits. If log-linear or logit row effects models fit the data, it can also be tested using the difference in likelihood-ratio statistics between the independence model and the row effects model. Each of the approaches gives a statistic that has an asymptotic null chi-squared distribution with $r-1$ degrees of freedom. Analogous tests can be formulated for multidimensional tables.

Computer Packages

Several computer packages can be used for the computational aspects of analyzing ordinal data. Some of these are large, general-purpose statistical packages that have components or options for categorical data*. For example, the widely available package BMDP has a program (4F) that, among other things, computes several measures of association and their asymptotic standard errors. The package GLIM* is particularly useful for fitting log-linear models, including the ordinal ones mentioned in the Log-linear Models section. Other programs have been designed specifically for categorical data* and can be used for certain ordinal methods. These include FREQ [13] for ML estimation of log-linear models, MULTIQUAL [3] for ML fitting of log-linear and logit models, and GENCAT [16], which can be used to fit a large variety of models using WLS (see also the FUNCAT program in the SAS package). (*See also* STATISTICAL SOFTWARE.)

Summary

More detailed surveys of methods for analyzing ordinal data are given by Semenya and Koch [19] and by Agresti [2]. Ordinal measures of association have been surveyed by Goodman and Kruskal [9, 10], Kruskal [15], and Kendall [14]. Summary discussions of methods for modeling ordinal variables were presented by Goodman [7], McCullagh [17], Clogg [5], and Agresti [1].

References

[1] Agresti, A. (1984). *J. Amer. Statist. Ass.*, **78**, 184–198.

[2] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley-Interscience, New York.

[3] Bock, R. D. and Yates, G. (1973). *Log-linear Analysis of Nominal or Ordinal Qualitative Data by the Method of Maximum Likelihood*. National Education Resources, Chicago.

[4] Bross, I. D. J. (1958). *Biometrics*, **14**, 18–38.

[5] Clogg, C. (1982). *J. Amer. Statist. Ass.*, **77**, 803–815.

[6] Darroch, J. N. and Ratcliff, D. (1972). *Ann. Math. Statist.*, **43**, 1470–1480.

[7] Goodman, L. A. (1979). *J. Amer. Statist. Ass.*, **74**, 537–552. (An easy-to-read development of log-linear models based on local odds ratios.)

[8] Goodman, L. A. (1979). *Biometrika*, **66**, 413–418.

[9] Goodman, L. A., and Kruskal, H. (1954). *J. Amer. Statist. Ass.*, **49**, 723–764. (A classic paper on measures of association for ordinal and nominal variables.)

[10] Goodman, L. A. and Kruskal, W. H. (1972). *J. Amer. Statist. Ass.*, **67**, 415–421.

[11] Grizzle, J. E., Starmer, C. F., and Koch, G. G., (1969). *Biometrics*, **25**, 489–504. (A good exposition of the use of weighted least squares for fitting a wide variety of models to categorical data.)

[12] Haberman, S. J. (1974). *Biometrics*, **30**, 589–600.

[13] Haberman, S. J. (1979). *Analysis of Qualitative Data*, Vol. 2: *New Developments*. Academic Press, New York. (One of the few categorical data books that devotes much space to models for ordinal variables, but not easy reading.)

[14] Kendall, M. G. (1970). *Rank Correlation Methods*, 4th ed. Charles Griffin, London.

[15] Kruskal, W. H. (1958). *J. Amer. Statist. Ass.*, **53**, 814–861.

[16] Landis, J. R., Stanish, W. M., Freeman, J. L., and Koch, G. G. (1976). *Computer Programs Biomed.*, **6**, 196–231.

[17] McCullagh, P. (1980). *J. R. Statist. Soc. B*, **42**, 109–42. (Discusses important issues to be considered in modeling ordinal response variables.)

[18] Nelder, J. A. and Wedderburn, R. W. M. (1972). *J. R. Statist. Soc. A*, **135**, 370–384.

[19] Semenya, K. and Koch, G. G. (1980). *Institute of Statistics Mimeo Series No. 1323*, University of North Carolina, Chapel Hill, NC. (A good survey of the use of weighted least squares for fitting various models to ordinal data.)

[20] Simon, G. (1974). *J. Amer. Statist. Ass.*, **69**, 971–976.

[21] Snell, E. J. (1964). *Biometrics*, **20**, 592–607.

[22] Williams, O. D. and Grizzle, J. E. (1972). *J. Amer. Statist. Ass.*, **67**, 55–63.

(ASSOCIATION, MEASURES OF
CONTINGENCY TABLES
GOODMAN–KRUSKAL TAU
AND GAMMA
LOGIT
NOMINAL DATA
ODDS-RATIO ESTIMATORS
RANKING PROCEDURES
RANK TESTS
SCALE TESTS)

ALAN AGRESTI

ORDINARY LEAST SQUARES (OLS)

See LEAST SQUARES

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)

The OECD is the Paris-based international organization of the industrialized, market-economy countries. Its membership includes the countries of Western Europe, Canada and the United States, Japan, Australia, and