Lecture 7: Testing Independence

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Testing Independence

- Previously, we looked at RR = OR = 1 to determine independence.
- Now, lets revisit the Pearson and Likelihood Ratio Chi-Squared tests.
- Pearson's Chi-Square

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Likelihood Ratio Test

$$G^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \log(\frac{O_{ij}}{E_{ij}})$$

- Since both X^2 and G^2 are distributed as approximately χ^2, in order to draw inference about the significance of both, we need the degrees of freedom.

Degrees of Freedom

- A way to think about degrees of freedom is to relate it to the number of “pieces” of information you need to complete a table.
- More specifically, Degrees of Freedom (df) equals

  df = Number of cells - Number of Constraints - Number of Parameters Estimated

- First, lets consider Pearson’s Chi-Square
- We will derive df for the Cross Sectional Design using this definition.

- For the general $I \times J$ contingency table, there are a total of IJ cells.
- Under the Multinomial sampling design, the only constraint is that $\sum p_{ij} = 1$ so there is only one constraint.
- Under the hypothesis on interest, we are interested in estimating the marginal probabilities.
  - Since the sample size is fixed, we only need to estimate $I - 1$ marginal row probabilities.
  - Namely $p_{1\cdot}, p_{2\cdot}, \ldots, p_{(I-1)\cdot}$
  - Likewise, we only need to estimate $J - 1$ column marginals.
- Thus,

$$\begin{aligned} df &= IJ - \text{Number of Constraints - Number of Parameters Estimated} \\ df &= IJ - 1 - ((I-1) + (J-1)) = IJ - I - J + 1 = (I-1)(J-1) \end{aligned}$$

Degrees of Freedom for the Product binomial Sampling

- Again, there are IJ cells in our $I \times J$ contingency table
- For the Prospective design, we have constraints that each rows probability sums to 1, so there are I constraints.
- Although we did not state it directly before, the hypothesis of interest is the "Homogeneity" hypothesis. That is, that $H_0 = p_{ij} = p_{\cdot j}$ for $j = 1, 2, \ldots J$. Therefore, there are $J - 1$ estimated marginal probabilities.
- Then the DF equals,

$$df = IJ - I - (J - 1) = IJ - I - J + 1 = (I - 1)(J - 1)$$

In summary for Pearson's Chi-Square

- For the remaining study design (Case-Control), the degrees of freedom can be shown to be $(I-1)(J-1)$.
- Therefore, regardless of the sample design, the df for any $I \times J$ contingency table using Pearson's Chi-Square is $(I-1)(J-1)$.
- For the 2×2 tables we have been studying,

$$df = (2-1) \times (2-1) = 1$$

Likelihood Ratio Test

- If you recall, we described the df for the likelihood ratio test as the difference in the number of parameters estimated under the alternative minus the number estimated under the null.
- Under the multinomial sampling design, the alternative model is that $p_{ij} \neq p_{i\cdot}p_{\cdot j}$ and as such, $\sum_i \sum_j p_{ij} = 1$. Thus, there is only one constraint and we estimate $IJ - 1$ cell probabilities.
- Under the null, we have $p_{ij} = p_{i\cdot}p_{\cdot j}$ which is determined by $(I-1)$ and $(J-1)$ marginals. Thus, we only estimate $[(I-1)+(J-1)]$ marginal probabilities.
- Thus, the DF of G^2 is

$$df = IJ - 1 - [(I-1)+(J-1)] = (I-1)(J-1)$$

.

Comparison of X^2 and G^2

- Pearson and the LRT have same limiting distribution. (both converge in distribution to χ^2 with $df = (I-1)(J-1)$ as $n \to \infty$)
- Pearson's is score based
- LRT combines the information of the null and alternative hypotheses
- So which one is best?

Choosing X^2 or G^2

- X^2 converges in distribution faster than G^2.
- When $n/IJ < 5$ (less than 5 per cell), G^2 usually is not a good estimate.
- When I or J is large, Pearson's usually is valid when some $E_{ij} < 5$ but most are greater than 5.
- Therefore, for the general $I \times J$ table, you can usually just use Pearson's Chi Square.
- We will now develop a test for small samples.

Small Samples

Question: Is there Gender Bias in Jury Selection?

		SELECTED FOR JURY: YES	SELECTED FOR JURY: NO	Total
GENDER	FEMALE	1	9	10
	MALE	11	9	20
	Total	12	18	30

The sampling distribution for this study design is hypergeometric.

However, we will adapt the study design into a small sample exact test.

- In this study, we COULD consider the column totals fixed by design (since the jury has to have 12 members), and the row totals random.
- Then, the columns are independent binomials.
- Using SAS

```
data one;
input sex $ jury $ count;
cards;
1FEMALE 1YES  1
1FEMALE 2NO   9
2MALE   1YES 11
2MALE   2NO   9
;

proc freq;
 table sex*jury/expected chisq;
weight  count;
run;
```

TABLE OF SEX BY JURY

SEX	JURY		
Frequency |			
Expected |			
Percent |			
Row Pct |			
Col Pct |	1YES	2NO	Total
1FEMALE	1	9	10
	4	6	
	3.33	30.00	33.33
	10.00	90.00	
	8.33	50.00	
2MALE	11	9	20
	8	12	
	36.67	30.00	66.67
	55.00	45.00	
	91.67	50.00	
Total	12	18	30
	40.00	60.00	100.00

STATISTICS FOR TABLE OF SEX BY JURY

Statistic	DF	Value	Prob
Chi-Square	1	5.6250	0.0177
Likelihood Ratio Chi-Square	1	6.3535	0.0117
Continuity Adj. Chi-Square	1	3.9063	0.0481
Mantel-Haenszel Chi-Square	1	5.4375	0.0197
Phi Coefficient		-0.4330	
Contingency Coefficient		0.3974	
Cramer's V		-0.4330	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

- A rule of thumb in SAS is that the Large Sample approximations for the likelihood ratio and Pearson's Chi-Square are not very good if the sample size is small

```
WARNING: 25% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.
```

- Suppose for a cross sectional, prospective, or case-control design:
  some of the cell counts are small (so that $E_{ij} < 5$), and you want to make inferences about the OR.
- A popular technique with small samples is to fix both margins of the (2×2) table, and use 'Exact Tests' and confidence intervals.

Exact Tests - For the 2×2 table

Suppose, then, for:

1. A prospective study (rows margins fixed) we further condition on the column margins
2. A case-control study (column margins fixed) we further condition on the rows margins
3. A cross sectional (total fixed) we condition on both row and column margins.
4. In all cases, we have a conditional distribution with row and column margins fixed.

Question

- What is the conditional distribution of Y_{11} given both row and column margins are fixed.
- First note, unlike the other distributions discussed, since the margins are fixed and known, we will show that this conditional distribution is a function of only one unknown parameter
- This follows from what we have seen:
- If the total sample size is fixed (cross sectional), we have 3 unknown parameters, (p_{11}, p_{12}, p_{21})
- If one of the margins is fixed (prospective, or case-control study), we have two unknown parameters, (p_1, p_2) or (π_1, π_2)
- Intuitively, given we know both margins, if we know one cell count (say Y_{11}), then we can figure out the other 3 cell counts by subtraction. This implies that we can characterize the conditional distribution by 1 parameter.
- Thus, given the margins are fixed, we only need to consider one cell count as random, and, by convention Y_{11} is usually chosen. (you could have chosen any of the 4 cell counts, though).

Can you complete all of the observed cell counts given the information available? Yes.

		Column 1	2	
Row	1	Y_{11}		$Y_{1.}$
	2			$Y_{2.}$
		$Y_{.1}$	$Y_{.2}$	$N = n_{..}$

- **Question:** Then, what is the conditional distribution of Y_{11} given both row and column margins are fixed.

$$P[Y_{11} = y_{11}|y_{1\cdot}, y_{\cdot 1}, y_{\cdot\cdot}, OR]$$

- After some tedious algebra, you can show it is non-central hypergeometric, i.e.,

$$P[Y_{11} = y_{11}|y_{1\cdot}, y_{\cdot 1}, y_{\cdot\cdot}, OR] =$$

$$\frac{\left(\begin{array}{c} y_{\cdot 1} \\ y_{11} \end{array}\right)\left(\begin{array}{c} y_{\cdot\cdot} - y_{\cdot 1} \\ y_{1\cdot} - y_{11} \end{array}\right)(OR)^{y_{11}}}{\sum_{\ell=0}^{y_{\cdot 1}}\left(\begin{array}{c} y_{\cdot 1} \\ \ell \end{array}\right)\left(\begin{array}{c} y_{\cdot\cdot} - y_{\cdot 1} \\ y_{1\cdot} - \ell \end{array}\right)(OR)^{\ell}}$$

  where, for all designs,

$$OR = \frac{O_{11}O_{22}}{O_{21}O_{12}},$$

- We denote the distribution of Y_{11} by

$$(Y_{11}|y_{1\cdot}, y_{\cdot 1}, y_{\cdot\cdot}) \sim HG(y_{\cdot\cdot}, y_{\cdot 1}, y_{1\cdot}, OR)$$

Notes about non-central hypergeometric

- Again, unlike the other distributions discussed, since the margins are fixed and known, the non-central hypergeometric is a function of only one unknown parameter, the OR.
- Thus, the conditional distribution given both margins is called non-central hypergeometric.
- Given both margins are fixed, if you know one of the 4 cells of the table, then you know all 4 cells (only one of the 4 counts in the table is non-redundant).
- Under the null H_0: OR=1, the non-central hypergeometric is called the central hypergeometric or just the hypergeometric.
- We will use the hypergeometric distribution (i.e., the non-central hypergeometric under H_0: OR=1) to obtain an 'Exact' Test for H_0: OR=1. This test is appropriate in small samples.

Fisher's Exact Test

Let's consider the following table with both Row and Column totals fixed.

		Column		
		1	2	
Row	1	Y_{11}	Y_{12}	$Y_{1\cdot}$
	2	Y_{21}	Y_{22}	$Y_{2\cdot}$
		$Y_{\cdot 1}$	$Y_{\cdot 2}$	$N = Y_{\cdot\cdot}$

Many define the $\{1, 1\}$ cell as the "Pivot Cell".

Before we consider the sampling distribution, lets consider the constraints on the Pivot Cell.

The Values L_1 and L_2

- We know that Y_{11} must not exceed the marginal totals, $Y_{\cdot 1}$ or $Y_{1\cdot}$.
- That is,

$$Y_{11} \leq Y_{\cdot 1} \text{ and } Y_{11} \leq Y_{1\cdot}.$$

- Therefore, the largest value Y_{11} can assume can be denoted as L_2 in which

$$L_2 = \min(Y_{\cdot 1}, Y_{1\cdot})$$

- Similarly, the minimum value of Y_{11} is also constrained.
- It is harder to visualize, but the minimum value Y_{11} can assume, denoted as L_1, is

$$L_1 = \max(0, Y_{1\cdot} + Y_{\cdot 1} - Y_{\cdot\cdot})$$

Example

Suppose you observe the following marginal distribution.

		Column 1	2	
Row	1	y_{11}		6
	2			3
		5	4	9

- We want to determine L_1 and L_2
- So that we can determine the values the Pivot Cell can assume.
- The values in which the Pivot Cell can assume are used in the significance testing.

Based on the previous slide's table, L_1 and L_2 are

$$\begin{aligned} L_1 &= \max(0, Y_{1\cdot} + Y_{\cdot 1} - Y_{\cdot\cdot}) \\ &= \max(0, 6 + 5 - 9) \\ &= \max(0, 2) \\ &= 2 \end{aligned}$$

and

$$\begin{aligned} L_2 &= \min(Y_{1\cdot}, Y_{\cdot 1}) \\ &= \min(6, 5) \\ &= 5 \end{aligned}$$

Therefore, the values that Y_{11} can assume are $\{2, 3, 4, 5\}$.

All Possible Contingency Tables

- Since each table is uniquely defined by the pivot cell, the following tables are all of the possible configurations.

$OR_E = 0.078$		Column 1	2	
Row	1	2	4	6
	2	3	0	3
		5	4	9

$OR = 0.5$		Column 1	2	
Row	1	3	3	6
	2	2	1	3
		5	4	9

$OR = 4$		Column 1	2	
Row	1	4	2	6
	2	1	2	3
		5	4	9

$OR_E = 25.7$ ***		Column 1	2	
Row	1	5	1	6
	2	0	3	3
		5	4	9

- Suppose the table observed is flagged with "***". When freq = 0, we use 0.5.
- How do we know if the Rows and Columns are independent?
- Note, as Y_{11} increases, so does the OR.

Test Statistics

- The probability of observing any given table is

$$P[Y_{11} = y_{11}|Y_{1\cdot}, Y_{2\cdot}, Y_{\cdot 1}, Y_{\cdot 2}] = \frac{\left(\begin{array}{c} y_{\cdot 1} \\ y_{11} \end{array}\right)\left(\begin{array}{c} y_{\cdot 2} \\ y_{12} \end{array}\right)}{\left(\begin{array}{c} y_{\cdot\cdot} \\ y_{1\cdot} \end{array}\right)}$$

- The probability of observing our table is

$$\begin{aligned} P[Y_{11} = 5|6, 3, 5, 4] &= \frac{\left(\begin{array}{c} 5 \\ 5 \end{array}\right)\left(\begin{array}{c} 4 \\ 1 \end{array}\right)}{\left(\begin{array}{c} 9 \\ 6 \end{array}\right)} \\ &= \frac{4}{84} \\ &= 0.0476 \end{aligned}$$

- We now need to develop tests to determine whether or not this arrangement supports or rejects independence.

One-sided Tests

- Suppose we want to test

$$\mathsf{H}_O{:}OR = 1 \qquad or \qquad E(Y_{11}) = y_{1\cdot}y_{\cdot 1}/y_{\cdot\cdot}$$

  versus

$$\mathsf{H}_A{:}OR > 1 \qquad or \qquad E(Y_{11}) > y_{1\cdot}y_{\cdot 1}/y_{\cdot\cdot}$$

- Let $y_{11,obs}$ be the observed value of Y_{11}; we will reject the null in favor of the alternative if $y_{11,obs}$ is large (recall from the example, as Y_{11} increases, so does the OR).
- Then, the exact $p-$value (one-sided) is the sum of the table probabilities in which the pivot cell is greater than or equal to the $Y_{11,obs}$.

- Or more specifically, The exact $p-$value looks at the upper tail:

$$p - \text{value} = P[Y_{11} \geq y_{11,obs} | \text{H}_O\text{:}OR = 1]$$

$$= \sum_{\ell=y_{11,obs}}^{L_2=\min(y_{\cdot 1}, y_{1\cdot})} \frac{\begin{pmatrix} y_{\cdot 1} \\ \ell \end{pmatrix} \begin{pmatrix} y_{\cdot 2} \\ y_{1\cdot} - \ell \end{pmatrix}}{\begin{pmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \end{pmatrix}}$$

- Note that ℓ increments the values of Y_{11} to produce the tables as extreme ($\ell = Y_{11,obs}$ and more extreme (approaching L_2)
- Note $y_{1\cdot} = y_{11} + y_{12}$ so $y_{12} = y_{1\cdot} - y_{11}$.

- Suppose we want to test

$$\text{H}_O\text{:}OR = 1 \qquad or \qquad E(Y_{11}) = y_{1\cdot}y_{\cdot 1}/y_{\cdot\cdot}$$

versus

$$\text{H}_A\text{:}OR < 1 \qquad or \qquad E(Y_{11}) < y_{1\cdot}y_{\cdot 1}/y_{\cdot\cdot}$$

- We will reject the null in favor of the alternative if $y_{11,obs}$ is small.
- Then, the exact $p-$value looks at the lower tail:

$$p - \text{value} \quad = \quad P[Y_{11} \leq y_{11,obs}|\text{H}_O\text{:}OR = 1]$$

$$= \quad \sum_{\ell=L_1=\max(0,y_{1\cdot}+y_{\cdot 1}-y_{\cdot\cdot})}^{y_{11,obs}} \frac{\begin{pmatrix} y_{\cdot 1} \\ \ell \end{pmatrix}\begin{pmatrix} y_{\cdot 2} \\ y_{1\cdot}-\ell \end{pmatrix}}{\begin{pmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \end{pmatrix}}$$

Fisher's Exact Test - Two-sided Test

- Suppose we want to test

$$\mathsf{H}_O{:}OR = 1 \qquad or \qquad E(Y_{11}) = y_{1\cdot}y_{\cdot 1}/y_{\cdot\cdot}$$

  versus

$$\mathsf{H}_A{:}OR \neq 1 \qquad or \qquad E(Y_{11}) \neq y_{1\cdot}y_{\cdot 1}/y_{\cdot\cdot}$$

- The exact $p-$value here is the exact 2-sided p-value is

$$P\left[\begin{array}{l}\text{seeing a result as likely or} \\ \text{less likely than the observed} \\ \text{result in either direction}\end{array}\middle|\ \mathsf{H}_0 : OR = 1\right].$$

In general, to calculate the 2-sided $p-$value,

1. Calculate the probability of the observed result under the null

$$\pi = P[Y_{11} = y_{11,obs} | \mathsf{H}_O{:}OR = 1]$$

$$= \frac{\begin{pmatrix} y_{\cdot 1} \\ y_{11,obs} \end{pmatrix} \begin{pmatrix} y_{\cdot\cdot} - y_{\cdot 1} \\ y_{1\cdot} - y_{11,obs} \end{pmatrix}}{\begin{pmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \end{pmatrix}}$$

2. Recall, Y_{11} can take on the values

$$\max(0, y_{1\cdot} + y_{\cdot 1} - y_{\cdot\cdot}) \leq Y_{11} \leq \min(y_{1\cdot}, y_{\cdot 1}),$$

Calculate the probabilities of all of these values,

$$\pi_\ell = P[Y_{11} = \ell | \mathsf{H}_O{:}OR = 1]$$

3. Sum the probabilities π_ℓ in (2.) that are less than or equal to the observed probability π in 1.

$$p - value = \sum_{\ell=\max(0, y_{1\cdot}+y_{\cdot 1}-y_{\cdot\cdot})}^{\min(y_{1\cdot}, y_{\cdot 1})} \pi_\ell I(\pi_\ell \leq \pi)$$

where

$$I(\pi_\ell \leq \pi) = \begin{cases} 1 \text{ if } \pi_\ell \leq \pi \\ 0 \text{ if } \pi_\ell > \pi \end{cases} .$$

Using our example “By Hand”

Recall, $P(Y_{11,obs} = 5) = 0.0476$. Below are the calculations of the other three tables.

$$\begin{aligned} P[Y_{11} = 2|6,3,5,4] &= \frac{\binom{5}{2}\binom{4}{4}}{\binom{9}{6}} \\ &= \frac{10}{84} \\ &= 0.1190 \end{aligned}$$

$$\begin{aligned} P[Y_{11} = 3|6,3,5,4] &= \frac{\binom{5}{3}\binom{4}{3}}{\binom{9}{6}} \\ &= \frac{40}{84} \\ &= 0.4762 \end{aligned}$$

$$
\begin{aligned}
P[Y_{11} = 4|6, 3, 5, 4] &= \frac{\begin{pmatrix} 5 \\ 4 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix}}{\begin{pmatrix} 9 \\ 6 \end{pmatrix}} \\
&= \frac{30}{84} \\
&= 0.3571
\end{aligned}
$$

- Then, for H$_A$:$OR < 1$,
  $p-$value = 0.1190 + 0.4762 + 0.3571 + 0.0476 = 1.0

- for H$_A$:$OR > 1$,
  $p-$value = 0.0476

- For H$_A$:$OR \neq 1$,
  $p-$value = 0.0476 (we observed the most extreme arrangement)

Using SAS

```
data test;
 input row $ col$ count;
cards;
1row 1col 5
1row 2col 1
2row 1col 0
2row 2col 3
;
run;
proc freq;
tables row*col/exact;
 weight count;
run;
```

Frequency Percent Row Pct Col Pct	1col	2col	Total
1row	5 55.56 83.33 100.00	1 11.11 16.67 25.00	6 66.67
2row	0 0.00 0.00 0.00	3 33.33 100.00 75.00	3 33.33
Total	5 55.56	4 44.44	9 100.00

Statistics for Table of row by col

Statistic	DF	Value	Prob
Chi-Square	1	5.6250	0.0177
Likelihood Ratio Chi-Square	1	6.9586	0.0083
Continuity Adj. Chi-Square	1	2.7563	0.0969
Mantel-Haenszel Chi-Square	1	5.0000	0.0253
Phi Coefficient		0.7906	
Contingency Coefficient		0.6202	
Cramer's V		0.7906	

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

```
       Fisher's Exact Test
-----------------------------------
Cell (1,1) Frequency (F)          5
Left-sided Pr <= F           1.0000
Right-sided Pr >= F          0.0476

Table Probability (P)        0.0476
Two-sided Pr <= P            0.0476

Sample Size = 9
```

General Notes about Fisher's Exact Test

- Fisher's Exact $p-$values is one of the most frequently used $p-$values you will find in the medical literature (for "good studies")
- However, Cruess (1989) reviewed 201 scientific articles published during 1988 in *The American Journal of Tropical Medicine and Hygiene* and found 148 articles with at least one statistical error. The most common error was found to be the use of a large sample χ^2 $p-$value when the sample was too small for the approximation.
- Since the values of Y_{11} is discrete (highly discrete given a small sample size such as in our example), the actual number of possible $p-$values is limited.
- For example, Given our example margins, $\{0.0476, 0.1666, 0.5237, 1.0\}$ are our only potential values.

- The hypergeometric (when $OR = 1$) is symmetrically defined in the rows and columns.

Variable (X)	Variable (Y) 1	2	
1	Y_{11}	Y_{12}	$Y_{1.}$
2	Y_{21}	Y_{22}	$Y_{2.}$
	$Y_{.1}$	$Y_{.2}$	$Y_{..}$

In particular, under H_0 : OR = 1

$$P[Y_{11} = y_{11}|OR = 1] = \frac{\begin{pmatrix} y_{\cdot 1} \\ y_{11} \end{pmatrix}\begin{pmatrix} y_{\cdot 2} \\ y_{21} \end{pmatrix}}{\begin{pmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \end{pmatrix}}$$

$$= \frac{\begin{pmatrix} y_{1\cdot} \\ y_{11} \end{pmatrix}\begin{pmatrix} y_{2\cdot} \\ y_{21} \end{pmatrix}}{\begin{pmatrix} y_{\cdot\cdot} \\ y_{\cdot 1} \end{pmatrix}}$$

Expected Value of Y_{11} under the null

- Recall, for the hypergeometric distribution, the margins $Y_{i\cdot}$, $Y_{\cdot j}$ and $Y_{\cdot\cdot}$ are assumed known and fixed.
- From the theory of the hypergeometric distribution, under the null of no association, the mean is

$$E(Y_{ij}|OR = 1) = \frac{y_{i\cdot}y_{\cdot j}}{y_{\cdot\cdot}}$$

- For other distributions, we could not write the expected value in terms of the possibly random $Y_{i\cdot}$ and/or $Y_{\cdot j}$. Since $(Y_{i\cdot}, Y_{\cdot j}, Y_{\cdot\cdot})$ are known for the hypergeometric, we can write the expected value in terms of them.

- Thus, the null H$_0$:$OR = 1$ can be rewritten as

$$\text{H}_0\text{:}E(Y_{ij}|OR = 1) = \frac{y_{i\cdot}y_{\cdot j}}{y_{\cdot\cdot}},$$

- Under no association,

$$E_{ij} = \frac{[i^{th} \text{ row total } (y_{i\cdot})] \cdot [j^{th} \text{ column total } (y_{\cdot j})]}{[\text{total sample size } (y_{\cdot\cdot})]},$$

  is the estimate of $E(Y_{ij})$ under the null of no association

- However, under independence, E_{ij} is the exact conditional mean (not an estimate) since $y_{i\cdot}$ and $y_{\cdot j}$ are both fixed.

Miscellaneous notes regarding X^2 Test

- Suppose we have the following

$$p_1 = .4$$

- and

$$p_2 = .6$$

- where p_1 and p_2 are the true success rates for a prospective study.
- Thus, the true odds ratio is

$$OR = \frac{.40 \cdot .80}{.20 \cdot .60} = 2\frac{2}{3} \quad = 2.666$$

Potential Samples

- Suppose we randomized 50 subjects (25 in each group) and observe the following table

	Success	Failure	Total
Group 1	10	15	25
Group 2	5	20	25
Total	15	35	50

- And use SAS to test $p_1 = p_2$

```
options nocenter;
data one;
 input row col count;
 cards;
 1 1 10
 1 2 15
 2 1 5
 2 2 20
 ;
 run;
 proc freq data=one;
  tables row*col/chisq  measures;
  weight count;
run;
```

Selected Results

```
The FREQ Procedure

       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        10
Left-sided Pr <= F          0.9689
Right-sided Pr >= F         0.1083

Table Probability (P)       0.0772
Two-sided Pr <= P           0.2165

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      2.6667        0.7525        9.4497
Cohort (Col1 Risk)             2.0000        0.7976        5.0151
Cohort (Col2 Risk)             0.7500        0.5153        1.0916

Sample Size = 50
```

Example Continued

- For this trial, we would fail to reject the null hypothesis (p=0.2165).
- However, our estimated odds ratio is 2.6666 and relative risk is 2.0
- What would happen if our sample size was larger?

```
data two;
 input row col count;
 cards;
1 1 40
1 2 60
2 1 20
2 2 80
;
run;
 proc freq data=two;
  tables row*col/chisq measures;
  weight count;
run;
```

Fisher's Exact Test

Cell (1,1) Frequency (F)	40
Left-sided Pr <= F	0.9995
Right-sided Pr >= F	0.0016
Table Probability (P)	0.0010
Two-sided Pr <= P	0.0032

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	2.6667	1.4166	5.0199
Cohort (Col1 Risk)	2.0000	1.2630	3.1670
Cohort (Col2 Risk)	0.7500	0.6217	0.9048

Sample Size = 200

Moral of the Story?

- Both examples have the exact same underlying probability distribution
- Both examples have the exact same estimates for OR and RR
- The statistical significance differed
- A Chi-square (or as presented Fisher's exact)'s $p-$value does not indicate how strong an association is in the data (i.e., a smaller p-value, say < 0.001, does not mean there is a "strong" treatment effect)
- It simply indicates that you have evidence for the alternative (i.e., $p_1 \neq p_2$).
- You must use a measure of association to quantify this difference

Generalized Odds Ratio

- For the 2×2 table, a single measure can summarize the association.
- The association could be the Odds Ratio or Relative Risk
- For the general $I \times J$ case, a single measure cannot summarize the association without loss of information.
- However, a set of odds ratios or another summary index (such as a correlation measure) can summarize the association

Note: “Loss of information” can be obtained by collapsing the categories into a 2×2 structure.

Example we will use

Agresti Table 2.1 - Page 37			
	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10845
Aspirin	5	99	10933

We want to estimate the association of Aspirin Use on MI.

Collapsed Categories

We could collapse the Fatal Attack and Nonfatal Attack categories together to obtain

	Myocardial Infarction	
	Fatal Attack or Nonfatal attack	No Attack
Placebo	189	10845
Aspirin	104	10933

Then, the OR of having a MI is

$$\begin{aligned} OR_{MI} &= \frac{189*10933}{104*10933} \\ &= 1.83 \end{aligned}$$

Thus, the odds of a MI are 1.83 times higher when taking placebo when compared to aspirin.

Generalized Odds Ratio

- There are $\left(\begin{array}{c} I \\ 2 \end{array} \right)$ pairs of rows
- and $\left(\begin{array}{c} J \\ 2 \end{array} \right)$ pairs of columns
- that can produce $\left(\begin{array}{c} I \\ 2 \end{array} \right) \left(\begin{array}{c} J \\ 2 \end{array} \right)$ estimates of the odds ratio
- We are going to consider three cases for the generalized odds ratio

Case 1

For rows a and b and columns c and d, the odds ratio ($\pi_{ac}\pi_{bd}/\pi_{bc}\pi_{ad}$) is the most loosely defined set of generalizes ORs.

There are $\begin{pmatrix} I \\ 2 \end{pmatrix}\begin{pmatrix} J \\ 2 \end{pmatrix}$ of this type.

For our example, lets compare Fatal MI to No MI.

$$OR_{\text{fatal vs. No MI}} = \frac{18 * 10933}{5 * 10845} = 3.63$$

That is, the odds of a having a fatal MI vs No MI are 3.63 times higher for the Placebo group when compared to the group taking Aspirin.

Case 2: Local ORs

The local ORs are obtained by comparing adjacent rows and columns.

That is,

$$OR_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}$$

For our example, we could obtain 2 local ORs

1. Fatal MI vs. Non Fatal MI ($OR = (18 \cdot 99)/(5 \cdot 171) = 2.08$)
2. Non Fatal MI vs. No MI ($OR = (171 \cdot 10933)/(99 \cdot 10845) = 1.74$)

Note: There are $(I-1)(J-1)$ local odds ratio.

Case 3: Last Column (Reference) OR

For the $I \times J$ table with I representing the last row and J representing the last column, then

$$\alpha_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}}, \quad i = 1, 2, \ldots, I-1, \quad j = 1, 2, \ldots, J-1$$

represents the OR obtained by referencing the last row and last column. For our example,

1. $\alpha_{11} = (18 * 10933)/(5 * 10933) = 3.62$
2. $\alpha_{12} = (171 * 10933)/(99 * 10845) = 1.74$

Summary of Generalized Methods

- The set of $(I-1)(J-1)$ ORs contain much redundant information
- As illustrated, many of the approaches provide the same result
- When interpreting the ORs, be mindful of the reference category and state it in the summary
- Independence is equivalent to all $(I-1)(J-1)$ ORs = 1