Encyclopedia of Research Design

Categorical Data Analysis

Contributors: Maria Kateri & Alan Agresti
Editors: Neil J. Salkind
Book Title: Encyclopedia of Research Design
Chapter Title: "Categorical Data Analysis"
Pub. Date: 2010
Access Date: October 17, 2013
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412961271
Online ISBN: 9781412961288
DOI: http://dx.doi.org/10.4135/9781412961288.n40
Print pages: 120-124

http://dx.doi.org/10.4135/9781412961288.n40
A categorical variable consists of a set of non-overlapping categories. Categorical data are *counts* for those categories. The measurement scale of a categorical variable is *ordinal* if the categories exhibit a natural ordering, such as opinion variables with categories from “strongly disagree” to “strongly agree.” The measurement scale is *nominal* if there is no inherent ordering. The types of possible analysis for categorical data depend on the measurement scale.

Types of Analysis

When the subjects measured are cross-classified on two or more categorical variables, the table of counts for the various combinations of categories is a *contingency table.* The information in a contingency table can be summarized and further analyzed through appropriate *measures of association* and models as discussed below. These measures and models differentiate according to the nature of the classification variables (nominal or ordinal).

Most studies distinguish between one or more *response variables* and a set of *explanatory variables.* When the main focus is on the association and interaction structure among a set of response variables, such as whether two variables are conditionally independent given values for the other variables, *loglinear models* are useful, as described in a later section. More commonly, research questions focus on effects of explanatory variables on a categorical response variable. Those explanatory variables might be categorical, quantitative, or of both types. *Logistic regression models* are then of **[p. 120 ↓]** particular interest. Initially such models were developed for *binary* (success–failure) response variables. They describe the *logit*, which is log[*P*(*Y* = 1)/*P*(*Y* = 2)], using the equation

$$\log[P(Y = 1)/P(Y = 2)] = a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where *Y* is the binary response variable and *x*
1

, …, *x*
p

the set of the explanatory variables. The logistic regression model was later extended to nominal and ordinal response variables. For a nominal response *Y* with *J* categories, the model simultaneously describes

$$\log[P(Y = 1)/P(Y = J)],$$
$$\log[P(Y = 2)/P(Y = J)], \ldots,$$
$$\log[P(Y = J - 1)/P(Y = J)].$$

For ordinal responses, a popular model uses explanatory variables to predict a logit defined in terms of a cumulative probability,

$$\log[P(Y \leq j)/P(Y > j)], j = 1, 2, \ldots, J - 1.$$

For categorical data, the binomial and multinomial distributions play the central role that the normal does for quantitative data. Models for categorical data assuming the binomial or multinomial were unified with standard regression and *analysis of variance* (ANOVA) models for quantitative data assuming normality were unified through the introduction of the *generalized linear model* (GLM). This very wide class of models can incorporate data assumed to come from any of a variety of standard distributions (such as the normal, binomial, and Poisson). The GLM relates a function of the mean (such as the log or logit of the mean) to explanatory variables with a linear predictor. Certain GLMs for counts, such as *Poisson regression* models, relate naturally to log linear and logistic models for binomial and multinomial responses.

More recently, methods for categorical data have been extended to include *clustered data*, for which observations within each cluster are allowed to be correlated. A very important special case is that of *repeated measurements*, such as in a longitudinal study in which each subject provides a cluster of observations taken at different times. One way this is done is to introduce a *random effect* in the model to represent each cluster, thus extending the GLM to a *generalized linear mixed model*, the *mixed*

referring to the model's containing both random effects and the usual sorts of fixed effects.

Two-Way Contingency Tables

Two categorical variables are *independent* if the probability of response in any particular category of one variable is the same for each category of the other variable. The most well-known result on two-way contingency tables is the test of the null hypothesis of independence, introduced by Karl Pearson in 1900. If X and Y are two categorical variables with I and J categories, respectively, then their cross-classification leads to a $I \times J$ table of observed frequencies $\mathbf{n} = (n_{ij})$. Under this hypothesis, the expected cell frequencies are values that have the same marginal totals as the observed counts but perfectly satisfy the hypothesis. They equal

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \qquad (1)$$

, where n is the total sample size $(n = \sum_{i,j} n_{ij})$ and $\pi_{i+}(\pi_{+j})$ is the ith row (jth column) marginal of the underlying probabilities matrix $\pi = (\pi_{ij})$. Then the corresponding maximum likelihood (ML) estimates equal

$$\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}$$

, where p_{ij} denotes the sample proportion in cell (i, j). The hypothesis of independence is tested through Pearson's chi-square statistic,

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \qquad (1)$$

The p value is the right-tail probability above the observed X^2 value. The distribution of X^2 under the null hypothesis is approximated by a $\chi^2_{(I-1)(J-1)}$, provided that the individual expected cell frequencies are not too small. In fact, Pearson claimed that the associated degrees of freedom (df) were $IJ-1$, and R. A. Fisher corrected this in 1922. Fisher later proposed a small-sample test of independence for 2 × 2 tables, now referred to as *Fisher's exact test.* This test was later extended to $I \times J$ tables as well as to more complex hypotheses in both two-way and multiway tables. When a contingency table has ordered row or column categories (ordinal variables), specialized methods can take advantage of that ordering.

[p. 121 ↓]

Ultimately more important than mere testing of significance is the estimation of the strength of the association. For ordinal data, measures can incorporate information about the direction (positive or negative) of the association as well.

More generally, models can be formulated that are more complex than independence, and expected frequencies m_{ij}

can be estimated under the constraint that the model holds. If

$$G^2 = 2\sum_{i,j} n_{ij} \ln\left(\frac{n_{ij}}{\hat{m}_{ij}}\right). \qquad (2)$$

are the corresponding maximum likelihood estimates, then, to test the hypothesis that the model holds, one can use the Pearson statistic (Equation 1) or the statistic that results from the standard statistical approach of conducting a *likelikood-ratio test*, which is

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

Under the null hypothesis, both statistics have the same large-sample chi-square distribution.

The special case of the 2 × 2 table occurs commonly in practice, for instance for comparing two groups on a success/fail–type outcome. In a 2 × 2 table, the basic measure of association is the *odds ratio.* For the probability table

$$\log(\hat{\theta}) \sim N\left(\log(\theta), \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

the odds ratio is defined as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y,$$
$$i = 1, \ldots, I,\ j = 1, \ldots, J.$$

. Independence corresponds to $\theta = 1$. Inference about the odds ratio can be based on the fact that for large samples,

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J. \tag{3}$$

The odds ratio relates to the *relative risk r.* In particular, if we assume that the rows of the above 2 × 2 table represent two independent groups of subjects (A and B) and the columns correspond to presence/absence of a disease, then the relative risk for this disease is defined as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j, \quad i = 1, \ldots, I, \; j = 1, \ldots, J, \tag{4}$$

, where

$$\pi_A = \frac{\pi_{11}}{\pi_{1+}}$$

is the probability of disease for the first group and π_B

is defined analogously. Since

$$\theta = r\frac{1-\pi_B}{1-\pi_A}$$

, it follows that $\theta \approx r$ whenever π_A

and π_B

are close to 0.

Models for Two-Way Contingency Tables

Independence between the classification variables *X* and *Y* (i.e., *m*
ij

= n#
i+

#
+j

, for all *i* and *j*) can equivalently be expressed in terms of a log linear model as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y,$$
$$i = 1, \ldots, I,\ j = 1, \ldots, J.$$

The more general model that allows association between the variables is

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad (3)$$
$$i = 1, \ldots, I,\ \ j = 1, \ldots, J.$$

Loglinear models describe the way the categorical variables and their association influence the count in each cell of the contingency table. They can be considered as a discrete analogue of ANOVA. The two-factor interaction terms relate to odds ratios describing the association. As in ANOVA models, some parameters are redundant in these specifications, and software reports estimates by assuming certain constraints.

The general model (Equation 3) does not impose any structure on the underlying association, and so it fits the data perfectly. Associations can be modeled through *association models.* The simplest such model, the *linear-by-linear association model*, is

relevant when both classification variables are ordinal. It replaces the interaction term Λ_{ij}^{XY} by the product $\#\mu_i \nu_j$, where μ_i and ν_j are known scores assigned to the row and column categories, respectively. This model,

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j, \quad i = 1, \ldots, I, \; j = 1, \ldots, J, \tag{4}$$

has only one parameter more than the independence model, namely φ. Consequently, the associated *df* are $(I-1)(J-1)-1$, and once it holds, independence can be tested conditionally on it by testing $\varphi = 0$ via a more powerful test with $df = 1$. The linear-by-linear association model (Equation 4) can equivalently be expressed in terms of the $(I-1)(J-1)$ local odds ratios

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} (I = 1, \ldots, I-1, j = 1, \ldots, J-1)$$

, defined by adjacent rows and columns of the table:

[p. 122 ↓]

$$\theta_{ij} = \exp[\phi(\mu_{i+1} - \mu_i)(\nu_{j+1} - \nu_j)], \quad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1. \qquad (5)$$

With equally spaced scores, all the local odds ratios are identical, and the model is referred to as *uniform association.* More general models treat one or both sets of scores as parameters. Association models have been mainly developed by L. Goodman.

Another popular method for studying the pattern of association between the row and column categories of a two-way contingency table is *correspondence analysis* (CA). It is mainly a descriptive method. CA assigns optimal scores to the row and column categories and plots these scores in two or three dimensions, providing thus a reduced rank display of the underlying association.

The special case of square *I* × *I* contingency tables with the same categories for the rows and the columns occurs with matched-pairs data. For example, such tables occur in the study of *rater agreement* and in the analysis of social mobility. A condition of particular interest for such data is *marginal homogeneity*, that $\pi_{i+} = \pi_{+i}$, *i* = 1, …, *I.* For the 2 × 2 case of binary matched pairs, the test comparing the margins using the chi-square statistic $(n_{12} - n_{21})^2/(n_{12} + n_{21})$

) is called *McNemar's test.*

Multiway Contingency Tables

The models described earlier for two-way tables extend to higher dimensions. For multidimensional tables, a number of models are available, varying in terms of the complexity of the association structure. For three variables, for instance, models include ones for which (a) the variables are mutually independent; (b) two of the variables are associated but are jointly independent of the third; (c) two of the variables are conditionally independent, given the third variable, but may both be associated with the third; and (d) each pair of variables is associated, but the association between each pair has the same strength at each level of the third variable. Because the number of possible models increases dramatically with the dimension, model selection methods become more important as the dimension of the table increases. When the underlying theory for the research study does not suggest particular methods, one can use the same methods that are available for ordinary regression models, such as stepwise selection methods and fit indices such as Akaike Information Criterion.

Loglinear models for multiway tables can include higher order interactions up to the order equal to the dimension of the table. Two-factor terms describe conditional association between two variables, three-factor terms describe how the conditional association varies among categories of a third variable, and so forth. CA has also been extended to higher dimensional tables, leading to *multiple CA.*

Historically, a common way to analyze higher way contingency tables was to analyze all the two-way tables obtained by collapsing the table over the other variables. However, the two-way associations can be quite different from conditional associations in which other variables are controlled. The association can even change direction, a phenomenon known as *Simpson's paradox.* Conditions under which tables can be collapsed are most easily expressed and visualized using graphical models that portray each variable as a node and a conditional association as a connection between two nodes. The patterns of associations and their strengths in two-way or multiway tables can also be illustrated through special plots called *mosaic plots.*

Inference and Software

Least squares is not an optimal estimation method for categorical data, because the variance of sample proportions is not constant but rather depends on the corresponding population proportions. Because of this, parameters for categorical data were estimated historically by the use of *weighted least squares*, giving more weight to observations having smaller variances. Currently, the most popular estimation method is *maximum likelihood*, which is an optimal method for large samples for any type of data. The *Bayesian* approach to inference, in which researchers combine the information from the data with their prior beliefs to obtain posterior distributions for the parameters of interest, is becoming more popular. For large samples, all these methods yield similar results.

Standard statistical packages, such as SAS, SPSS (an IBM company, formerly called PASW® Statistics), Stata, and S-Plus and R, are well suited for analyzing categorical data. Such packages now **[p. 123 ↓]** have facility for fitting GLMs, and most of the standard methods for categorical data can be viewed as special cases of such modeling. Bayesian analysis of categorical data can be carried out through WINBUGS. Specialized software such as the programs StatXact and LogXact, developed by Cytel Software, are available for small-sample exact methods of inference for contingency tables and for logistic regression parameters.

Maria Kateri and Alan Agresti

http://dx.doi.org/10.4135/9781412961288.n40
See also

Further Readings

Agresti, A. (2002). Categorical data analysis (2nd ed.). New York: Wiley.

Agresti, A. (2007). An introduction to categorical data analysis (2nd ed.). New York: Wiley.

Agresti, A. (2010). Analysis of Ordinal Categorical Data. New York: Wiley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press.

Congdon, P. (2005). Bayesian Models for Categorical Data. New York: Wiley.

Goodman, L. A. Some useful extensions of the usual correspondence analysis and the usual log-linear models approach in the analysis of contingency tables with or without missing entries. International Statistical Review, (1986). vol. 54, pp. 243–309.

Kateri, M. (2008). Categorical data. In S. Kotz (Ed.), Encyclopedia of statistical sciences (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Websites

Cytel Software: http://www.cytel.com

WINBUGS: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml