Contingency Tables

Conditional Probabilities and Proportions

Two by Two Tables

Do you use the internet to buy electronic equipment?

	Yes	No	Total
Females	20	80	100
Males	45	55	100

Proportions or Conditional Probabilities

$$\hat{p}_F = .20, \hat{p}_M = .45$$

$$Ho: p_F = p_M$$
$$Ha: p_F \neq p_M$$

Independence

- Our question can be restated as whether
  - P(yes | female) is equal to P(yes | male)
- This is the same as saying that
  - P(yes | female) = P(yes) or that
  - P(yes | male)= P(yes)
- The question then is whether the probability of answering yes is affected by gender.

Significance Test of Proportions

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\hat{\sigma}_{(p_1 - p_2)}}$$

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)}$$

Example

$$z = \frac{.45 - .20}{\sqrt{.325(1-.325)\left(\frac{1}{100} + \frac{1}{100}\right)}}$$

$$z = \frac{.25}{.0662} = 3.77$$

The Hypotheses

- Ho: Gender and Buying behavior are independent. (Rejected)
- Ha: Gender and Buying behavior are dependent

Normal Approximation to the Binomial

- The significance test is based on the Binomial Distribution can be approximated with the normal when n is large and p is not too small.
- Binomial: mean=n*p
  - variance= n*p*(1-p)

Another Approach

- Because of the relationship between the normal and the chi-squared distribution, we could have used the chi-squared table to evaluate the results

$$z^2 = \chi^2$$

Chi-Squared Distribution

- Skewed
- Mean= df (degrees of freedom)
- Variance= 2*df
- For independent z's:

$$\chi^2_{df} = \sum_{i=1}^{df} z_i^2$$

Another Example

Gender	Democrat	Independent	Republican	Total
Female	279	73	225	577
Male	165	47	191	403

Conditional Probabilities

Gender	Democrat	Independent	Republican	Total
Female	.483	.127	.390	1.00
Male	.409	.117	.474	1.00

Hypotheses

- Ho: the two conditional distributions are equal.
- Ha: the two conditional distributions are not equal.

Chi-square Test

- Because the second example involves a 2 by 4 contingency table, we can no longer use the z-test. Instead, we must use the chi-square test:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Expected Frequencies

- The chi-square test involves a comparison of the expected frequencies (under independence) and the observed.
- f_e = N* $P(R_i)P(C_j)$=(row total * col total)/n
- f_o= observed frequency in the cell
- df = (# of rows -1)*(# of cols-1)

Computation

$$\chi^2 = \frac{(279-261.4)^2}{261.4} + \ldots + \frac{(191-171.1)^2}{171.1}$$
$$= 1.19 + \ldots + 2.31$$
$$= 7.0$$

$$f_e = \left(\frac{444}{980}\right)577 = 261.4$$

Critical Value

- Looking at the Chi-squared table, we find that (for df=2, and alpha=.05) the chi-squared value is 5.99.
- Because the observed value is larger than the critical value, we reject the Null.
- Party affiliation is dependent of gender.

Party Identification

```
data party;
input gender $ party $ count @@; cards;
f dem 279 f ind 73 f rep 225
m dem 165 m ind 47 m rep 191
;
proc freq; weight count;
tables gender*party / chisq measures
expected; run;
```

Statistics for Table of gender by party

Statistic	DF	Value	Prob
Chi-Square	2	7.0095	0.0301
Likelihood Ratio Chi-Square	2	7.0026	0.0302
Mantel-Haenszel Chi-Square	1	6.7581	0.0093
Phi Coefficient		0.0846	
Contingency Coefficient		0.0843	
Cramer's V		0.0846	

Association

Statistic	Value	ASE
Gamma	0.1470	0.0559
Kendall's Tau-b	0.0796	0.0306
Stuart's Tau-c	0.0858	0.0330

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault
Stranger	12	379	727
Acquaintance or Relative	39	106	642

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12	379	727	1118
Acquaintance or Relative	39	106	642	787
Column Total	51	485	1369	1905

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12	379	727	1118
Acquaintance or Relative	39	106	642	787
Column Total	51	485	1369	1905

$$E = \frac{\text{(row total) (column total)}}{\text{(grand total)}}$$

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12 (29.93)	379	727	1118
Acquaintance or Relative	39	106	642	787
Column Total	51	485	1369	1905

$$E = \frac{\text{(row total) (column total)}}{\text{(grand total)}}$$

$$E = \frac{(1118)(51)}{1905} = 29.93$$

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12 (29.93)	379 (284.64)	727 (803.43)	1118
Acquaintance or Relative	39 (21.07)	106 (200.36)	642 (565.57)	787
Column Total	51	485	1369	1905

$$E = \frac{\text{(row total) (column total)}}{\text{(grand total)}}$$

$$E = \frac{(1118)(51)}{1905} = 29.93 \qquad E = \frac{(1118)(485)}{1905} = 284.64$$

etc.

Is the type of crime independent of whether the criminal is a stranger?

$$X^2 = \sum \frac{(O - E)^2}{E}$$

	Homicide	Robbery	Assault
Stranger	12 (29.93) [10.741]	379 (284.64) [31.281]	727 (803.43) [7.271]
Acquaintance or Relative	39 (21.07) [15.258]	106 (200.36) [44.439]	642 (565.57) [10.329]

(E) $\left[\frac{(O - E)^2}{E}\right]$

Upper left cell: $\frac{(O - E)^2}{E} = \frac{(12 - 29.93)^2}{29.93} = 10.741$

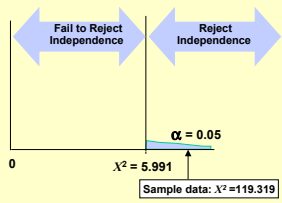
Test Statistic $X^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

Test Statistic $X^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

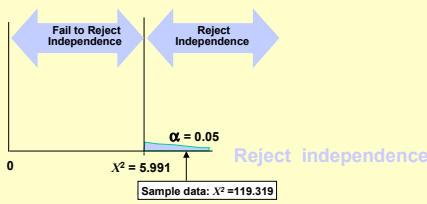
Critical Value $X^2 = 5.991$



Test Statistic $X^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

Critical Value $X^2 = 5.991$



Reject independence

Requirements for the Chi-squared tests

- The test requires random sample or a stratified random samples. The samples should be large.
- Stratified sampling– the population is divided into strata (groups) that share the same characteristics, then we draw a random sample from each stratum.

Assumptions

- Two categorical variables
- fe $\geq$ 5 (for all cells)
- We use the right tail probability from the chi-squared table
- No repeated measures

Class Data: Do you the internet to buy electronic equipment?

Gender	Yes	No
Female	1	14
Male	8	4

SAS Setup for the Class Data

```
Data Class; input buy $ sex $ count;
cards;
y f 1
n f 14
y m 8
n m 4
;
Proc freq; weight count;
tables sex*buy / exact chisq measures expected;
run;
```

Association

- Like other statistical tests, the significance of chi-squared test depends on the size of the sample.
- So, we must ask: How strong is the association?
- When both of the variables are nominal:
  - We can compare the conditional probability distributions.
  - we can look at the odds ratio.

Odds of yes, for the class data

- Odds= P(success) / P(failure)
- P(Yes)=9/27=1/3
- P(No)=2/3
- Odds=(1/3)/(2/3)=1/2=.5

Odds of No

- Odds=P(No)/(1-P(no))
- Or, Odds=P(No)/P(yes)=(2/3)/(1/3)=2/1=2
- A No is twice as likely as Yes

Odds Ratio

- Females Odds of yes:
  - Odds=(1/15)/(14/15)=1/14
  - We expect one YES for every 14 NO's
- Males odds of yes
  - Odds=(8/12)/(4/12)=8/4=2
  - We expect two YESes for every NO
- Odds ratio= males / females = (2/1)/(1/14)=28
- The male odds (for saying yes) are 28 times the female odds.

Gender	Yes	No
Female	1	14
Male	8	4

Females: the odds of yes are 1 to 14, or .071 to 1.

Males: the odds of yes are 8 to 4, or 2 to 1.

Class Data

Physician's Health Aspirin Study, Another Example

	Aspirin	Placebo	Total
Heart Attack	139	239	**378**
Stroke	119	98	**217**
Healthy	10,779	10,697	**21,476**
Total	**11,037**	**11,034**	**22,071**

Chi-squared Test

```
Chi-Square = 28.8000, p <.0001
```

Risk and Odds for Placebo

- Risk and odds of heart attack for individuals in the placebo group:
  - Risk=239/11,034=.0217
    - (2.17% of the placebo group had a heart attack)
  - Odds=239/10,795=.0221 (to one)
    - (Or 1 to 45.25, if you are a physician for every 46 persons that you treat one is likely to have a heart attack)

Heart attack and Aspirin

- Risk and Odds of heart attack for individuals taking aspirin:
  - Risk=139/11,037=.0125%
    - (1.25% of the aspirin group had a heart attack)
  - Odds=139/10,898=.0128 (to one)
    - (Or 1 to 78.125, if you are a physician for every 79 people that you treat with aspiring one is likely to have a heart attack)

Odds Ratio

- Odds ratio (of having a heart attack if you take aspirin)=.0128/.0221=.5791
- The odds of having a heart attack are reduced by .58 when you take aspirin.
- Or, the odds of having a heart attack when you do not take aspirin are 1.73 times the odds of having a heart attack when you take aspirin.

Relative Risk

- Relative Risk (of heart attack for aspirin takers):
  - Relative Risk=.0125/.0217=.576
  - (The risk of a heart attack is reduced by about 57% when you take aspirin.)

Handling Ordinal Variables

- Ordinal variables- data can be arranged in some order, but differences between data points either can not be determined or are meaningless.

Contingency Tables in which both variables are ordinal

- Compute a measures of association, then check whether it is significant. (We could compute a chi-square, but it would be less powerful.)
- Measures:
  - Gamma
  - Kendall's Tau b
  - Sommer's d (assumes an response and explanatory variables)

Gamma

- Gamma is based only on the number of concordant and discordant pairs. It ignores tied pairs. If the variables are independent then gamma is close to zero. Gamma is appropriate only for ordinal variables. It ranges from –1 to 1.

Tau- b

- Tau-b: it is similar to gamma except that uses a correction for ties. Tau-b is only appropriate for ordinal variables. It ranges from –1 to 1. It is more stable than gamma under different categorizations.

Gamma test

- When the two ordinal variables are independent and n is large (the number of concordant and discordant pairs are each greater than 50).
- $$z = \frac{\hat{\gamma}}{\hat{\sigma}}$$

Job Satisfaction, an example from the book

	Dissatisfied	Moderately Satisfied	Very Satisfied
<$5000	6	13	3
5000 to 25000	9	37	12
>25000	3	13	8

Low-High on the Left Upper Corner

```
data satis; input income $ sat $ count
@@;cards;
h d 3 h ms 13 h vs 8
m d 9 m ms 37 m vs 12
l d 6 l ms 13 l vs 3
;
proc freq order=data; weight count;
tables income*sat / chisq measures;
run;
```

Gamma=-0.2873, ASE=0.1506, chisq=4.1

Low-Low on the Left Upper Corner

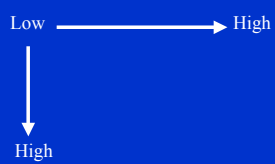
```
data satis; input income $ sat $ count
@@;cards;
l d 6 l ms 13 l vs 3
m d 9 m ms 37 m vs 12
h d 3 h ms 13 h vs 8
;
proc freq order=data; weight count;
tables income*sat / chisq measures;run;
```

Gamma=0.2873, ASE=0.1506, chisq=4.1

Z=.2873/.1506=1.91

Notice the difference in the gamma values.

Proper order for Ordinal Data



Discordant and Concordant

- A pair of observations is concordant if the subject who is higher on one variable also is higher on the other variable.
- A pair of observations is discordant if the subject who is higher in one variable is lower on the other variable.

Example, 10 concordant pairs

	Low	High
Low	2	3 concordant
High	4	5

12 discordant pairs

	Low	High
Low	2	3
High	4	Discordant 5

Gamma

- C=total number of concordant pairs
- D=total number of discordant pairs
- Gamma=(C-D)/(C+D)

Gamma=(10-12)/(10+12)=-2/22=-.09
ASE is much more difficult to obtain.

Analysis of mixed ordinal-nominal tables

- If the nominal variable has only two categories, then you can use a gamma.
- If the nominal variable has more than two categories and the ordinal variable has a few categories, then you can use ANOVA.

The End

Chapter 8