Analysis of Categorical Data

Three-Way Contingency Table

Outline

- Three way contingency tables
- Simpson's paradox
- Marginal vs. conditional independence
- Homogeneous association
- Cochran-Mantel-Haenszel Methods

Three-Way Contingency Tables

- Partial Tables
 - Make 2-way tables of $X \times Y$ at variaous levels of Z. This effectively removes the effect of Z by holding it constant.
 - The associations of partial tables are called *conditional* associations because we are looking at X and Y conditional on a fixed level of Z.
 - Focus is on relationship between variables X and Y at fixed levels of another variable $Z = 1, \ldots, K$.
- Marginal Tables
 - Sum the counts from the same cell location of partial tables.
 The idea is to form an X, Y table by summing over Z.
 - Marginal tables can be quite misleading: Simpson's Paradox.

Simpson's Paradox: Example 1

Table 1: Admission to Graduate School (Verducci)

		Accepted	Rejected	_
Science	Male	60	15	
	Female	25	5	

		Accepted	Rejected
Arts	Male	10	15
	Female	30	40

- X = Sex: Male, Female
- Y = Admission: Accepted, Rejected
 - Z = College: Science, Arts

Example 1 (Cont'd)

• Condition on Z.

- $\diamond O_{XY(Sci)} = 4/5 < 1$
- $\bullet \quad O_{XY(Art)} = 8/9 < 1$
- $O_{XY} = 21/11 > 1$

Condition on X.

- $\diamond \ O_{ZY(M)} = 6$
- $\diamond O_{ZY(F)} = 20/3$
- $\diamond \quad O_{ZY} = 187/12$
- **Condition on** Y.
 - $\diamond O_{ZX(Acc)} = 36/5$
 - $\bullet \ O_{ZX(Rej)} = 8$

$$\bullet \ O_{ZX} = 7$$

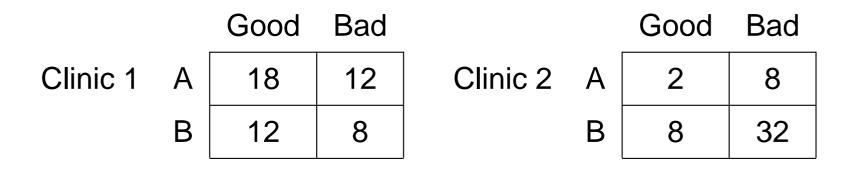
Simpson's Paradox (Cont'd)

Paradox

- In each College, women have a greater acceptance rate than do men;
- Overall, men have a greater acceptance rate than do women;
- Resolution
 - The sciences have a much higher acceptance rate than do the arts
 - Most men apply to sciences; women to arts
 - Simpson's paradox happens when there are different associations in partial and marginal tables.

Marginal vs. Conditional Independence

- If X and Y are independent in each partial table, controlling for Z, then X and Y are conditionally independent.
- If X and Y are conditionally independent at each level of Z, but may still not be marginally independent
- Example 2 : Clinic and Treatment



Example 2 (Cont'd)

Condition on Z.

- $\bullet \ O_{XY(C1)} = 1$
- $\diamond \quad O_{XY(C2)} = 1$
- $\bullet \ O_{XY} = 2$

Condition on X.

- $\diamond \ O_{ZY(A)} = 6$
- $\diamond \ O_{ZY(B)} = 6$
- $\bullet \ O_{ZY} = 6$

Condition on Y.

- $\diamond \ O_{ZX(Good)} = 6$
- $\diamond \ O_{ZX(Bad)} = 6$

$$\bullet \ O_{ZX} = 6$$

Example 2 (Summary)

X and Y are conditionally independent at each level of Z, but they are not marginally independent. This happens because, across levels of Z,

- there is a reversal in the odds of success:
 - 3:2 in Clinic 1
 - 1:4 in Clinic 2
- There is a reversal in prevalence of treatment:
 - Clinic 1 uses Treatment A the most
 - Clinic 2 uses Treatment B the most

Homogeneous Association

- Effect of X on Y is the same at all levels of Z.
- Happens when the conditional odds ratio using any two levels of X and any two levels of Y is the same at all levels of Z:

$$O_{XY(1)} = \ldots = O_{XY(K)}$$

- Conditional Independence is a special case, when these all equal 1.
- In the case when K=2, homogeneous association implies that the other conditional odds ratios will also be the same:

$$O_{ZY(1)} = O_{ZY(2)}$$
 and $O_{ZX(1)} = O_{ZX(2)}$

For 3-way tables of larger dimensions, homogeneous association generalizes to the model of no-three way interaction.

Example 3: Bipoloar Children Trtment

- 200 families with a bipolar child
 - 100 randomized to immediate "treatment"
 - 100 randomized to 1-year waitlist
- Outcome Variable: Social functioning at one year into the study
 - 100 good and 100 bad
- Moderating Variable: Both biological parents as caregivers
 - 100 Yes and 100 No

		Good	Bad			Good	Bad
Intact	imm	60	20	Not Intact	imm	15	5
Family	wait	10	10	Family	wait	40	40

Example 3 (Cont'd)

Condition on Z.

- $\diamond O_{XY(IF)} = 3$
- $\diamond \ O_{XY(NIF)} = 3$
- $\diamond O_{XY} = 3$

Condition on X.

- $\diamond O_{ZY(imm)} = 1$
- $\bullet \ O_{ZY(wait)} = 1$
- $\bullet \quad O_{ZY} = 1.9$

Condition on Y.

- $\diamond O_{ZX(Good)} = 16$
- $\blacklozenge \ O_{ZX(Bad)} = 16$

$$\bullet \quad O_{ZX} = 16$$

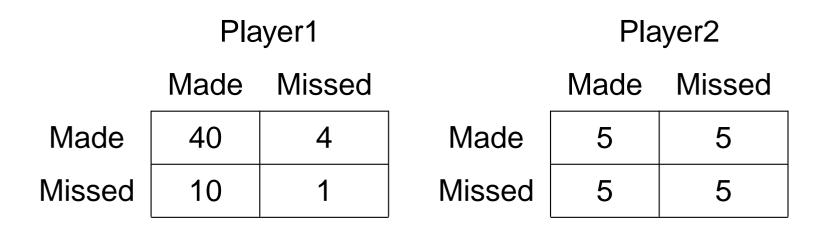
CMH Test

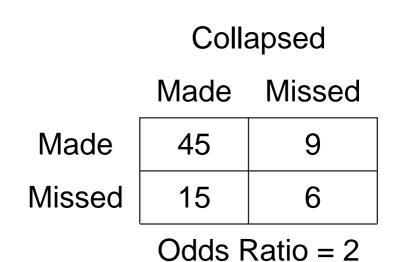
Motivation: Is there an association between X and Y?

- Can't just collapse table [why not?]
- Assume there is a common odds ratio θ at each level of Z
- Hypotheses
 - Null hypothesis $H_0: \theta = 1$
 - Alternative hypothesis $H_1: \theta < 1$ or $\theta > 1$
- Evidence
 - Condition on the margins of XY table at each level of Z
 - Only need to consider one entry n_{11k} at level k of Z
 (k = 1,...,K)
 - Under the null hypothesis, {n_{11k}} are independent
 hypergeometric random variables

Why Not?

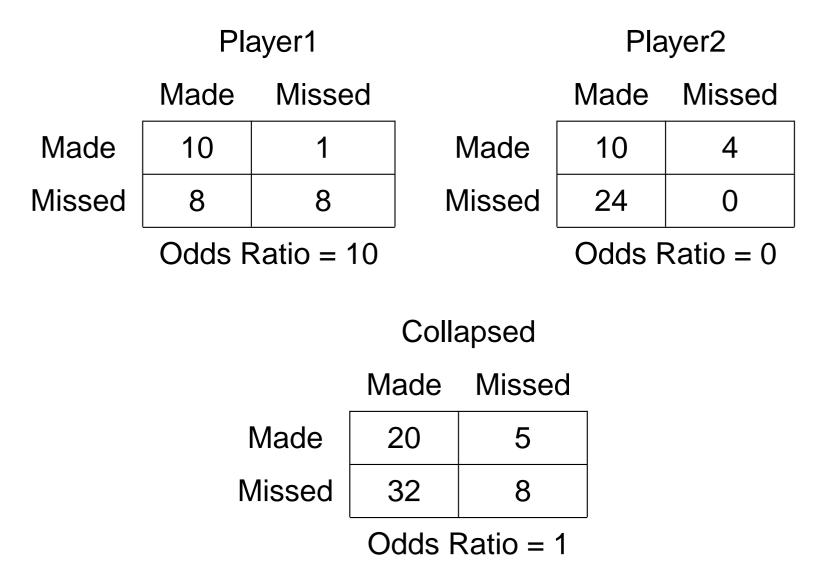
Could wrongly find association: Example 4





Example 4 (Cont'd): Why Not?

Could wrongly mistake diverse association for no association



CMH Test

Under the null hypothesis, $\{n_{11k}\}$ are independent hypergeometric random variables

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$
$$Var(n_{11k}) = \frac{n_{1+k}n_{1+k}n_{+1k}n_{+1k}}{n_{++k}^2(n_{++k}-1)}$$

CMH Test Statistics

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

- Important: In the numerator, sum before squaring
- Under the null hypothesis $CMH \sim \chi_1^2$

CMH Test (Cont'd)

- The CMH test is a powerful summary of evidence against the hypothesis of conditional independence, as long as the sample associations fall primarily in a single direction.
- Mantel-Haenszel Estimator for Common Odds Ratio

$$\hat{\theta}_{MH} = \frac{\sum_{k} \left(\frac{n_{11k}n_{22k}}{n_{++k}}\right)}{\sum_{k} \left(\frac{n_{12k}n_{21k}}{n_{++k}}\right)}$$

Example 5: Coronary Artery Disease