Lecture 16

Chi-Squared Test, Relative Risk, and Odds Ratio

Text: Chapter 10

STAT 8020 Statistical Methods II
October 15, 2020

Whitney Huang
Clemson University

Agenda

1. Chi-Squared Test

2. Relative Risk and Odds Ratio

Review: χ^2-Test for Independence

1. Calculate the χ^2 statistic:

$$\chi^2_{obs} = \sum \text{partial } \chi^2 \text{ value}$$

2. Calculate the degrees of freedom (df)

$$df = (\#\text{of rows} - 1) \times (\#\text{of columns} - 1)$$

3. Find the χ^2 critical value with respect to α

4. Draw the conclusion:

Reject H_0 if χ^2_{obs} is bigger than the χ^2 critical value $\Rightarrow$
There is an statistical evidence that there is a relationship between the two variables at α level

Handedness/Gender Example Revisited

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

Is the percentage left-handed men in the population different from the percentage of left-handed women?

Example

A 2011 study was conducted in Kalamazoo, Michigan. The objective was to determine if parents' marital status affects children's marital status later in their life. In total, 2,000 children were interviewed. The columns refer to the parents' marital status. Use the contingency table below to conduct a χ^2 test from beginning to end. Use α = $.10$

(Observed)	Married	Divorced	Total
Married	581	487	
Divorced	455	477	
Total			

Example Cont'd

1. Define the Null and Alternative hypotheses:

   H_0 : there is no relationship between parents' marital status and childrens' marital status.

   H_a : there is a relationship between parents' marital status and childrens' marital status

2. Calculate the marginal totals, and the grand total

(Observed)	Married	Divorced	Total
Married	581	487	1068
Divorced	455	477	932
Total	1036	964	2000

Example Cont'd

3. Calculate the expected cell counts

(Expected)	Married	Divorced
Married	$\frac{1068 \times 1036}{2000} = 553.224$	$\frac{1068 \times 964}{2000} = 514.776$
Divorced	$\frac{932 \times 1036}{2000} = 482.776$	$\frac{932 \times 964}{2000} = 449.224$

Example Cont'd

3. Calculate the expected cell counts

(Expected)	Married	Divorced
Married	$\frac{1068\times1036}{2000} = 553.224$	$\frac{1068\times964}{2000} = 514.776$
Divorced	$\frac{932\times1036}{2000} = 482.776$	$\frac{932\times964}{2000} = 449.224$

4. Calculate the partial χ^2 values

partial χ^2	Married	Divorced
Married	$\frac{(581-553.224)^2}{553.224} = 1.39$	$\frac{(487-514.776)^2}{514.776} = 1.50$
Divorced	$\frac{(455-482.776)^2}{482.776} = 1.60$	$\frac{(477-449.224)^2}{449.224} = 1.72$

Example Cont'd

3. Calculate the expected cell counts

(Expected)	Married	Divorced
Married	$\frac{1068\times1036}{2000} = 553.224$	$\frac{1068\times964}{2000} = 514.776$
Divorced	$\frac{932\times1036}{2000} = 482.776$	$\frac{932\times964}{2000} = 449.224$

4. Calculate the partial χ^2 values

partial χ^2	Married	Divorced
Married	$\frac{(581-553.224)^2}{553.224} = 1.39$	$\frac{(487-514.776)^2}{514.776} = 1.50$
Divorced	$\frac{(455-482.776)^2}{482.776} = 1.60$	$\frac{(477-449.224)^2}{449.224} = 1.72$

5. Calculate the χ^2 statistic
$\chi^2 = 1.39 + 1.50 + 1.60 + 1.72 = 6.21$

Example Cont'd

3. Calculate the expected cell counts

(Expected)	Married	Divorced
Married	$\frac{1068 \times 1036}{2000} = 553.224$	$\frac{1068 \times 964}{2000} = 514.776$
Divorced	$\frac{932 \times 1036}{2000} = 482.776$	$\frac{932 \times 964}{2000} = 449.224$

4. Calculate the partial χ^2 values

partial χ^2	Married	Divorced
Married	$\frac{(581-553.224)^2}{553.224} = 1.39$	$\frac{(487-514.776)^2}{514.776} = 1.50$
Divorced	$\frac{(455-482.776)^2}{482.776} = 1.60$	$\frac{(477-449.224)^2}{449.224} = 1.72$

5. Calculate the χ^2 statistic
$\chi^2 = 1.39 + 1.50 + 1.60 + 1.72 = 6.21$
6. Calculate the degrees of freedom (df)
The df is $(2-1) \times (2-1) = 1$

Example Cont'd

3. Calculate the expected cell counts

(Expected)	Married	Divorced
Married	$\frac{1068\times1036}{2000}=553.224$	$\frac{1068\times964}{2000}=514.776$
Divorced	$\frac{932\times1036}{2000}=482.776$	$\frac{932\times964}{2000}=449.224$

4. Calculate the partial χ^2 values

partial χ^2	Married	Divorced
Married	$\frac{(581-553.224)^2}{553.224}=1.39$	$\frac{(487-514.776)^2}{514.776}=1.50$
Divorced	$\frac{(455-482.776)^2}{482.776}=1.60$	$\frac{(477-449.224)^2}{449.224}=1.72$

5. Calculate the χ^2 statistic
   $\chi^2 = 1.39 + 1.50 + 1.60 + 1.72 = 6.21$
6. Calculate the degrees of freedom (df)
   The df is $(2-1)\times(2-1)=1$
7. Find the χ^2 critical value with respect to α from the χ^2 table
   The $\chi^2_{\alpha=0.1,df=1} = 2.71$

Example Cont'd

3. Calculate the expected cell counts

(Expected)	Married	Divorced
Married	$\frac{1068\times1036}{2000} = 553.224$	$\frac{1068\times964}{2000} = 514.776$
Divorced	$\frac{932\times1036}{2000} = 482.776$	$\frac{932\times964}{2000} = 449.224$

4. Calculate the partial χ^2 values

partial χ^2	Married	Divorced
Married	$\frac{(581-553.224)^2}{553.224} = 1.39$	$\frac{(487-514.776)^2}{514.776} = 1.50$
Divorced	$\frac{(455-482.776)^2}{482.776} = 1.60$	$\frac{(477-449.224)^2}{449.224} = 1.72$

5. Calculate the χ^2 statistic
$\chi^2 = 1.39 + 1.50 + 1.60 + 1.72 = 6.21$
6. Calculate the degrees of freedom (df)
The df is $(2-1) \times (2-1) = 1$
7. Find the χ^2 critical value with respect to α from the χ^2 table
The $\chi^2_{\alpha=0.1,df=1} = 2.71$
8. Draw your conclusion:
We reject H_0 and conclude that there is a relationship between parents' marital status and childrens' marital status.

Example

The following contingency table contains enrollment data for a random sample of students from several colleges at Purdue University during the 2006-2007 academic year. The table lists the number of male and female students enrolled in each college. Use the two-way table to conduct a χ^2 test from beginning to end. Use $\alpha = .01$

(Observed)	Female	Male	Total
Liberal Arts	378	262	640
Science	99	175	274
Engineering	104	510	614
Total	581	947	1528

Example Cont'd

(Expected)	Female	Male
Liberal Arts	$\frac{640\times581}{1528} = 243.35$	$\frac{640\times947}{1528} = 396.65$
Science	$\frac{274\times581}{1528} = 104.18$	$\frac{274\times947}{1528} = 169.82$
Engineering	$\frac{614\times581}{1528} = 233.46$	$\frac{614\times947}{1528} = 380.54$

partial χ^2	Female	Male
Lib Arts	$\frac{(378-243.35)^2}{243.35} = 74.50$	$\frac{(262-396.65)^2}{396.65} = 45.71$
Sci	$\frac{(99-104.18)^2}{104.18} = 0.26$	$\frac{(175-169.82)^2}{169.82} = 0.16$
Eng	$\frac{(104-233.46)^2}{233.46} = 71.79$	$\frac{(510-380.54)^2}{380.54} = 44.05$

$$\chi^2 = 74.50 + 45.71 + 0.26 + 0.16 + 71.79 + 44.05 = \boxed{236.47}$$

The $df = (3-1)\times(2-1) = 2 \Rightarrow$ Critical value $\chi^2_{\alpha=.01,df=2} = \boxed{9.21}$

Therefore we **reject** H_0 (at .01 level) and conclude that there is a relationship between gender and major.

R Code & Output

```
table <- matrix(c(378, 99, 104,
                  262, 175, 510), 3, 2)
colnames(table) <- c("Female", "Male")
rownames(table) <- c("Liberal Arts", "Science",
"Engineering")
table
```

```
             Female Male
Liberal Arts    378  262
Science          99  175
Engineering     104  510
```

```
chisq.test(table)
```

```
        Pearson's Chi-squared test

data:  table
X-squared = 236.47, df = 2, p-value <
2.2e-16
```

Take Another Look at the Example

(Proportion)	Female	Male	Total
Liberal Arts	.59 (.65)	.41 (.28)	(.42)
Science	.36 (.17)	.64 (.18)	(.18)
Engineering	.17 (.18)	.83 (.54)	(.40)

Rejecting $H_0 \Rightarrow$ conditional probabilities are not consistent with marginal probabilities

Example: Comparing Two Population Proportions

Let $p_1 = \mathrm{P}(Female|Liberal\ Arts)$ and $p_2 = \mathrm{P}(Female|Science)$.

$n_1 = 640, X_1 = 378$, $n_2 = 274, X_2 = 99$

- $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
- $z_{obs} = \frac{.59 - .36}{\sqrt{\frac{.52\times.48}{640} + \frac{.52\times.48}{274}}} = 6.36 > z_{0.025} = 1.96$
- We do have enough statistical evidence to conclude that $p_1 \neq p_2$ at .05% significant level.

R Code & Output

```
prop.test(x = c(378, 99), n = c(640, 274),
          correct = F)
```

```
        2-sample test for equality of
        proportions without continuity
        correction

data:  c(378, 99) out of c(640, 274)
X-squared = 40.432, df = 1, p-value =
2.036e-10
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1608524 0.2977699
sample estimates:
   prop 1    prop 2
0.5906250 0.3613139
```

Example: Test for Homogeneity

Let $p_1 = \mathrm{P}(Liberal\ Arts)$, $p_2 = \mathrm{P}(Science)$, $p_3 = \mathrm{P}(Engineering)$

- The Hypotheses:

  $H_0 : p_1 = p_2 = p_3 = \frac{1}{3}$

  H_a : At least one is different

- The Test Statistic:

$$\chi^2_{obs} = \frac{(640 - 509.33)^2}{509.33} + \frac{(274 - 509.33)^2}{509.33} + \frac{(614 - 509.33)^2}{509.33}$$
$$= 33.52 + 108.73 + 21.51 = 163.76 > \chi^2_{.05,df=2} = 5.99$$

- Rejecting H_0 at .05 level

R Code & Output

```
chisq.test(x = c(640, 274, 614), p = rep(1/3, 3))
```

```
        Chi-squared test for given
        probabilities

data:  c(640, 274, 614)
X-squared = 163.76, df = 2, p-value
< 2.2e-16
```

The Lady Tasting Tea Experiment

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. [...] [It] consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of that the test will consist, namely, that she will be asked to taste eight cups, that these shall be four of each kind [...]. — Fisher, 1935.

R Code & Output


```
TeaTasting <-
matrix(c(3, 1, 1, 3), nrow = 2,
       dimnames = list(Guess = c("Milk", "Tea"),
                       Truth = c("Milk", "Tea")))
TeaTasting
```

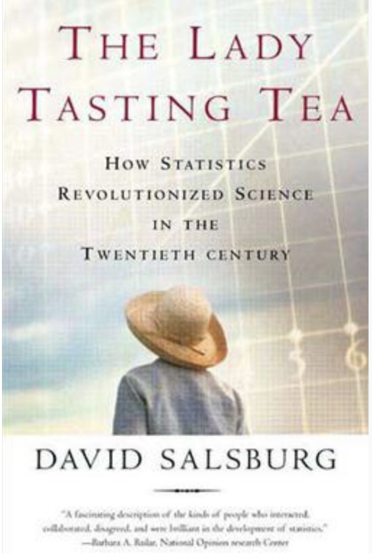
```
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3
```

```
fisher.test(TeaTasting, alternative = "greater")
```

```
        Fisher's Exact Test for Count Data

data:  TeaTasting
p-value = 0.2429
alternative hypothesis: true odds ratio is greater
than 1
```

The Lady Tasting Tea



Aspirin Use and Myocardial Infarction

The table below is from a report on the relationship between aspirin use and heart attract by the Physicians' Health Study Research Group at Harvard Medical School (*New Engl. J. Med.* **318**: 262-264 ,1988).

	Heart Attack		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Here we want to know if the use of aspirin effectively reduces the heart attack rate. To do so we are going to introduce relative risk and odds ratio.

Relative Risk and Odds Ratio

The relative risk (RR) is defined to be the ratio

$$\text{RR} = \frac{p_1}{p_2},$$

where p_i is the probability of "success" for the i$^{\text{th}}$ group.

The odds ratio (θ) is defined as

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)},$$

where Ω is the odds for the i$^{\text{th}}$ group.

Aspirin Use and Heart Attack Revisited: Inference for θ

	Heart Attack		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Point Estimation:

$$\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}},$$

where n_{i1} is the number of successes and n_{i2} is the number of the failures of the i^{th} group. For this example, we have

$$\hat{\theta} = \frac{189 \times 10933}{10845 \times 104} = 1.83$$

$\Rightarrow$ the odds of heart attack for those taking placebo was 1.83 times the odds of those taking aspiring.

Inference for θ Cont'd

Interval Estimation:
Confidence interval is constructed in the nature log scale. We have the Wald confidence interval for $\log\theta$

$$\log\hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\log\hat{\theta}},$$

where the estimated standard error for $\log\hat{\theta}$ is

$$\hat{\sigma}_{\log\hat{\theta}} = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2}.$$

Exponentiating its end points (i.e., the lower limit and the upper limit) provides a confidence interval for θ.

Aspirin Use and Heart Attack: Confidence Interval for θ

Suppose we want to construct a 95% CI for θ:

- $\log \hat{\theta} = 0.6054$
- $\hat{\sigma}_{\log \hat{\theta}} = \left(\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}\right)^{1/2} = 0.1228$
- Margin of error: $z_{0.025} \times \hat{\sigma}_{\log \hat{\theta}} = 1.96 \times 0.1228 = 0.2407$
- CI on the nature log scale:
  $[0.6054 - 0.2407, 0.6054 + 0.2407] = [0.3647, 0.8461]$
- CI on the original scale:
  $[\exp(0.3647), \exp(0.8461)] = [1.4401, 2.3305]$