Contingency Tables
Conditional Probabilities and Proportions

Two by Two Tables
Do you use the internet to buy electronic equipment?

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>Males</td>
<td>45</td>
<td>55</td>
<td>100</td>
</tr>
</tbody>
</table>

Proportions or Conditional Probabilities

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

\[ H_0 : p_F = p_M \]

\[ H_a : p_F \neq p_M \]

Independence

- Our question can be restated as whether
  - \( P(\text{yes} | \text{female}) \) is equal to \( P(\text{yes} | \text{male}) \)
- This is the same as saying that
  - \( P(\text{yes} | \text{female}) = P(\text{yes}) \) or that
  - \( P(\text{yes} | \text{male}) = P(\text{yes}) \)
- 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

<table>
<thead>
<tr>
<th>Gender</th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>279</td>
<td>73</td>
<td>225</td>
<td>577</td>
</tr>
<tr>
<td>Male</td>
<td>165</td>
<td>47</td>
<td>191</td>
<td>403</td>
</tr>
</tbody>
</table>

Conditional Probabilities

<table>
<thead>
<tr>
<th>Gender</th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>.483</td>
<td>.127</td>
<td>.390</td>
<td>1.00</td>
</tr>
<tr>
<td>Male</td>
<td>.409</td>
<td>.117</td>
<td>.474</td>
<td>1.00</td>
</tr>
</tbody>
</table>
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.
- \( f_e = N \times P(R_i)P(C_j) = \text{(row total} \times \text{col total)}/n \)
- \( f_o = \text{observed frequency in the cell} \)
- \( df = \text{(# of rows -1) x (# of cols-1)} \)

Computation

\[ \chi^2 = \frac{(279 - 261.4)^2}{261.4} + \ldots + \frac{(191 - 171.1)^2}{171.1} \]

\[ = 1.19 + \ldots + 2.31 \]

\[ = 7.0 \]

\[ f_e = \left( \frac{444}{980} \right) \times 377 = 261.4 \]

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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>2</td>
<td>7.0095</td>
<td>0.0301</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>2</td>
<td>7.0026</td>
<td>0.0302</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>6.7581</td>
<td>0.0093</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.0846</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.0843</td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.0846</td>
<td></td>
</tr>
</tbody>
</table>

Association

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.1470</td>
<td>0.0559</td>
</tr>
<tr>
<td>Kendall’s Tau-b</td>
<td>0.0796</td>
<td>0.0306</td>
</tr>
<tr>
<td>Stuart’s Tau-c</td>
<td>0.0858</td>
<td>0.0330</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Homicide</th>
<th>Robbery</th>
<th>Assault</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stranger</td>
<td>12</td>
<td>379</td>
<td>727</td>
<td>1118</td>
</tr>
<tr>
<td>Acquaintance</td>
<td>39</td>
<td>106</td>
<td>642</td>
<td>51</td>
</tr>
<tr>
<td>Column Total</td>
<td>51</td>
<td>485</td>
<td>1369</td>
<td>1905</td>
</tr>
</tbody>
</table>

Row Total

Column Total

\[ E = \frac{(\text{row total}) \times (\text{column total})}{(\text{grand total})} \]

\[ E = \frac{(1118)(51)}{1905} = 29.93 \]
Is the type of crime independent of whether the criminal is a stranger?

<table>
<thead>
<tr>
<th></th>
<th>Homicide</th>
<th>Robbery</th>
<th>Assault</th>
<th>Row Total</th>
</tr>
</thead>
</table>
| Stranger | 12  
(29.93) | 379  
(208.64) | 727  
(603.43) | 1118 |
| Acquaintance or Relative | 29  
(21.07) | 156  
(205.36) | 642  
(965.57) | 727 |
| Column Total | 51  
(86.48) | 545  
(651.32) | 1389  
(1578.97) | 1905 |

\[ E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} \]

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

\[ E = \frac{(1118)(485)}{1905} = 284.64 \]

\[ \text{etc.} \]

Test Statistic \( \chi^2 = 119.319 \)

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

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?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>Male</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

SAS Setup for the Class Data

```sas
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 = \( P(\text{success}) / P(\text{failure}) \)
- \( P(\text{Yes}) = 9/27 = 1/3 \)
- \( P(\text{No}) = 2/3 \)
- \( \text{Odds} = (1/3)/(2/3) = 1/2 = .5 \)

Odds of No

- \( \text{Odds} = P(\text{No})/(1-P(\text{no})) \)
- Or, \( \text{Odds} = P(\text{No})/P(\text{yes}) = (2/3)/(1/3) = 2/1 = 2 \)
- 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)=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.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>Male</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

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.

Physician’s Health Aspirin Study, Another Example

<table>
<thead>
<tr>
<th></th>
<th>Aspirin</th>
<th>Placebo</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Attack</td>
<td>139</td>
<td>239</td>
<td>378</td>
</tr>
<tr>
<td>Stroke</td>
<td>119</td>
<td>98</td>
<td>217</td>
</tr>
<tr>
<td>Healthy</td>
<td>10,779</td>
<td>10,697</td>
<td>21,476</td>
</tr>
<tr>
<td>Total</td>
<td>11,037</td>
<td>11,034</td>
<td>22,071</td>
</tr>
</tbody>
</table>

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).

- \[ Z = \frac{\hat{\gamma}}{\sigma} \]

---

**Job Satisfaction, an example from the book**

<table>
<thead>
<tr>
<th></th>
<th>Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$5000</td>
<td>6</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>5000 to 25000</td>
<td>9</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>&gt;25000</td>
<td>3</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

---

**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`

\[ Z = \frac{0.2873}{0.1506} = 1.91 \]

Notice the difference in the gamma values.

---

**Proper order for Ordinal Data**

- **Low** \( \rightarrow \) **High**

---

**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

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>High</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

12 discordant pairs

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>High</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

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 = -0.09

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