Common Ordinal Analyses: Loglinear Models and Measures of Association

To date, we have covered a variety of statistical tests, some of them designed for continuous dependent variables and some designed for discrete dependent variables (binary, nominal categories). At the beginning the course, I gave an overview of the most common statistical tests (see the handout "Levels of Measurement and Choosing the Correct Statistical Test"), but we have not discussed analyses that can be used for a dependent variable with just a few ordinal categories (e.g., three or four). These special ordinal analyses are usually just used in the social sciences when the dependent variable only has a few values that can be ordered, such as a survey question with response options "never," "sometimes,", "always." For more ordinal categories on the dependent variable, most researchers in social and behavioral sciences tend to use analyses designed for continuous dependent variables. There seems to be fairly good evidence from simulation studies that suggests that if there are five or more ordered categories, there will be relatively little statistical inaccuracy in treating these ordinal variables as continuous (e.g., Johnson & Creech, 1983; Muthén & Kaplan, 1985; Zumbo & Zimmerman, 1993; Taylor, West, & Aiken, 2006). There are exceptions to this rule, particularly when continuous variables (of any number of ordered categories) are especially skewed or kurtotic and sample sizes are small. We will also discuss some additional remedies, such as nonparametric tests, robust standard errors, and bootstrapping that work well under many circumstances soon. With a dependent variable that is not binary but has fewer than five ordinal categories (i.e., 3 or 4), there are several analyses specifically for ordinal variables that are useful to know about. A few common ordinal analyses are summarized below:[1]

Independent Variable is:	Ordinal Dependent Variable
Nominal	Wilcoxon-Mann-Whitney (two groups) Kruskal-Wallis (ANOVA for ranks) Sign test (within-subjects, two measures) Friedman test (within-subjects, multiple groups) **Loglinear models** Ordinal regression models (ordinal logistic and probit)
Ordinal	Spearman's rho, **gamma**, **Kendall's tau, Somer's d,** Jonckheere-Terpstra, **loglinear models**, ordinal regression models (ordinal logistic and probit),
Continuous	Ordinal regression models (ordinal logistic and probit)

In this handout, I focus on the several of the simplest statistical tests that are the most commonly used for ordinal and rank data—loglinear models and ordinal measures of association. As the table suggests, ordinal regression models, which are primarily implemented with logistic or probit regression models, can be used whether the independent variable is nominal, ordinal, or continuous. Because these regression modeling approaches can be used to answer research questions for all three types of independent variables, ordinal regression modeling is perhaps more useful than any of the other approaches when the dependent variable has few ordinal values. We will discuss these models next term, however. Some of the other models noted in the table above, such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis, Sign test, and Friedman tests, usually fall under the category of "nonparametric" tests but can be used for ordinal dependent variables. I will discuss them in the subsequent handout "Nonparametric Statistics."

Loglinear Analysis

Loglinear models are an alternative method of analyzing contingency tables, with the ability to test many of the same hypotheses we have discussed up to this point. It is different framework for conceptualizing categorical data that has many advantages because of its flexibility for constructing tests for more complex designs with ordinal variables.

Natural logarithms. Before discussing loglinear models, we will need to review a few basic principles about logarithms, because natural log transformations are integral to the approach. The natural logarithm

[1] Adapted from Table 6.2, p. 135, Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. New York: Routledge

is complementary or inverse function to the exponential transformation. Though there are several variants of the logarithm function, we will focus only on the natural logarithm which is the inverse of the exponent function. The natural logarithm is denoted either as $\log_e$ ($\log$ base e), $\ln$, or in many statistical texts just $\log$. I am not fond of just using "log" because of the potential ambiguity, but to remain consistent with your reading, I will always be referring to the natural logarithm ($\ln$ or $\log_e$) when using $\log$ below. The exponential function indicated by e (and also by $\exp$; named after the Swiss mathematician Leonhard Euler) raises the constant, $e \approx 2.71828$, to some power, say x. In general, we could state that $e^{\log x} = x$. If x = 4.5, then $e^x = \exp(x) = e^{4.5} = 90.017$. The natural log reverses this transformation, $\log_e(90.017) = \log(90.017) = 4.5$. The natural log of an integer will always be a positive number, and the log of a positive decimal between 0 and 1 will always be negative. The log of a negative number is undefined.

There are a couple of special algebraic rules that will be used in discussing loglinear models. The log of a product is equal to the sum of the logs of each individual value or variable, known as the *product rule*.

$$\log(xy) = \log(x) + \log(y)$$

Incidentally, the analysis is called *loglinear*, because of this additive rule. The addition of terms on the right hand side of the equation implies a linear combination of the terms.

The basic loglinear model. Let's start with a simple problem in which we seek to learn whether there is a relationship between two binary variables, just as with the 2 × 2 contingency chi-square. The basic loglinear model is sometimes called the *independence model*, because its form assumes that the two variables in the analysis are independent or not correlated. This is not a new type of hypothesis, but later we will see how the loglinear model can be extended to ordinal variables. Your reading (Green, 1988) introduces a new notation for expected frequencies here. Instead of using E_{ij} for the expected count in a cell, the Greek letter "mu" is used, μ_{ij}. (Note that in this context i and j subscripts refer to row or column numbers, respectively, not individuals and groups). One way to express the computation for the expected frequencies for the Pearson χ^2 would be $\mu_{ij} = N\pi_{i+}\pi_{+j}$., with μ_{ij} representing the expected frequency, π_{i+} representing a row proportion and π_{+j} representing a column proportion. This equation looks a little different from the expected frequencies we computed for the 2 × 2 chi-square (see the handout "2 X 2 Contingency Chi-square"), where the expected frequency for a cell was $E_{ij} = R_i C_j / N$, but they are algebraically equivalent. The loglinear model expresses everything in terms of natural logs, so the log of the expected frequency for one cell is $log(\mu_{ij}) = log(N\pi_{i+}\pi_{+j})$. Using the product rule, we can see how one basic loglinear model, the independence model, partitions the expected frequencies for a cell into three components.

$$\log(\mu_{ij}) = \log(N) + \log(\pi_{i+}) + \log(\pi_{+j})$$

Above, μ_{ij} is the expected count for one cell, N is the total sample size, π_{i+} is the corresponding marginal row proportion, and π_{+j} is the corresponding column marginal proportion. The equation is true if X and Y are independent.[2] Each natural log term is often re-expressed as a set of *parameters*, using the Greek symbol λ ("lambda").

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

The superscript is not an exponent, but just denotes that the parameter pertains to the X (row) or the Y (column) variable in the contingency table. Each lambda represents each log term in the previous equation, $\lambda = \log(N)$, $\lambda_i^X = \log(\pi_{i+})$, and $\lambda_j^Y = \log(\pi_{j+})$.

[2] Although X and Y are used in this notation, one does not need to assume that X is an independent variable and Y is a dependent variable. Like chi-square or correlation, loglinear models are symmetric statistics in the sense that it is not necessary to posit a causal direction.

As an example, we can compute the log of the expected frequency for the first cell in the YouGov poll table from the 2 × 2 chi-square example, where the marginal proportion is equal to the marginal n divided by the total N, $\pi_{i+} = n_{i+}/N$ and $\pi_{+j} = n_{+j}/N$

	Did not vote	Voted	
All other ages	183	824	$n_{1+} = 1007$
Youngest	35	50	$n_{2+} = 85$
	$n_{+1} = 218$	$n_{+2} = 874$	$N = 1092$

$$\log(\mu_{ij}) = \log(N) + \log(\pi_{i+}) + \log(\pi_{+j})$$
$$\log(1092) + \log(1007/1092) + \log(218/1092)$$
$$= 7.00 + (-.08) + (-1.61) = 5.31$$

This value does not mean too much by itself, but if we use the exponent function to undo the log, then $e^{5.31} = 202.35$. This value happens to be equal (within rounding) to the expected value using the expected value formula for the Pearson χ^2, which, in the "2 X 2 Contingency Chi-square" handout, we computed as $E_{11} = (1007 * 218)/1092 = 201.03$.

Two-way contingency table tests

The independence model holds true under the null hypothesis that the conditional proportions in each row are equal (i.e., that the odds ratio is 1.0). If the null hypothesis is not true, then there must be something added to the right hand side of the equation that would produce a value larger or smaller than $\log(\mu_{ij})$. The extra term on the righthand side of the equation, λ_{ij}^{XY}, is called the *association term*[3] and represents the degree of non-independence, which, just as with chi-square, quantifies the departure of the observed frequencies from the expected frequencies.

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

Output from software packages includes statistical tests of each of the terms. The test of λ_i^X is a test of whether the frequencies differ on the first variable, the test of λ_j^Y is a test of whether the frequencies differ on the second variable, and the test of λ_{ij}^{XY} is a test of whether there is dependence or an association between the two variables.

Fit and model comparisons through nested tests. Adding this last term to the model produces the so-called *saturated model*, in which there is the same number of parameters as there are observed frequencies (i.e., the fit is perfect). Testing hypotheses in loglinear modeling sometimes involves the comparison of a model with more parameters to a model with fewer parameters. Although one could construct other tests by comparing the saturated model to a model that drops another parameter (e.g., λ_i^2), nearly always the interest in 2 × 2 contingency tables is with the comparison of the saturated and the independence model. Although a kind of Pearson χ^2 could also be used for this test, loglinear models are most often tested with the likelihood ratio test, G^2, with $df = (I - 1)(J - 1)$.

$$G^2 = 2\sum_i^I \sum_j^J N_{ij} \log\left(\frac{N_{ij}}{\mu_{ij}}\right)$$

[3] This term is also commonly referred to as the interaction term, because the value of Y depends on the value of X. I think the analogy to the factorial ANOVA is potentially confusing, however, because we have a circumstance with only two variables rather than three variables (two independent and one dependent variable). The two-way table in chi-square is like a t-test in which the proportions in two groups are compared (rather than means).

The G^2 value will usually be very similar to the Pearson chi-square, but it is a different way of computing the fit of the expected to the observed values. Like the Pearson chi-square, the loglinear model for the 2 × 2 contingency table is easily extended to $I \times J$ tables. For these larger designs, the table can be divided up into a set of 2 × 2 subtables for follow-up tests.[4]

Ordinal loglinear models

The general loglinear model described above 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.[5]

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

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. Significance of each component ("parameter") is estimated. Overall model fit is also obtained, based on either G^2 (more commonly) or Pearson χ^2, and are similarly constructed to those discussed more generally for loglinear models.[6] These fit tests can be used to compare different models. The models are estimated with a maximum likelihood process. The test is a single-df test [as opposed to the $(I - 1)(J - 1)$ in the nominal categorical loglinear model]. β can be standardized for a more general interpretation of its magnitude.

Measures of Association: Gamma and Kendall's Tau

Gamma (Goodman & Kruskal, 1954) and Kendall's tau-b (1948) are measures of association for ranked data. Ranks and ordinal data are closely associated, and both of these statistics are similar to Spearman's rho for ranked values. Gamma and Kendall's tau-b are usually used for fewer ordinal categories, however, even though they are just as applicable to rankings.[7] 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 one 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 (as with the connection between Spearman's rho and the Pearson correlation coefficient).

There are a number of other related or similar tests. 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 can be formulated as a symmetric (i.e., no distinction between independent and dependent variable) or as 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

---

[4] The loglinear model for two-way contingency tables can be extended to three-way contingency tables as well (e.g., two independent variables and one dependent variable): $\log\left(\mu_{ijk}\right) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{jk}^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$, which does correspond to the two-way factorial ANOVA (with all binary variables).

[5] Some authors, like Green (1988), give the $u_i v_j$ term in terms of deviation scores, $\left(u_i - \overline{u}\right)$ and $\left(v_i - \overline{v}\right)$ instead.

[6] 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 (which are printed in SPSS). With larger ordinal tables you will usually see considerable differences, as each is testing a different hypothesis.

[7] You will also hear about other versions of Kendall's tau: 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.

and polychoric correlations compute associations assuming a continuous, latent, normal distribution underlying the observed ordinal values. We will discuss these measures of association in connection with the generalized linear model and probit analysis in the next course.

References

Green, J.A. (1988). Loglinear Analysis of Cross-Classified Ordinal Data: Applications in Developmental Research. *Child Development, 59*, 1-25.

Johnson, D.R., & Creech, J.C. (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review, 48*, 398-407.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189.

Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. New York: Routledge.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and psychological measurement*, *66*(2), 228-239.

Zumbo, B.D., & Zimmerman, D.W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34,* 390-400.