ST3241 Categorical Data Analysis I
Two-way Contingency Tables

Odds Ratio and Tests of Independence

Inference For Odds Ratio (p. 24)

- For small to moderate sample size, the distribution of sample odds ratio $\hat{\theta}$ is highly skewed.
- For $\theta = 1$, $\hat{\theta}$ cannot be much smaller that θ, but it can be much larger than θ with nonnegligible probability.
- Consider log odds ratio, $\log \theta$
- X and Y are independent implies $\log \theta = 0$.

Log Odds Ratio

- Log odds ratio is symmetric about zero in the sense that reversal of rows or reversal of columns changes its sign only.
- The sample log odds ratio, $\log \hat{\theta}$ has a less skewed distribution and can be approximated by the normal distribution well.
- The asymptotic standard error of $\log \hat{\theta}$ is given by

$$ASE(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Confidence Intervals

- A large sample confidence interval for $\log\theta$ is given by

$$\log(\hat{\theta}) \pm z_{\alpha/2} ASE(\log\hat{\theta})$$

- A large sample confidence interval for θ is given by

$$\exp[\log(\hat{\theta}) \pm z_{\alpha/2} ASE(\log\hat{\theta})]$$

Example: Aspirin Usage

- Sample Odds Ratio $= 1.832$
- Sample log odds ratio, $\log \hat{\theta} = \log(1.832) = \cancel{0.2629}$ 0.605
- ASE of $\log \hat{\theta}$

$$\sqrt{\frac{1}{89} + \frac{1}{10933} + \frac{1}{10845} + \frac{1}{104}} = 0.123$$

- 95% confidence interval for $\log \theta$ equals

$$0.605 \pm 1.96 \times 0.123$$

- The corresponding confidence interval for θ is $(e^{0.365}, e^{0.846})$ or (1.44,2.33).

Recall SAS Output

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence	Limits
Case-Control (Odds Ratio)	1.8321	1.4400	2.3308
Cohort (Col1 Risk)	1.8178	1.4330	2.3059
Cohort (Col2 Risk)	0.9922	0.9892	0.9953

Sample Size = 22071

A Simple R Function For Odds Ratio

```
> odds.ratio <-
  function(x, pad.zeros = FALSE, conf.level=0.95) {
    if(pad.zeros) {
      if(any(x==0)) x<-x+0.5
}
    theta<-x[1,1]*x[2,2]/(x[2,1]*x[1,2])
    ASE<-sqrt(sum(1/x))
    CI<-exp(log(theta) +
      c(-1,1)*qnorm(0.5*(1+conf.level))*ASE)
    list(estimator=theta, ASE=ASE,
conf.interval=CI, conf.level=conf.level) }
```

Notes (p. 25)

- Recall the formula for sample odds ratio

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

- The sample odds ratio is 0 or ∞ if any $n_{ij} = 0$ and it is undefined if both entries in a row or column are zero.
- Consider the slightly modified formula

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}$$

- In the ASE formula also, n_{ij}'s are replaced by $n_{ij} + 0.5$.

Observations

- A sample odds ratio 1.832 does not mean that p_1 is 1.832 times p_2.
- A simple relation:

$$OddsRatio = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = RelativeRisk \times \frac{1-p_2}{1-p_1}$$

- If p_1 and p_2 are close to 0, the odds ratio and relative risk take similar values.
- This relationship between odds ratio and relative risk is useful.

Example: Smoking Status and Myocardial Infarction

Ever Smoker	Myocardial Infarction	Controls
Yes	172	173
No	90	346

- Odds Ratio=? (3.8222)
- How do we get relative risk? (2.4152)

Chi-Squared Tests (p.27)

- To test H_0 that the cell probabilities equal certain fixed values $\{\pi_{ij}\}$.
- Let $\{n_{ij}\}$ be the cell frequencies and n be the total sample size.
- Then $\mu_{ij} = n\pi_{ij}$ are the expected cell frequencies under H_0.
- Pearson (1900)'s chi-squared test statistic

$$\chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

Some Properties

- This statistic takes its minimum value of zero when all $n_{ij} = \mu_{ij}$.
- For a fixed sample size, greater differences between $\{n_{ij}\}$ and $\{\mu_{ij}\}$ produce larger χ^2 values and stronger evidence against H_0.
- The χ^2 statistic has approximately a chi-squared distribution with appropriate degrees of freedom for large sample sizes.

Likelihood-Ratio Test (p. 28)

- The likelihood ratio

$$\Lambda = \frac{\text{maximum likelihood when } H_0 \text{ is true}}{\text{maximum likelihood when parameters are unrestricted}}$$

- In mathematical notation, if $L(\theta)$ denotes the likelihood function with θ as the set of parameters and the null hypothesis is $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, the likelihood ratio is given by

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in (\Theta_0 \cup \Theta_1)} L(\theta)}$$

Properties

- Likelihood ratio cannot exceed 1.
- Small likelihood ratio implies deviation from H_0.
- Likelihood ratio test statistic is $-2\log\Lambda$, which has a chi-squared distribution with appropriate degrees of freedom for large samples.
- For a two-way contingency table, this statistic reduces to

$$G^2 = 2\sum n_{ij} log(\frac{n_{ij}}{\mu_{ij}})$$

- The test statistics χ^2 and G^2 have the same large sample distribution under null hypothesis.

Tests of Independence (p.30)

- To test: $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j.
- Equivalently, $H_0 : \mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$.
- Usually, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown.
- We estimate them, using sample proportions

$$\hat{\mu_{ij}} = np_{i+}p_{+j} = n\frac{n_{i+}n_{+j}}{n^2} = \frac{n_{i+}n_{+j}}{n}$$

- These $\{\hat{\mu}_{ij}\}$ are called estimated expected cell frequencies

Test Statistics

- Pearson's Chi-square test statistic

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

- Likelihood ratio test statistic

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} log(\frac{n_{ij}}{\hat{\mu}_{ij}})$$

- Both of them have large sample chi-squared distribution with $(I-1)(J-1)$ degrees of freedom.

Party Identification By Gender (p.31)

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	279	73	225	577
	(261.4)	(70.7)	(244.9)	
Males	165	47	191	403
	(182.6)	(49.3)	(171.1)	
Total	444	120	416	980

Example: Continued $\cdots$

- The test statistics are: $\chi^2 = 7.01$ and $G^2 = 7.00$
- Degrees of freedom $= (I-1)(J-1) = (2-1)(3-1) = 2$.
- p-value $= 0.03$.
- Thus, the above test statistics suggest that party identification and gender are associated.

SAS Codes: Read The Data

```
data Survey;
length Party $ 12;
input Gender $ Party $ count;
datalines;
Female Democrat 279
Female Independent 73
Female Republican 225
Male Democrat 165
Male Independent 47
Male Republican 191
;
run;
```

SAS Codes: Use `Proc Freq`

```
proc freq data=survey order=data;
weight count;
  tables gender*party / chisq expected
  nopercent norow nocol;
run;
```

Output

The FREQ Procedure

Table of Gender by Party

Gender Party

Frequency

Expected	Democrat	Independent	Republican	Total
Female	279	73	225	577
	261.42	70.653	244.93	
Male	165	47	191	403
	182.58	49.347	171.07	
Total	444	120	416	980

Output

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	

Sample Size = 980

R Codes

```
>gendergap<-matrix(c(279,73,225,165,47,191),
 byrow=T,ncol=3)
>dimnames(gendergap)<-
 list(Gender=c("Female","Male"),
 PartyID=c("Democrat","Independent",
 "Republican"))
>gendergap
        PartyID
 Gender   Democrat   Independent   Republican
 Female        279            73          225
   Male        165            47          191
```

R Codes

```
>chisq.test(gendergap)

 Pearson's Chi-squared test
data:  gendergap
X-squared = 7.0095, df = 2,p-value = 0.03005
```

An Alternative Way

```
>Gender<-c("Female","Female","Female","Male",
 "Male","Male")
>Party<-c("Democrat","Independent", "Republican",
 "Democrat","Independent", "Republican")
>count<-c(279,73,225,165,47,191)
>gender1<-data.frame(Gender,Party,count)
>gender<-xtabs(count Gender+Party, data=gender1)
>gender
>summary(gender)
```

Output

```
          Party
  Gender  Democrat  Independent  Republican
  Female  279       73           225
  Male    165       47           191
Call:  xtabs(formula = count  Gender + Party,
  data = gender1)
Number of cases in table:  980
Number of factors:  2
Test for independence of all factors:
    Chisq = 7.01, df = 2, p-value = 0.03005
```

Table of Expected Cell Counts

```
> rowsum<-apply(gender,1,sum)
> colsum<-apply(gender,2,sum)
> n<-sum(gender)
> gd<-outer(rowsum,colsum/n,
  make.dimnames=T)
```

Table of Expected Cell Counts

```
> gd
        Democrat   Independent   Republican
 Female  261.4163      70.65306    244.9306
   Male  182.5837      49.34694    171.0694
```

Residuals (p.31)

- To understand better the nature of evidence against H0, a cell by cell comparison of observed and estimated frequencies is necessary.
- Define, adjusted residuals

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

- If H_0 is true, each r_{ij} has a large sample standard normal distribution.
- If r_{ij} in a cell exceeds 2 then it indicates lack of fit of H_0 in that cell.
- The sign also describes the nature of association.

Computing Residuals in R

```
> rowp<-rowsum/n      %Row marginal prob.
> colp<-colsum/n      %Column marginal prob.
> pd<-outer(1-rowp,1-colp,
  make.dimnames=T)
> resid<-(gender-gd)/sqrt(gd*pd)
> resid
```

Residuals Output

	Party		
Gender	Democrat	Independent	Republican
Female	2.2931603	0.4647941	-2.6177798
Male	-2.2931603	-0.4647941	2.6177798

Some Comments (p.33)

- Pearson's χ^2 tests only indicate the degree of evidence for an association, but they cannot answer other questions like nature of association etc.
- These χ^2 tests are not always applicable. We need large data sets to apply them. Approximation is often poor when $n/(IJ) < 5$.
- The values of χ^2 or G^2 do not depend on the ordering of the rows. Thus we ignore some information when there is ordinal data.

Testing Independence For Ordinal Data (p.34)

- For ordinal data, it is important to look for types of associations when there is dependence.
- It is quite common to assume that as the levels of X increases, responses on Y tend to increases or responses on Y tends to decrease to ward higher levels of X.
- The most simple and common analysis assigns scores to categories and measures the degree of *linear trend* or correlation.
- The method used is known as "Mantel-Haenszel Chi-Square" test (Mantel and Haenszel 1959).

Linear Trend Alternative to Independence

- Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote scores for the rows.
- Let $v_1 \leq v_2 \leq \cdots \leq v_J$ denote scores for the columns.
- The scores have the same ordering as the category levels.
- Define the correlation between X and Y as

$$r = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} u_i v_j n_{ij} - (\sum_{i=1}^{I} u_i n_{i+})(\sum_{j=1}^{J} v_j n_{+j})/n}{\sqrt{[\sum_{i=1}^{I} u_i^2 n_{i+} - (\sum_{i=1}^{I} u_i n_{i+})^2/n][\sum_{j=1}^{J} v_j^2 n_{+j} - (\sum_{j=1}^{J} v_j n_{+j})^2/n]}}$$

Test For Linear Trend Alternative

- Independence between the variables implies that its true value equals zero.
- The larger the correlation is in absolute value, the farther the data fall from independence in this linear dimension.
- A test statistic is given by $M^2 = (n-1)r^2$.
- For large samples, it has approximately a chi-squared distribution with 1 degrees of freedom.

Infant Malformation and Mothers Alcohol Consumption

Alcohol Consumption	Malformation Absent(0)	Malformation Present(1)	Total
0	17,066	48	17,114
<1	14,464	38	14,502
1-2	788	5	793
3-5	126	1	127
≥6	37	1	38

Infant Malformation and Mothers Alcohol Consumption

Alcohol Consumption	Malformation Absent(0)	Present(1)	Total	Percent Present	Adjusted Residual
0	17,066	48	17,114	0.28	-0.18
<1	14,464	38	14,502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

Example: Tests For Independence

- Pearson's $\chi^2 = 12.1$, $d.f. = 4$, p-value $= 0.02$.
- Likelihood Ratio Test, $G^2 = 6.2$, $d.f. = 4$, p-value $=.19$.
- The two tests give inconsistent signals.
- The percent present and adjusted residuals suggest that there may be a linear trend.

Test For Linear Trend

- Assign scores, $v_1 = 0, v_2 = 1$ and $u_1 = 0, u_2 = 0.5, u_3 = 1.5, u_4 = 4.0$ and $u_5 = 7.0$.
- We have, $r = 0.014, n = 32,574$ and M^2 =6.6 with p-value = 0.01.
- It suggests strong evidence of a linear trend for infant malformation with alcohol consumption of mothers.

SAS Codes

```
data infants;
input malform alcohol count @@;
datalines;
 1     0   17066     2     0   48
 1   0.5   14464     2   0.5   38
 1   1.5     788     2   1.5    5
 1   4.0     126     2   4.0    1
 1   7.0      37     2   7.0    1
;
run;
proc format;
  value malform 2='Present' 1='Absent';
  value Alcohol 0='0' 0.5='<1' 1.5='1-2' 4.0='3-5'
7.0='>=6';
run;
```

SAS Codes

```
proc freq data = infants;
  format malform malform.  alcohol alcohol.;
  weight count;
  tables alcohol*malform / chisq cmh1 norow
  nocol nopercent;
run;
```

Partial Output

Statistic	DF	Value	Prob
Chi-Square	4	12.0821	0.0168
Likelihood Ratio Chi-Square	4	6.2020	0.1846
Mantel-Haenszel Chi-Square	1	6.5699	0.0104
Phi Coefficient		0.0193	
Contingency Coefficient		0.0193	
Cramer's V		0.0193	

Partial Output

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.5699	0.0104

Total Sample Size = 32574

Notes

- The correlation r has limited use as a descriptive measure of tables.
- Different choices of monotone scores usually give similar results.
- However, it may not happen when the data are very unbalanced, i.e. when some categories have many more observations than other categories.
- If we had taken $(1, 2, 3, 4, 5)$ as the row scores in our example, then $M^2 = 1.83$ and p-value $= 0.18$ gives a much weaker conclusion.
- It is usually better to use one's own judgment by selecting scores that reflect distances between categories.

SAS Codes

```
data infantsx;
input malform alcoholx count @@;
datalines;
 1   0   17066      2   0   48
 1   1   14464      2   1   38
 1   2     788      2   2    5
 1   3     126      2   3    1
 1   4      37      2   4    1
;
run;
proc freq data = infantsx;
weight count;
tables alcoholx*malform / cmh1 norow nocol nopercent;
run;
```

Partial Output

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	1.8278	0.1764

Total Sample Size = 32574

Fisher's Tea Tasting Experiment (p.39)

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Example: Tea Tasting

- To test whether she can tell accurately.
- To test H_0: $\theta = 1$ against $H_1 : \theta > 1$.
- We cannot use previously discussed tests as we have a very small sample size.

Fisher's Exact Test

- For a 2×2 table, under the assumption of independence, i.e. $\theta = 1$, the conditional distribution of n_{11} given the row and column totals is hypergeometric.
- For given row and column marginal totals, the value for n_{11} determines the other three cell counts. Thus, the hypergeometric formula expresses probabilities for the four cell counts in terms of n_{11} alone.

Fisher's Exact Test

- When $\theta = 1$, the probability of a particular value n_{11} for that count equals

$$p(n_{11}) = \frac{\begin{pmatrix} n_{1+} \\ n_{11} \end{pmatrix} \begin{pmatrix} n_{2+} \\ n_{+1} - n_{11} \end{pmatrix}}{\begin{pmatrix} n \\ n_{+1} \end{pmatrix}}$$

- To test independence, the p-value is the sum of hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome.

Example: Tea Tasting

- The outcomes at least as favorable as the observed data is $n_1 1 = 3$ and 4 given the row and column totals.
- Hence,

$$p(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} = .2286,$$

$$p(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70} = .0143.$$

Example: Tea Tasting

- Therefore, p-value $= P(3) + P(4) = 0.243$.
- There is not much evidence against the null hypothesis of independence.
- The experiment did not establish an association between the actual order of pouring and the guess.

SAS Codes: Exact Test

```
data tea;
  input poured $ guess $ count @@;
  datalines;
Milk Milk 3     Milk Tea 1
Tea Milk 1     Tea Tea 3
;
proc freq data=tea order=data;
  weight count;
  tables poured*guess / exact;
run;
```

Partial Output

```
Fisher's Exact Test
---------------------------------
Cell (1,1) Frequency (F)          3
Left-sided Pr <= F           0.9857
Right-sided Pr >= F          0.2429
Table Probability (P)        0.2286
Two-sided Pr <= P            0.4857
Sample Size = 8
```

R Codes

```
> Poured<-c("Milk","Milk","Tea","Tea")
> Guess<-c("Milk","Tea","Milk","Tea")
> count<-c(3,1,1,3)
> teadata<-data.frame(Poured,Guess,count)
> tea<-xtabs(count Poured+Guess,data=teadata)
> fisher.test(tea,alternative="greater")
```

Output

```
Fisher's Exact Test for Count Data
data:  tea p-value = 0.2429
alternative hypothesis:  true odds ratio is
  greater than 1
95 percent confidence interval:
0.3135693     Inf
sample estimates:
odds ratio
    6.408309
```