CATEGORICAL DATA ANALYSIS

S.R. Lipsitz
Medical University of the South Carolina, USA

G.M. Fitzmaurice
Harvard School of Public Health, Boston, MA, U.S.A

Keywords: Bernoulli experiment, proportion, contingency table, likelihood ratio test, chi-squared test, confounding, log-linear model, logistic regression, multinomial regression, Poisson regression, clustered categorical data

Contents

Summary

Modeling and inferential procedures (estimation and precision assessment) are discussed for a single proportion, thereafter in the context of cross-classified data, i.e., contingency tables, is discussed. First, the situation of two by two tables is given some treatment and then the more general context of *R* by *C* tables is considered. Particular emphasis is put on testing the null hypothesis of no association (independence) between the row and column classification. Next, the setting of multiple contingency tables is considered. Attention is given to the log-linear model and to the multinomial regression model. When the covariate is continuous, these methods naturally extend to logistic regression, a method at the intersection of generalized linear modeling and categorical data analysis. When the outcome is a count, one often uses Poisson regression. Finally, correlated categorical data settings are briefly discussed.

1. Introduction

In this chapter we present an overview of many of the statistical methods commonly used for the analysis of categorical or discrete “outcome” data. A discrete random variable is defined as one that takes on a finite number of values (e.g., “success” and “failure” in the case of a Bernoulli random variable) or a countably infinite number of

values (e.g., the number of premature births in a health district during 1999). For example, consider the data in Table 1 which are from a clinical trial of patients with cancer. In this study each patient was randomly assigned to one of two treatments (here denoted A and B) and the investigators were interested in determining which of the two treatments was superior in terms of curing the patients of disease. In this simple illustration, the discrete outcome has two levels, "cured" or "not cured". Table 1 is commonly referred to as a 2 × 2 contingency table. Much of the statistical theory underlying the analysis of categorical data is more easily formulated for 2 × 2 contingency tables. Indeed, methods for the analysis of 2 × 2 contingency tables provide the cornerstone for many of the advanced statistical methods required for more complicated problems. These include extensions for analyzing outcomes with more than 2 levels (e.g., "not cured", "partially cured", and "cured"), which may or may not be ordered; the former are referred to as an ordinal variables, the latter are referred to as nominal variables. In addition, there can be more than 2 levels of the experimental treatment or exposure variable (e.g., A, B, C, and D) and other factors or covariates (e.g., age, gender, health status before treatment) that influence the outcome variable.

	RESPONSE		
TREATMENT	Not Cured	Cured	TOTAL
A	49	26	75
B	64	11	75
TOTAL	113	37	150

Table 1: Illustrative data from a clinical trial of patients with cancer

Some of the most widely used probability distributions for discrete outcomes include the Bernoulli, binomial, hypergeometric, multinomial, and Poisson distributions. Throughout this chapter we assume the reader has very little prior knowledge of these probability distributions. The chapter is organized as follows. We begin with a discussion of inference for a single probability or proportion. This is followed by a description of methods for analyzing 2 × 2 contingency tables, and the extensions to $R \times C$ contingency tables (i.e., contingency tables with R rows and C columns). We also discuss the analysis of sets of 2 × 2 tables, and describe the Cochran-Mantel-Haenszel test. Finally, we present an overview of regression models for categorical data, including log-linear models for count data, logistic regression models for binary outcomes, multinomial logistic regression models for nominal and ordinal outcomes, and Poisson regression for rates. We also mention some issues that need to be considered when these regression models are applied to clustered categorical data.

2. Inference for a Single Proportion

In this section we discuss inference for a single proportion or probability. In order to motivate the methods, consider the following example from a cancer clinical trial. Phase II cancer clinical trials are usually designed to examine whether a single new treatment produces favorable results (e.g., in terms of the proportion of successes). Patients in the study receive a single treatment and the outcome for each patient can be denoted

$$Y_i = \begin{cases} 1 & \text{if new treatment is a success, (e.g. shrinks tumor),} \\ 0 & \text{if new treatment is a failure, (e.g. does not shrink tumor),} \end{cases}$$

for $i = 1, \ldots, n$ independent patients. The probability of success is denoted by $p = \text{pr}(Y_i = 1)$ and the probability of failure by $1 - p = \text{pr}(Y_i = 0)$. The distribution of the number of successes among the n patients, $Y = \sum_{i=1}^{n} Y_i$, can be used to form test statistics and a confidence interval for p. The probability distribution function for Y is binomial

$$\text{pr}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Furthermore, it can be shown that the maximum likelihood estimate of p is the proportion of successes, $\hat{p} = \frac{Y}{n}$. In large samples (say, $n > 30$, and $Y \geq 5$), $\hat{p}$ has an approximate normal distribution with mean p and variance $\frac{p(1-p)}{n}$. A 95% confidence interval for p is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Note, however, that even though $0 \leq p \leq 1$, the endpoints of this confidence interval are not restricted to be in [0,1]. When p is close to 0 or 1 (so that $\hat{p}$ will usually be close to 0 or 1), and/or in relatively small samples, the endpoints can fall outside of [0,1].

To circumvent this problem, we can instead base the confidence interval for p on the "logit" or "log-odds" of success, $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. Here, $\frac{p}{1-p}$ is the "odds" of success versus failure. Note that if $\theta = \text{logit}(p)$, then $p = \frac{e^{\theta}}{1+e^{\theta}}$. Thus, if we can obtain a 95 % confidence interval for $\theta = \text{logit}(p)$ with endpoints $\left[\hat{\theta}_{\text{L}}, \hat{\theta}_{\text{U}}\right]$ then by transforming the endpoints of the confidence interval for θ we can obtain a 95 % confidence interval for p of the form

$$\left[\frac{\exp\hat{\theta}_{\text{L}}}{1+\exp\hat{\theta}_{\text{L}}}, \frac{\exp\hat{\theta}_{\text{U}}}{1+\exp\hat{\theta}_{\text{U}}}\right].$$

This ensures that the endpoints of the resulting confidence interval for p are within [0,1]. From maximum likelihood theory, and application of a technique known as the delta method, the appropriate $\hat{\theta}_L$ and $\hat{\theta}_U$ are given by,

$$\hat{\theta}_L = \text{logit}(\hat{p}) - 1.96\sqrt{\frac{1}{y} + \frac{1}{n-y}} \quad \text{and} \quad \hat{\theta}_U = \text{logit}(\hat{p}) + 1.96\sqrt{\frac{1}{y} + \frac{1}{n-y}}.$$

Alternatively, when sample sizes are relatively small (say, $n < 30$), an exact confidence interval can be obtained that is based on the binomial distribution for Y. Finally, hypotheses tests for p equaling a specified value, say p_0, can be conducted using either large sample theory for the approximate normal distribution of $\hat{p}$ or via exact methods based on the binomial distribution for Y.

-
-
-

Bibliography

Agresti, A. (1990), *Categorical Data Analysis*. New York: Wiley. [This book provides a comprehensive survey of statistical methods for analyzing categorical data.]

Breslow, N.E. and Day, N.E. (1990). *Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case-Control Studies*. Lyon: World Health Organization. [This is a classic textbook on the analysis of data from case-control studies, including logistic regression and methods for analysing contingency table data]

Fisher, R.A. (1934). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd. [A basic early reference, where Fisher's exact method is proposed]

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719--748. [A basic article in which the Cochran-Mantel-Haenszel test is developed.]

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. 2nd edition. London: Chapman & Hall. [A key reference to generalized linear models.]

Pendergast J.F., Gange S.J., Newton M.A., Lindstrom M.J., Palta M. and Fisher M.R. (1996) A survey of methods for analyzing clustered binary response data. *International Statistical Review*, **64**, 89-118. [A comprehensive review article on methods for correlated binary data]

Biographical Sketches

Stuart R. Lipsitz is Professor at the Medical University of the South Carolina. He holds a doctoral degree from the Harvard School of Public Health, Boston, Massachusetts. His areas of interest include

repeated categorical data, with emphasis on marginal and on generalized estimating equations related methods, and models for the analysis of incomplete longitudinal data.

Garrett Fitzmaurice is Associate Professor of Biostatistics at Harvard University. His methodological research is focused on a number of inter-related areas, including methods for analyzing discrete longitudinal data, models for the joint analysis of mixed discrete and continuous outcomes, missing data problems, methods for detecting and adjusting for overdispersion, and statistical issues in psychiatric epidemiology and mental health research. Much of his collaborative research is focused on applications to mental health research, broadly defined.