Sections 2.3, 2.4

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

In describing a relationship between categorical variables X and Y, one should be aware of possible confounding between (X, Y) and another variable Z.

Your text gives an example on studying the relationship between spousal secondhand smoke and lung cancer. Obvious variables to control for (i.e. stratify by) are age, socioeconomic status, smoke exposure elsewhere, etc.

For now, let's assume Z is also categorical with K levels. The outcome is now n_{ijk}, the number out of n such that $(X, Y, Z) = (i, j, k)$.

Initially consider K different 2×2 partial tables of counts; for $Z = k$ we have

	$Y = 1$	$Y = 2$
$X = 1$	n_{11k}	n_{12k}
$X = 2$	n_{21k}	n_{22k}

Each table may have a different association between X and Y, perhaps estimated by $\hat{\theta}_{XY(k)} = n_{11k}n_{22k}/[n_{12k}n_{21k}]$, and this association will usually change with levels of $Z = k$.

The *marginal table* ignores the role of Z and collapses the table:

	$Y = 1$	$Y = 2$
$X = 1$	n_{11+}	n_{12+}
$X = 2$	n_{21+}	n_{22+}

The association in this marginal table may be similar to that observed for some levels of $Z = k$, *or not*.

Death penalty example

Here's a $2 \times 2 \times 2$ contingency table:

		Death penalty	No death penalty
White victim	White defendant	53	414
	Black defendant	11	37
Black victim	White defendant	0	16
	Black defendant	4	139

These data are $n = 674$ convicted murder cases in Florida from 1976 to 1987.

We are interested in the association between X = defendant race (black or white) and Y = death penalty (yes or no).

Death penalty example

Let's look at the table collapsed over the victim's race:

	Death penalty	No death penalty
White defendant	53	430
Black defendant	15	176

The probability of a death penalty is estimated to be $53/484 = 0.11$ versus 0.08 for white versus black defendants. Ignoring the victim's race leads us to believe that *whites* are more likely to get the death penalty.

However, when we stratify by the victim's race, these probabilities are 0.11 and 0.23 (white versus black defendants) for white victims and 0.00 and 0.03 for black victims. In both cases *black defendants* are more likely to be given the death penalty.

This illustrates the importance of *adjusting* for concomitant, often confounding variables (victim's race) that may be associated with both the response (death penalty) and a predictor (defendant's race).

Simpson's paradox

This is an example of *Simpson's paradox*, nicely illustrated in Figure 2.2. This happens because whites strongly tend to kill whites & although less strong, blacks tend to kill blacks. Let D =death penalty, D_w & D_b defendant, and V_b & V_w victim.

$$\begin{aligned} P(D|D_w) &= P(D|D_w, V_b)P(V_b|D_w) + P(D|D_w, V_w)P(V_w|D_w) \\ &\doteq 0.00(0.03) + 0.11(0.97) = 0.11. \end{aligned}$$

$$\begin{aligned} P(D|D_b) &= P(D|D_b, V_b)P(V_b|D_b) + P(D|D_b, V_w)P(V_w|D_b) \\ &\doteq 0.03(0.75) + 0.23(0.25) = 0.08. \end{aligned}$$

For example, $P(D|D_w)$ is a weighted average of $P(D|D_w, V_b)$ and $P(D|D_w, V_w)$ and the conditional weights $P(V_b|D_w) = 0.03$ and $P(V_w|D_w) = 0.97$ favor the larger value because *whites tend to kill whites*.

2.3.3 Conditional and marginal odds ratios

Consider a $2 \times 2 \times K$ table. Within a fixed level k of Z, let the odds ratio for (X, Y) be

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

where $\mu_{ijk} = n\pi_{ijk}$ is the expected cell number in the table. Of course these can be estimated by replacing expected frequencies by MLEs $\hat{\mu}_{ijk}$.

There are $k = 1, \ldots, K$ of these conditional odds ratios. The marginal odds ratio from the collapsed table is given by

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

Conditional odds ratios, cont.

Now consider $I \times J \times K$ tables. If

$$P(X = i, Y = j | Z = k) = P(X = i | Z = k) P(Y = j | Z = k),$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$ then X is (conditionally) independent of Y given $Z = k$. We write $X \perp Y | Z = k$.

If this holds for all $k = 1, \ldots, K$ then X is independent of Y given Z, $X \perp Y | Z$. This is equivalent to (2.9) on page 52.

Assuming a single multinomial applies to all counts $[n_{ijk}]_{I \times J \times K}$, independence implies that

$$\frac{\pi_{ijk}}{\pi_{++k}} = \frac{\pi_{i+k}}{\pi_{++k}} \times \frac{\pi_{+jk}}{\pi_{++k}},$$

or $\pi_{ijk}\pi_{++k} = \pi_{i+k}\pi_{+jk}$.

Conditional odds ratios, cont.

Note: Conditional independence does not imply marginal independence! That is, $X \perp Y|Z$ does not imply $X \perp Y$. This is true in general for any (X, Y, Z). For $2 \times 2 \times K$ tables, $X \perp Y|Z$ if and only if $\theta_{XY(k)} = 1$ for $k = 1, \ldots, K$; i.e. if the relative rates of success do not change with levels of Z.

Example: The following is a stratified table containing the (virtually always unknown) μ_{ijk} where $i = 1, 2$ indicates treatment, $j = 1, 2$ indicates outcome, and $k = 1, 2$ indicates clinic.

		Success	Failure
Clinic 1	Treatment A	18	12
	Treatment B	12	8
Clinic 2	Treatment A	2	8
	Treatment B	8	32

Here $\theta_{XY(1)} = \theta_{XY(2)} = 1$: X and Y are conditionally independent within a clinic. We conclude X and Y are not associated.

Conditional odds ratios, cont.

However, when we examine the marginal table

	Success	Failure
Treatment A	20	20
Treatment B	20	40

we obtain $\theta_{XY} = 2$, the odds of success are twice as great with treatment A instead of B.

What is happening here?

Loosely: Clinic 1 has a better overall success rate ($P(S|C_1) = 0.6$) than clinic 2 ($P(S|C_2) = 0.2$) – perhaps clinic 1 serves a more vital population. Also, clinic 1 tends to use treatment A more than B. So the collapsed results are weighted by clinic 1's more frequent use of A and better success rate.

Bottom line: it does not matter which treatment you receive, but you should try to get into clinic 1!

2.3.5 Homogeneous association

When $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$ the association between X and Y is the same for each fixed value of $Z = k$. This is called homogeneous association.

If additionally, $\theta_{XY(k)} = 1$ for each $Z = k$ then $X \perp Y | Z$.

Example: $X =$ smoking (yes, no), $Y =$lung cancer (yes, no), and $Z =$age (< 45, $45 - 65$, > 65). If $\theta_{XY(1)} = 1.2$, $\theta_{XY(2)} = 3.9$, and $\theta_{XY(3)} = 8.8$, then the association between smoking and lung cancer strengthens with age. X and Y are conditionally dependent on age Z.

The Cochran-Mantel-Haenszel statistics tests $H_0 : X \perp Y | Z$, coming up in Section 6.4.

2.4.1 Association in $I \times J$ tables

For 2×2 tables, θ summarizes the association between X and Y. For larger two-dimensional tables, i.e. $I \times J$ tables, we need to generalize this idea. There are $(I-1)(J-1)$ *local odds ratios*

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \text{ for } i = 1, \ldots, I-1 \text{ and } j = 1, \ldots, J-1.$$

θ_{ij} is relative odds of $Y = j$ versus $Y = j+1$ when $X = i$ versus $X = i+1$. All possible odds ratios each 2×2 table obtained from any two of the $\binom{I}{2}$ rows and any two $\binom{J}{2}$ columns from the $I \times J$ table can be obtained from the $(I-1)(J-1)$ local odds ratios $\{\theta_{ij}\}$.

For example, say $I = J = 3$. Then there are $(3-1)(3-1) = 4$ local odds ratios θ_{11}, θ_{12}, θ_{21}, θ_{22}.
Homework problem: Obtain

$$\theta = \frac{P(Y=3|X=2)/P(Y=1|X=2)}{P(Y=3|X=3)/P(Y=1|X=3)},$$

as a function of

$$\theta_{11} = \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)},$$

$$\theta_{12} = \frac{P(Y=2|X=1)/P(Y=3|X=1)}{P(Y=2|X=2)/P(Y=3|X=2)},$$

$$\theta_{21} = \frac{P(Y=1|X=2)/P(Y=2|X=2)}{P(Y=1|X=3)/P(Y=2|X=3)},$$

$$\theta_{22} = \frac{P(Y=2|X=2)/P(Y=3|X=2)}{P(Y=2|X=3)/P(Y=3|X=3)}.$$

Ordinal data

When X and Y are *ordinal*, we often examine odds of cumulative probabilities of the form

$$\theta = \frac{P(Y \leq 3|X \leq 1)/P(Y > 3|X \leq 1)}{P(Y \leq 3|X > 1)/P(Y > 3|X > 1)}.$$

For example if Y is the answer to "We should increase funding to public schools" (strongly disagree, disagree, ambivalent, agree, strongly agree) and X is education level (high school, undergraduate, graduate degree), this would be the odds of a random subject not agreeing for more money for schools given the subject has a high school versus these odds with a college degree.

Of these types of odds are the same across a table, $\theta = [P(Y \leq j|X \leq i)/P(Y > j|X \leq i)]/[P(Y \leq j|X > i)/P(Y > j|X > i)]$ for all i and j, then θ is termed a global odds ratio. It is a single number that summarizes association in an $I \times J$ table.

Another single statistic that summarizes association for ordinal (X, Y) uses the idea of concordant and discordant pairs. Consider data from the 2006 General Social Survey:

Age	Job satisfaction		
	Not satisfied	Fairly satisfied	Very satisfied
< 30	34	53	88
$30 - 50$	80	174	304
> 50	29	75	172

Clearly, job satisfaction tends to increase with age. How to summarize this association?

One measure of positive association is the probability of concordance.

Concordance

Consider two independent, randomly drawn individuals (X_1, Y_1) and (X_2, Y_2). This pair is concordant if either $X_1 < X_2$ and $Y_1 < Y_2$ simultaneously, or $X_1 > X_2$ and $Y_1 > Y_2$ simultaneously. An example would be (< 30 years, Fairly satisfied) and ($30 - 50$ years, Very satisfied). This pair indicates some measure of increased satisfaction with salary.

The probability of concordance Π_c is

$$\begin{aligned} P(X_2 > X_1, Y_2 > Y_1 \text{ or } X_2 < X_1, Y_2 < Y_1) &= P(X_2 > X_1, Y_2 > Y_1) \\ &\quad + P(X_2 < X_1, Y_2 < Y_1) \\ &= 2P(X_2 > X_1, Y_2 > Y_1) \end{aligned}$$

Using iterated expectation we can show

$$P(X_2 > X_1, Y_2 > Y_1) = \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} \left(\sum_{h=i+1}^{I} \sum_{k=j+1}^{J} \pi_{hk} \right).$$

2.4.5 γ statistic

Similarly, the probability of discordance is Π_d given by $2P(X_1 > X_2, Y_1 < Y_2)$. For pairs that are untied on both variables (i.e. they do not share the same age or satisfaction categories), the probability of concordance is $\Pi_c/(\Pi_c + \Pi_d)$ and the probability of discordance is $\Pi_d/(\Pi_c + \Pi_d)$. The difference in these is the *gamma* statistic

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

We have $-1 \leq \gamma \leq 1$. $\gamma = 1$ only if $\Pi_c = 1$, all pairs are perfectly concordant. Let C be the number of concordant pairs and D be

the number of discordant pairs; p. 57 shows how to count these. An estimator is $\hat{\gamma} = \frac{C-D}{C+D}$. For the job satisfaction data, $\hat{\gamma} = (C - D)/(C + D) = (99566 - 73943)/(99566 + 73943) = 0.15$, a weak, positive association between job satisfaction and age. Among untied pairs, the proportion of concordance is 15% greater than discordance.

Modifications to γ

- $\hat{\gamma}$ ignores ties. Kendall's $\hat{\tau}_b$ corrects for ties and has same interpretation; as with γ, $-1 \leq \tau_b \leq 1$.
- Stuart's $\hat{\tau}_c$ is Kendall's $\hat{\tau}_b$ corrected for sample size.
- Somer's D, $D(C|R)$ and $D(R|C)$ are assymetric versions of $\hat{\tau}_b$ looking at either the column variable or the row variable as the dependent outcome.

2.4.8 Polychoric correlation

Another measure of association for two ordinal variables.

For ordinal (X, Y), we can envision underlying *latent* continuous variables (Z_1, Z_2) that determine (X, Y) according to cutoffs.

$$X = i \Leftrightarrow \alpha_{i-1} < Z_1 < \alpha_i,$$

and

$$Y = j \Leftrightarrow \beta_{j-1} < Z_2 < \beta_j,$$

where

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{I-1} < \alpha_I = \infty,$$

and

$$-\infty = \beta_0 < \beta_1 < \cdots < \beta_{J-1} < \beta_J = \infty.$$

Polychoric correlation, cont.

Assume that (Z_1, Z_2) are bivariate normal $N_2(\mathbf{0}_2, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$; then there are $1 + (I - 1) + (J - 1)$ parameters to estimate: ρ, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{I-1})$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{J-1})$.

The parameter ρ is called the *polychoric correlation* between X and Y, and can be estimated via maximum likelihood.

A measure for nominal outcomes $\lambda_{Y|X}$

- Treat Y as an outcome and X as a predictor.
- If we ignore X, our best prediction for Y is the j with largest marginal π_{+j}, denoted $\pi_m = \max_{j=1,\dots,J}\{\pi_{+j}\}$. The error in this choice is $1 - \pi_m$.
- If we know $X = i$, then our best prediction for Y is the j with the largest π_{ij}, denoted $\pi_{c|i} = \max_{j=1,\dots,J}\{\pi_{ij}\}$. The error in this choice is $1 - \sum_{i=1}^{I} \pi_{c|i} = 1 - \pi_c$ when all the X's are weighted the same.
- The measure of association proposed by Goodman and Kruskall (1954) is the reduction in error when considering X in the prediction of Y vs. ignoring X:

$$\lambda_{Y|X} = \frac{(1-\pi_m)-(1-\pi_c)}{1-\pi_m} = \frac{\pi_c - \pi_m}{1-\pi_m}.$$

- SAS calls this $\lambda(C|R)$ or $\lambda(R|C)$ depending on whether the row or the column is the outcome variable.
- $\lambda_{Y|X}$ gives the proportion of error in predicting Y that can be eliminated by using a known value of X; $0 \le \lambda_{Y|X} \le 1$.