ST3241 Categorical Data Analysis

Review I

Some Topics Covered

- Introduction to Categorical Data
- Two-way Contingency Tables
- Three-way Contingency Tables
- Generalized Linear Models
- Logistic Regression
- Log-linear Models

Categorical Data

- A *categorical* variable is one for which the measurement scale consists of a set of categories
- One and only one category should be applied to each subject.
- Ordinal variable: Categories are ordered
- Nominal variable: Categories can not be ordered.

Probability Distributions Involved

- Poisson Distribution

$$P[Y = y] = \frac{e^{-\mu}\mu^y}{y!}, y = 0, 1, 2, \cdots$$

- Binomial Distribution

$$P[Y = y] = \frac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y}, y = 0, 1, \cdots, N$$

- Multinomial distribution: more than 2 outcomes

Inferences

- Parameter estimation–Maximum likelihood estimation (MLE)
- Hypothesis testing: z-test (CLT-based) or proportion test
- Confidence interval construction: CLT-based or exact distribution based

Contingency Table

- Let X: I levels and Y: J levels, be two categorical variables
- Display the IJ possible combinations of outcomes in a rectangular table having I rows and J columns to form a two-way *contingency table* or an $I \times J$ table.
- Similarly, a table which cross classifies three variables is called a $three - way\ table$.

Some Notations, Definitions $\cdots$

- $\pi_{ij} = P[X = i, Y = j]$: cell probability that (X, Y) falls in the (i, j)-th cell
- The probabilities $\{\pi_{ij}\}$ form the joint distribution of X and Y.
- Row marginal distribution of X $\pi_{i+} = \sum_{j=1}^{J} \pi_{ij}$
- Column marginal distribution of Y $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$

Notations For The Data

- Cell counts are $\{n_{ij}\}$, with $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$
- Cell proportions are $p_{ij} = \frac{n_{ij}}{n}$
- The marginal frequencies are row totals $\{n_{i+}\}$and column totals $\{n_{+j}\}$

Independence

- Two variables are statistically independent if all joint probabilities equal the product of their marginal probabilities $\pi_{ij} = \pi_{i+}\pi_{+j}$, for $i = 1, \cdots, I$ and $j = 1, \cdots, J$
- Conditional distributions of Y are identical at each level of X.

Probability Model For A 2×2 Table

- Poisson Model: Each of the 4 cell counts are independent Poisson random variables
- Binomial Model: Marginal totals of X are fixed and Conditional distributions of Y at each level of X are binomial.
- Multinomial Model: Total sample size is fixed and the distribution of 4 cell counts are multinomial

Comparing Proportions in 2×2 Tables

- Assume a binomial model with the two categories of Y as *success* and *failure*.
- Let π_1 = Probability of success in row 1 and π_2 = Probability of success in row 2.
- The difference in probabilities $\pi_1 - \pi_2$ compares the success probabilities in two rows.

Sample Difference of Proportions

- The sample proportion difference $p_1 - p_2$ estimates the population proportion difference $\pi_1 - \pi_2$.
- Under the independence assumption, the estimated standard error of $p_1 - p_2$ is

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

- A large sample $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is then

$$p_1 - p_2 \pm z_{\alpha/2}\hat{\sigma}(p_1 - p_2)$$

where $z_{\alpha/2}$ denotes the table value of $N(0,1)$.

Relative Risk

- In 2×2 tables, the relative risk is the ratio of the success probabilities for the two groups π_1/π_2.
- Sample relative risk $= p_1/p_2$.
- Its distribution is heavily skewed and cannot be approximated by normal distribution well unless the sample sizes are quite large.
- A large sample confidence interval is given by

$$\exp\left[\log(\frac{p_1}{p_2}) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_{1+}p_1} + \frac{1-p_2}{n_{2+}p_2}}\right]$$

Odds Ratio

- The odds of success for Row 1 is $Odds_1 = \pi_1/(1-\pi_1)$
- The odds of success for Row 2 is $Odds_2 = \pi_2/(1-\pi_2)$
- Odds Ratio

$$\theta = \frac{Odds_1}{Odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

- When both variables are responses, the odds ratio can be defined using the joint probability as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

and called $cross-product\ ratio$.

Sample Odds Ratio

- Sample odds ratio is defined as

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

- For small to moderate sample size, the distribution of sample odds ratio $\hat{\theta}$ is highly skewed.
- The sample log odds ratio, $\log\hat{\theta}$ has a less skewed distribution and can be approximated by the normal distribution well.
- The asymptotic standard error of $\log\hat{\theta}$ is given by

$$ASE(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Confidence Intervals

- A large sample confidence interval for $\log \theta$ is given by

$$\log(\hat{\theta}) \pm z_{\alpha/2} ASE(\log \hat{\theta})$$

- A large sample confidence interval for θ is given by

$$\exp[\log(\hat{\theta}) \pm z_{\alpha/2} ASE(\log \hat{\theta})]$$

Tests of Independence

- To test: $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j.
- Equivalently, $H_0 : \mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$.
- Usually, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown.
- We estimate them, using sample proportions

$$\hat{\mu_{ij}} = np_{i+}p_{+j} = n\frac{n_{i+}n_{+j}}{n^2} = \frac{n_{i+}n_{+j}}{n}$$

- These $\{\hat{\mu}_{ij}\}$ are called estimated expected cell frequencies

Test Statistics

- Pearson's Chi-square test statistic

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

- Likelihood ratio test statistic

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} log(\frac{n_{ij}}{\hat{\mu}_{ij}})$$

- Both of them have large sample chi-squared distribution with $(I-1)(J-1)$ degrees of freedom.

Residuals

- To understand better the nature of evidence against H0, a cell by cell comparison of observed and estimated frequencies is necessary.
- Define, adjusted residuals

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

- If H_0 is true, each r_{ij} has a large sample standard normal distribution.
- If r_{ij} in a cell exceeds 2 then it indicates lack of fit of H_0 in that cell.
- The sign also describes the nature of association.

Testing Independence For Ordinal Data

- For ordinal data, it is important to look for types of associations when there is dependence.
- It is quite common to assume that as the levels of X increases, responses on Y tend to increases or responses on Y tends to decrease to ward higher levels of X.
- The most simple and common analysis assigns scores to categories and measures the degree of *linear trend* or correlation, known as "Mantel-Haenszel Chi-Square" test (Mantel and Haenszel 1959).

Linear Trend Alternative to Independence

- Let $u_1 \leq u_2 \leq \cdots \leq u_I$ and $v_1 \leq v_2 \leq \cdots \leq v_J$ denote scores for the rows and columns.
- The scores have the same ordering as the category levels.
- Define the correlation between X and Y as

$$r = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} u_i v_j n_{ij} - (\sum_{i=1}^{I} u_i n_{i+})(\sum_{j=1}^{J} v_j n_{+j})/n}{\sqrt{[\sum_{i=1}^{I} u_i^2 n_{i+} - (\sum_{i=1}^{I} u_i n_{i+})^2/n][\sum_{j=1}^{J} v_j^2 n_{+j} - (\sum_{j=1}^{J} v_j n_{+j})^2/n]}}$$

Test For Linear Trend Alternative

- Independence between the variables implies that its true value equals zero.
- The larger the correlation is in absolute value, the farther the data fall from independence in this linear dimension.
- A test statistic is given by $M^2 = (n-1)r^2$.
- For large samples, it has approximately a chi-squared distribution with 1 degrees of freedom.

Fisher's Exact Test

- For a 2×2 table, under the assumption of independence, i.e. $\theta = 1$, the conditional distribution of n_{11} given the row and column totals is hypergeometric.
- For given row and column marginal totals, the value for n_{11} determines the other three cell counts. Thus, the hypergeometric formula expresses probabilities for the four cell counts in terms of n_{11} alone.

Fisher's Exact Test

- When $\theta = 1$, the probability of a particular value n_{11} for that count equals

$$p(n_{11}) = \frac{\begin{pmatrix} n_{1+} \\ n_{11} \end{pmatrix} \begin{pmatrix} n_{2+} \\ n_{+1} - n_{11} \end{pmatrix}}{\begin{pmatrix} n \\ n_{+1} \end{pmatrix}}$$

- To test independence, the p-value is the sum of hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome.

Three-way Table

Three-way Table for X, Y, Z

- To find association between X and Y by controlling other covariates that can influence the association.
- We study the effect of X on Y by fixing such covariates constant.
- In other words, study the association between X and Y given the levels of Z.

Partial Tables

- Two-way tables between X and Y at separate levels of Z.
- The two-way contingency table obtained by combining the partial tables is called the $X - Y$ marginal table.
- Each cell count in the marginal table is a sum of counts from the same cell location in the partial tables.
- The marginal table, rather than controlling Z, ignores it and does not contain any information about Z.

Notes

- The associations in partial tables are called conditional associations.
- Conditional associations in partial tables can be quite different from associations in marginal tables
- The result that a marginal association can have different direction from the conditional associations is called *Simpson's paradox*.

Conditional Odds Ratios

- Consider $2 \times 2 \times K$ tables, where K denotes the number of levels of a control variable Z.
- Let $\{n_{ijk}\}$ denote the observed frequencies and let $\{\mu_{ijk}\}$ denote their expected frequencies.
- Within a fixed level k of Z,

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

  describes conditional $X - Y$ association.
- We refer to them as the $X - Y$ *conditional odds ratios*.

Marginal Odds Ratio

- Expected frequencies in the $X - Y$ marginal table is:

$$\mu_{ij+} = \sum_{k=1}^{K} \mu_{ijk}$$

- The $X - Y$ marginal odds ratio is defined as

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

- Similar formulas with μ_{ijk} substituted by n_{ijk}s provide sample estimates of $\theta_{XY(k)}$ and θ_{XY}.

Sample Odds Ratios

- Sample Conditional Odds Ratio:

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

- Sample $X - Y$ Marginal Odds Ratio:

$$\hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$$

Marginal vs. Conditional Independence

- If X and Y are independent in each partial table, then X and Y are said to be conditionally independent given Z.
- All conditional odds ratios between X and Y are then equal to 1.
- Conditional independence of X and Y, given Z, does not imply marginal independence of X and Y.
- That is, odds ratios between X and Y equal to 1 at each level of Z, the marginal odds ratio may differ from 1.

Homogeneous Association

- There is *homogeneous $X - Y$ association* in a $2 \times 2 \times K$ if the conditional odds ratios between X and Y are identical at all levels of Z.
- That is,$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$.
- In such a situation, a single number describes the conditional association.

Notes

- *Conditional independence* is a special case of homogeneous association, where each conditional odds ratio equals 1.0.
- *Homogeneous association* is a symmetric property, applying to any pair of variables viewed across the levels of the third.
- If it occurs, there is said to be *no interaction* between two variables in their effects on the third variable.

Cochran-Mantel-Haenszel Test

- To Test: X and Y are conditionally independent given Z.
- So, $H_0 : \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1.0$.
- In the k-th partial table, the row totals are n_{1+k}, n_{2+k} and column totals are n_{+1k}, n_{+2k}.
- Given both these totals, n_{11k} has a hypergeometric distribution and that determines all other cell counts in the k-th partial table.

Cochran-Mantel-Haenszel Test C Continued

- Under the null hypothesis of independence,

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n},$$

$$Var(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{{n_{++k}}^2(n_{++k}-1)}, k = 1, \cdots, K$$

- The test statistic is given by

$$CMH = \frac{[\sum_{k=1}^{K}(n_{11k} - \mu_{11k})]^2}{\sum_{k=1}^{K} Var(n_{11k})}$$

- This is called the $Cochran - Mantel - Haenszel$ (CMH) statistic.
- It has a large sample chi-squared distribution with $df = 1$.

Notes

- CMH takes larger values when $(n_{11k} - \mu_{11k})$ is consistently positive or consistently negative.
- This test is inappropriate when the association varies widely among the partial tables.

Estimation of Common Odds Ratio

- Assume, homogeneous association, that is,
  $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$.
- The Mantel-Haenszel estimator of that common value equals

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} (n_{11k} n_{22k} / n_{++k})}{\sum_{k=1}^{K} (n_{12k} n_{21k} / n_{++k})}$$

Standard Error

- The squared standard error for log of MH estimator is:

$$\hat{\sigma}^2(log(\hat{\theta}_{MH})) = \frac{\sum_{k=1}^{K}(n_{11k}+n_{22k})(n_{11k}n_{22k})/n_{++k}^2}{2(\sum_{k=1}^{K} n_{11k}n_{22k}/n_{++k})^2}$$

$$+ \frac{\sum_{k=1}^{K}[(n_{11k}+n_{22k})(n_{12k}n_{21k}) + (n_{12k}+n_{21k})(n_{11k}n_{22k})]/n_{++k}^2}{2(\sum_{k=1}^{K} n_{11k}n_{22k}/n_{++k})(\sum_{k=1}^{K} n_{12k}n_{21k}/n_{++k})}$$

$$+ \frac{\sum_{k=1}^{K}(n_{12k}+n_{21k})(n_{12k}n_{21k})/n_{++k}^2}{2(\sum_{k=1}^{K} n_{12k}n_{21k}/n_{++k})^2}$$

Testing Homogeneity of Odds Ratios

- To test for homogeneous association in $2 \times 2 \times K$ tables, $H_0 : \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$.
- The Breslow-Day test statistic has the form:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$$

  where $\hat{\mu}_{ijk}$ is the expected cell frequency.

- The formula for computing $\hat{\mu}_{ijk}$ is complicated.

Notes:

- Under null hypothesis, the Breslow-Day statistic has a large sample chi-squared distribution with degrees of freedom $K - 1$.
- The sample size should be relatively large in each partial table. .. It can be computed using standard statistical software.