CATEGORICAL DATA ANALYSIS I: CROSS TABS TESTS OF ASSOCIATION THE ODDS RATIO

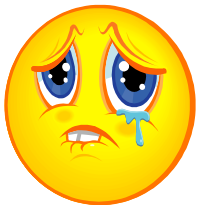
TABLES AND CROSSTABS

WHICH APPROACH?

	Predictors		
Response	**Categorical**	**Continuous**	**Categorical and Continuous**
Continuous	ANOVA	Linear Reg	ANCOVA (really just regression with dummy variables)
Categorical	Contingency Table Analysis (Logistic)	Logistic	Logistic

		
	72%	28%
	72%	28%

Can't plot binary variables (like can with continuous)
Instead, set up rows and columns, look for associations

		
	82%	18%
	60%	40%

Frequency Tables

A frequency table shows the number of observations that fall in certain categories or intervals. A one-way frequency table examines one variable.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

PROC FREQ
tables var1

Crosstabulation Tables

A *crosstabulation* table shows the number of observations for each combination of the row and column variables.

	column 1	column 2	…	column c
row 1	$cell_{11}$	$cell_{12}$	…	$cell_{1c}$
row 2	$cell_{21}$	$cell_{22}$	…	$cell_{2c}$
…	…	…	…	…
row r	$cell_{r1}$	$cell_{r2}$	…	$cell_{rc}$

```
PROC FREQ
   tables var1 * var2
```

Notes on Cross Tab Output

- proc freq good way to catch miscoding
- Default 4 # each cell (freq, % tot, row %, col %)
- Look if row %'s vary by column category
  - E.g. 68% men vs. 58% women spent < $100
  - 48% high income vs 32% low income spent > $100
- Default alphanumeric sorting
  - e.g. 'high, low, medium" – needs to be addressed

TESTS OF ASSOCIATION

CHISQ

- $\chi 2 = \sum (O-E)^2/E$
  - underlying assumption of normality
- where E = row total * column total / grand total
  - derived from (column total / grand total)* row total
- Ho: observed = expected
  - p-value for a $\chi 2$ = prob of observing test statistic at least as large given no association
- Chisq – for general associations between nominal variables
- MH chisq – for ordinal variables (takes order into account)

Notes about CHISQ

- NOT a measure of the strength of an association
- for *frequencies*, not percentages or proportions
- Cochrane's Rule: 80% > 5, no cells <1
  - - for a 2x2 table, *none of the cells should have expected counts less than 5*
- option after TABLES statement
  - *TABLES var_a*var_b / CHISQ*
  - Options after CHISQ e.g. *expected nocol nopercent*

Fisher's Exact Test

- 2x2 tables with small cells
- $r_1!\ r_2!\ c_1!\ c_2!\ /\ n!\ f_{11}!\ f_{12}!\ f_{21}!\ f_{22}!$
  - Where r and c are the row and column totals; n is the grand total, f is the cell frequency
- 4 Steps:
  - lay out the table
  - lay out all the more extreme tables
  - calculate the exact probability for each small frequency 2x2 table
  - Sum all of these probabilities
- in SAS:
  - Pearson Chi-Square (PCHI)
  - EXACT statement for tables

Exact *p*-Values for Pearson Chi-Square

Small values, asymptotic chi square not appropriate

Observed Table

0	3	3
2	2	4
2	5	

Exact *p*-Values for Pearson Chi-Square

Observed Table		
0	3	3
2	2	4
2	5	

$\chi^2 = 2.100$
prob = .286

Possible Table 1		
1	2	3
1	3	4
2	5	

$\chi^2 = 0.058$
prob = .571

Possible Table 2		
2	1	3
0	4	4
2	5	

$\chi^2 = 3.733$
prob = .143

Assume column and row values fixed
Calculate probability for each table

Detecting Ordinal Associations: The Mantel-Haenszel $\chi 2$

- ***assumes a linear association***
  - **looks at the *average trend***
  - **if not a linear trend, can give you the wrong answer**
  - **again, *not a measure of the strength of the association.***

- **different from Cochran-Mantel Haenszel test for stratified 2x2 tables**

- **Mean Score Statistic - Another test statistic available in SAS, recommended for nominal by ordinal associations**

Spearman Correlation

- *strength of an association*
- uses ranks of data
- range betw -1 and 1
- MEASURES option under PROC FREQ
  - CL give confidence limits

THE ODDS RATIO

Probabilities and Odds:

- Probability
  - 0 to 1
  - ratio number chances to the total chances possible.
- Odds
  - number of chances for (or against) versus the number of chances against (or for).
- convert odds to probability
  - sum the odds to represent the total number of chances of an event occurring
  - probability = odds / 1+odds
- To convert a probability to an odds
  - remove the event from the total chances and set them against each other as a ratio
  - odds = probability / 1-probability

Why do we need an Odds Ratio?

- Cohort studies: P[disease | exposure]
- Case-control : P[exposure | disease]
- e.g. Lung Cancer Example
  - Cohort studies: P[disease | exposure] = P[lung Cancer | smoking]
  - Case-control : P[exposure | disease]= P[smoking | lung Cancer]

Lung Cancer				
Smoking		Yes	No	
	Yes	100	900	1000
	No	50	1950	2000
		150	2850	3000

The Disease Odds Ratio

- P[disease | exposure] / P[no disease | exposure] = P [disease | exposure] / 1- P[disease | exposure] = (a/ a+b) / (b/a+b) = (100/1000) / (900/1000) = .1/.9 = 0.11
-
- Can similarly get the odds of disease given no exposure = (c/c+d) / (d / c + d) = (50/2000) / (1950/2000) = .025/0.927 = 0.026
-
- OR = 0.11/0.026 = 4.3 (very close to our RR, would be closer the rarer the outcome)
-
- So odds ratio = [(a/a+b) / (b/a+b)] / [(c/c+d) / (d / c + d)] = (a/b) / (c/d) = ad/bc

The Exposure Odds Ratio

- **Case-Control can't get the odds of D given E vs. no E…**
- ***can get…odds of E given D vs. odds E given no D***
  - **odds E given D (= P[E | D] / P[no E | D] = P[E | D] / 1 - P[E | E]**
    - *= (a/a+c) / (c/a+c) = (100/150) / (50/150) = 2*
  - **odds E given no D = P[E | no D] / P[no E |no D] = [E | no D] / 1 - P[E | no D]**
    - *= (b/b+d) / (d/d+c) = (900/2850) / (1950/2850) = 0.46*
- **OR = 2/0.46 = 4.3**
  - **invariant across design type**
  - **can approximate prospective results, from retrospective data**
  - **a very cool thing.**
- **MEASURES under PROC FREQ. For a 2x2 table will also give you the Odds Ratio along with its confidence limits**

DEMONSTRATIONS: CROSSTABS, REORDERING, FISHER'S , SPEARMAN'S, ODDS RATIO

The Expenditure Data Set

- Outcome = healthcare expenditure
  - 1= high service users ($100 or more)
  - 0= low service users (LT $100)
- Predictors = Gender, Income (SES), Age
- Types of Variables
  - Nominal, Ordinal, Interval Ratio
  - Gender? 3-Level SES?

Pre-term Labor ?➔LBW Data Set

- 189 Obs
  - low='Indicator for Birth Weight‘
  - mother_age='Mother"s age‘
  - mother_wt='Weight at Last Menstrual Period‘
  - socio='Socio-Economic Status‘
  - alcohol='Did the mother drink during pregnancy?‘
  - hist_hyp='History of Hypertension‘
  - prev_pretrm='Previous Preterm Labors‘
  - uterine_irr='Uterine Irritability‘
  - phy_visit='Physician Visit in 1st Trimester';

PUTTING IT TOGETHER: THE EXPENDITURE DATA SET