Analysis of Ordinal Contingency Tables

When variables have few categories that can be rank ordered in value, there is usually an advantage to using the ordinal nature of the scale in the analysis as compared with loglinear models, Pearson χ^2, or the likelihood ratio test for nominal categories. Examples might include variables with responses such as, “no,” “maybe,” and “yes”; “never,” “sometimes,” and “always”; “strongly disagree,” “disagree,” “strongly disagree.” For binary variables, there is really no advantage to thinking about variables as ordinal. Although we could interpret a “yes”-“no” response to an attitude question as having higher values (more favorability) if the response is “yes,” there is not any meaningful difference from the analyses we have already used which compare these two values as nominally different. At the other end of the spectrum, there is diminishing benefit from using a special ordinal analysis approach when variables have 5 or more ordinal categories instead of using normal theory analyses such as regression. For variables in between these two extremes, there are a number of ordinal analysis options that a researcher can choose from. I provide a brief introduction to a few of these below, but we will return to this topic when we discuss the more commonly employed and even more flexible analysis approaches of ordinal logistic and probit analysis.

Loglinear

The general loglinear model which I outlined in the previous handout “Loglinear Models” can be extended to take into account a rank ordering present in an $I \times J$ dimensional table. In a two-way table, the saturated model for the most common model, the *linear-by-linear association* model, uses a regression coefficient, indicated below by $\beta u_i v_j$, instead of the more general association term, λ_{ij}^{XY}, that was used in the nominal category loglinear model previously discussed.

$$\log\left(\mu_{ij}\right) = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$$

The first two parameters on the right hand side are based on the marginal proportions and are the same as in the prior loglinear model. The final term is the regression coefficient, using u_i and v_j to indicate the association between ordinal row and column frequencies. The regression coefficient is interpreted similarly to the logistic regression case with a log proportional increase in Y for each unit increase in X, or in the linear-by-linear model, it also represents a log proportional increase in X for each unit increase in Y. Significance tests are based on either G^2 (more commonly) or Pearson χ^2, similarly constructed to those discussed more generally for loglinear models, and are based on maximum likelihood estimation.[1] The test is a single-df test [as opposed to the $(I\text{-}1)(J\text{-}1)$ in the nominal categorical loglinear model]. β can be standardized for more general interpretation of its magnitude.

The odds ratio for any two adjacent cells (*local odds ratio*) can be computed. Using the a, b, c, d fourfold notation, the log of the odds ratio for adjacent cells is

$$\log\left(\frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}}\right) = \beta\left(u_c - u_a\right)\left(v_d - v_b\right)$$

If one assumes a *uniform association* (Goodman, 1979) in which the increment from one level of the row variable is equal across all levels and the increment in one level of the column variable is equal across all levels, a single odds ratio is e^{β} describes the ordinal association between X and Y.

[1] For the 2 × 2 case, you will see linear-by-linear association test statistics and p-values that are nearly identical to the Pearson chi-squared and likelihood ratio chi-squared tests. With larger ordinal tables you will usually see considerable differences, as each is testing a different hypothesis.

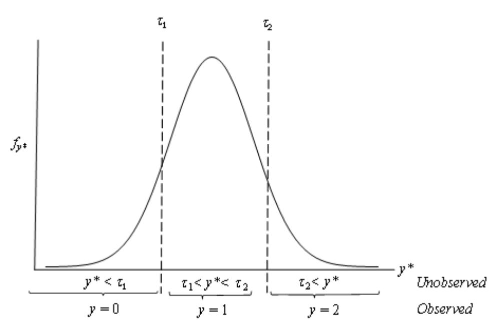
The model can be modified to only assume integer scoring in one direction, where either the column variable Y is integer scored ($Y = 1, 2, 3...$) or the row variable X is integer scored. Such models, called the *row-effect* or *column-effect model* (Goodman, 1979), respectively, are less restrictive than the linear-by-linear association model.

Tetrachoric and Polychoric Correlations

Pearson (1901) developed a correlation for binary variables, called a tetrachoric correlation, that estimates the correlation between two underlying continuous variables. With two normally distributed variables split at the mean/median, there is an information loss when their correlation is estimated using phi (or the usual Pearson correlation coefficient) that can be substantial [MacCallum and colleagues (2002) show the proportion of variance shared among the two dichotomized variables may be even five times smaller than the variance shared between their two continuous counterparts]. The tetrachoric correlation approach assumes two normally distributed variables and a bivariate normal distribution for the theoretical continuous variables, sometimes referred to as latent variables.[2] The computation of the tetrachoric correlation is complex and is best done with maximum likelihood (Olsson, 1979), but there are a several simpler approximations. Below is one from Goodman (1981), but there are several improvements upon this approximation (e.g., see Bonett & Price, 1985).

$$\hat{\rho}_{TC} \approx \frac{\sqrt{1+4\phi^2}-1}{2\phi}, \text{ where } \phi \text{ is } \phi = \frac{p}{1-p^2}$$

Tetrachoric correlations for two binary variables will tend to be larger than the phi correlation because of the correction for information loss. There are several terms for correlations derived from the same idea for different combinations of variable types. A similar type of correlation between a dichotomous and continuous variable is known as a polyserial correlation, and similar type of correlation between two ordinal variables is known as a polychoric correlation. The term polychoric correlation tends to be used generically for all of these, because the tetrachoric correlation is a special case of the polychoric correlation. For the polychoric correlation for ordinal variables, there is an underlying continuous distribution assumed, where there are thresholds, τ_j, corresponding the observed values (e.g., 0,1, 2) that divide up the underlying distribution.



The polychoric correlation can be connected to the multiplicative association term in loglinear models for ordinal variables (Wickens, 1989) and to ordinal probit models (covered later) and to the weighted least squares for ordinal variables in structural equation modeling (Muthén, 1978).

Gamma and Kendall's

Gamma (Goodman & Kruskal, 1954) and Kendall' tau (1948) are measures of association for ranked data. Ranks and ordinal data are closely associated, and both of these statistics are similar to

[2] An exact normal distribution may not be necessary in practice for the tetrachoric correlations to be reasonable estimates. The degree of deviation from normality and the sample size will likely be important in the accuracy of the tetrachoric estimate (Greer et al., 2003). Note also that Cohen (1983) shows that, when the underlying continuous variables are nonnormal, the information loss in the phi or Pearson's r for two dichotomized variables (in terms of the reduced proportion of shared variance) will be even greater than when the underlying variables are normally distributed.

Spearman's rho for ranked values. Gamma and Kendall's tau are usually used for fewer ordinal categories as opposed to ranked scores with many possible ranking values. Technically, a rank occurs if cases are sorted or ordered according to some measured or unmeasured variable. We could take students heights from class and instead of representing their heights by inches or centimeters, we simply order them from shortest to tallest. The distance between the scores now is only 1 unit even though the difference between any two students' heights might be more than one inch. In this sense, we have ordinal data. A more ordinal example is if we had ratings of perceptions of student height as "short" "average" and "tall." Ties exist when any two cases have the same rank. When the number of different observed ranks becomes large and the number of cases becomes large, there is often no important difference in statistical comparisons made with parametric tests.

If we construct a square, $I \times I$ table, we can designate all cells as concordant to a particular cell only if the scores on both variables are higher. Any other cells are discordant. For example, for variable X with scores 1, 2, 3, and Y, with scores 1, 2, 3, compared with cell (1,1), the following cells are concordant: (2,2), (2,3), (3,2), and (3,3). For cell (2,1), only the following are concordant: (3,2), (3,3). Below is the computation for C and D for a 3×3 table, which creates products and sums of concordant and discordant cell counts.

$$C = n_{11}\left(n_{22} + n_{23} + n_{32} + n_{33}\right) + n_{12}\left(n_{23} + n_{33}\right) + n_{21}\left(n_{32} + n_{33}\right) + n_{22}\left(n_{33}\right)$$
$$D = n_{13}\left(n_{21} + n_{22} + n_{31} + n_{32}\right) + n_{12}\left(n_{21} + n_{31}\right) + n_{23}\left(n_{31} + n_{32}\right) + n_{22}\left(n_{33}\right)$$

Gamma (γ) is then a simple ratio of the difference between C and D out of the sum total of the C and D frequencies.

$$\gamma = \frac{(C - D)}{(C + D)}$$

Gamma has values that range between -1 and +1 like the correlation coefficient. The special case of gamma for a 2 × 2 table is called Yule's Q, representing the proportion of cases that are concordant. Kendall's tau (or tau-a) is a special case with all pairs on the denominator. Kendall's tau also has values between -1 and +1.

$$\tau_a = \frac{(C - D)}{\binom{n}{2}}$$

where

$$\binom{n}{2} = \frac{n!}{2!(n-2)!}$$

tau-a does not adjust for ties. tau-b alters the denominator computation to adjust for ties, so it is better suited when there are a large number of ties. tau-c is designed for rectangular tables.

There are a number of other related or similar tests.[3] One can compute a partial tau which statistically controls for other variables. The partial tau is not particularly commonly used because other ordinal regression techniques are usually used for this purpose. Somer's d is an asymmetric measure of association for ordinal variables with versions for either X or Y as an explanatory variable. For similar reasons, asymmetrical relationships are usually examined with ordinal regression models. Tetrachoric and polychoric correlations compute associations assuming a continuous, latent, normal distribution

[3] Spearman's rho is often described as a rank-order statistic, but it is equivalent to the Pearson correlation when there are no ties.

underlying the observed ordinal values. We will discuss these measures of association in connection with the generalized linear model and probit analysis later in the course.

Measures of association for ranks or ordinal variables, including Spearman's rho, Gamma, Kendall's tau, or Sommer's d, are often referred to as "nonparametric" tests (Gibbons & Fielden, 1993). The term "nonparametric" is intended to refer to statistical tests that do not make some or any of the standard assumptions of parametric tests. The distinction is not so clear cut. Sheskin (2011), author of a large tome on nonparametric tests, states "In truth, nonparametric tests are really not assumption free, and, in view of this, some sources Marascuilo and McSweeney (1977) suggest that it might be more appropriate to employ the term 'assumption freer' rather than nonparametric in relation to such tests" (p. 109). These tests are not necessarily more efficient than other tests of association and are not necessarily less impacted by distributional violations. Because Spearman's rho is equivalent to Pearson's correlation coefficient, its statistical test is no more robust to distributional or statistical estimation problems. One reason that these tests may be categorized as nonparametric is that a method sometimes used for handling distributional problems with a continuous variable is to convert its raw values to ranks (e.g., imagine an extreme outlier that then simply becomes the highest rank score) and then use one the above-mentioned measures of association. Here, it is this conversion to ranks that is the key to any "nonparametric" robustness that might be gained with these measures of association.

References and Further Reading

Agresti, A. (2013). *Categorical data analysis, third edition*. New York: Wiley.
Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.
Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30*(2), 213-225.
Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.
Gibbons, J. D., & Fielden, J. D. G. (1993). *Nonparametric measures of association* (QASS No. 91). Sage.
Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association, 74*(367), 537-552.
Goodman, L. A. (1981). Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika, 68(*2), 347-355.
Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.
Greer, T., Dunlap, W. P., & Beatty, G. O. (2003). A Monte Carlo evaluation of the tetrachoric correlation coefficient. *Educational and Psychological Measurement, 63*(6), 931-950.
Kendall, M.G. (1948). *Rank Correlation Methods*. London: C. Griffin & Co.
MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19.
Muthén, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association, 74*(368), 807-811.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*(4), 443-460.
Pearson, K. (1901). XI. Mathematical contributions to the theory of evolution.—VII: On the correlation of characters not quantifiably measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195A, 1-47.
Powers, D. A., Xie, Y., & Xie, Y. (2000). "Chapter 4: Loglinear models for statistical tables" in *Statistical methods for categorical data analysis* (Vol. 106). New York: Academic Press.
Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures, Fifth Edition*. Boca Raton, FL: CRC Press.
Wickens, T. D. (1989). *Multiway contingency tables analysis*. Social Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.