Lecture 14

Categorical Data Analysis II

Text: Chapter 10

STAT 8020 Statistical Methods II
October 8, 2020

Whitney Huang
Clemson University

Notes

Hypothesis Testing for p

1. State the null and alternative hypotheses:

$$H_0: p = p_0 \text{ vs. } H_a: p > \text{ or } \neq \text{ or } < p_0$$

2. Compute the test statistic:

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

3. Make the decision of the test:
   Rejection Region/ P-Value Methods

4. Draw the conclusion of the test:
   We (do/do not) have enough statistical evidence to conclude that (H_a in words) at α significant level.

Notes

Bird Flu Example Revisited

> Among 900 randomly selected registered voters nationwide, 63% of them are somewhat or very concerned about the spread of bird flu in the United States. Conduct a hypothesis test at .01 level to assess the research hypothesis: $p > .6$.

Notes

Recap: Inference for p

- Point estimate:
$$\hat{p} = \frac{x}{n}$$
where x is the number of "successes" in a sample with sample size n, and the probability of success, p, is the parameter of interest

- $100(1-\alpha)\%$ confidence interval:
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}}$$

- Hypothesis Testing:
$H_0 : p = p_0$ vs. $H_a : p >$ or $\neq$ or $< p_0$
$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$
Under $H_0 : p = p_0$, $\quad z^* \sim \mathrm{N}(0,1)$

Notes

Another CI for p: Wilson Score Confidence Interval

- The actual coverage probability of $100(1-\alpha)\%$ CI $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}}$ is usually falls below $(1-\alpha)$ ☹

- E.B. Wilson proposed one solution in 1927
Idea: Solving $\frac{p-\hat{p}}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_{\alpha/2}$ for p
$$\Rightarrow (p-\hat{p})^2 = z^2_{\alpha/2}\frac{p(1-p)}{n}$$

$100(1-\alpha)\%$ Wilson Score Confidence Interval:
$$\frac{X + \frac{z^2_{\alpha/2}}{2}}{n + z^2_{\alpha/2}} \pm \frac{z_{\alpha/2}}{n + z^2_{\alpha/2}}\sqrt{\frac{X(n-X)}{n} + \frac{z^2_{\alpha/2}}{4}}$$

Notes

Example

Suppose we would like to estimate p, the probability of being vegetarian (for all the CU student). We take a sample with sample size $n = 25$ and none of them are vegetarian (i.e., $x = 0$). Construct a 95% CI for p.

Notes

Rule of Three: An Approximate 95% CI for p When $\hat{p} = 0$ or 1

When $\hat{p} = 0$, we have

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 0 \pm z_{\alpha/2} \times 0 = (0, 0)$$

Similarly, when $\hat{p} = 1$, we have

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 1 \pm z_{\alpha/2} \times 0 = (1, 1)$$

These Wald CIs degenerate to a point , which do not reflect the estimation uncertainty. Here we could apply the rule of three to approximate 95% CI:

$$\begin{cases} (0, 3/n), & \text{if } \hat{p} = 0 \\ (1 - 3/n, 1), & \text{if } \hat{p} = 1 \end{cases}$$

Categorical Data Analysis II

CLEMSON UNIVERSITY

Notes

Comparing Two Population Proportions p_1 and p_2

- We often interested in comparing two groups, e.g., does a particular treatment increase the survival probability for cancer patients ?
- We would like to infer $p_1 - p_2$, the difference between two population proportions ⇒ point estimate, interval estimate, hypothesis testing

Categorical Data Analysis II

CLEMSON UNIVERSITY

Notes

Notation

- Parameters
  - p_1, p_2: population proportions
  - $p_1 - p_2$: the difference between two population proportions
- Sample Statistics
  - n_1, n_2: sample sizes
  - $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$: sample proportions

$$\Rightarrow \hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(\hat{p}_1)(1-\hat{p}_1)}{n_1} + \frac{(\hat{p}_2)(1-\hat{p}_2)}{n_2}}$$

Categorical Data Analysis II

CLEMSON UNIVERSITY

Notes

Point/Interval Estimation for $p_1 - p_2$

- Point estimate:

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

- $100(1-\alpha)\%$ CI based on CLT:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{(\hat{p}_1)(1-\hat{p}_1)}{n_1} + \frac{(\hat{p}_2)(1-\hat{p}_2)}{n_2}}$$

Notes

Hypothesis Testing for $p_1 - p_2$

1. State the null and alternative hypotheses:

   $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 >$ or $\neq$ or < 0

2. Compute the test statistic:

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}},$$

   where $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$

3. Make the decision of the test:

   Rejection Region/ P-Value Methods

4. Draw the conclusion of the test:
   We (do/do not) have enough statistical evidence to conclude that (H_a in words) at α% significant level.

Notes

Example

A Simple Random Simple of 100 CU graduate students is taken and it is found that 79 "strongly agree" that they would recommend their current graduate program. A Simple Random Simple of 85 USC graduate students is taken and it is found that 52 "strongly agree" that they would recommend their current graduate program. At 5 % level, can we conclude that the proportion of "strongly agree" is higher at CU?

Notes

Binomial Experiments and Inference for p

- Fixed number of n trials (sample size), each trial is an independent event (simple random sample)
- Binary outcomes ("success/failure"), where the probability of success, p, for each trial is constant
- The number of successes $X \sim \text{Bin}(n, p)$

We use a random sample x to infer p, the population proportion, using $\hat{p} = \frac{x}{n}$

Notes

Multinomial Experiments and Inference for $\boldsymbol{p} = (p_1, \cdots, p_K)$

- Fixed number of n trials, each trial is an independent event
- K possible outcomes, each with probability $p_k, k = 1, \cdots, K$ where $\sum_{k=1}^{K} p_k = 1$
- $(X_1, X_2, \cdots, X_K) \sim \text{Multi}(n, p_1, p_2, \cdots, p_K)$

We use a random sample $\boldsymbol{x} = (x_1, x_2, \cdots, x_K)$ to infer $\{p_k\}_{k=1}^{K}$, the event probabilities

Question: How many parameters here?

Notes

Example: Multinomial Probability

Suppose that in a three-way election for a large country, candidate 1 received 20% of the votes, candidate 2 received 35% of the votes, and candidate 3 received 45% of the votes. If ten voters are **selected randomly**, what is the probability that there will be exactly two supporter for candidate 1, three supporters for candidate 2 and five supporters for candidate 3 in the sample?

$$\text{P}(X_1 = 2, X_2 = 3, X_3 = 5) = \frac{10!}{2!3!5!}(0.2)^2(0.35)^3(0.45)^5 \approx 0.08$$

Notes

Example: Estimating Multinomial Parameters

> If we **randomly select** ten voters, two supporter for candidate 1, three supporters for candidate 2 and five supporters for candidate 3 in the sample. What would our best guess for the population proportion each candidate would received?

Notes

Pearson's χ^2 Test

- The Hypotheses:
  $H_0 : p_1 = p_{1,0}; p_2 = p_{2,0}; \cdots, p_K = p_{K,0}$
  H_a : At least one is different
- The Test Statistic:

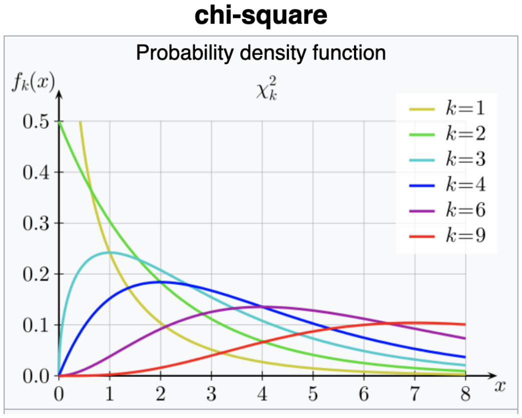
$$\chi^2_* = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k},$$

  where O_k is the observed frequency for the k_{th} event and E_k is the expected frequency under H_0
- The Null Distribution: $\chi^2_* \sim \chi^2_{df=K-1}$
- Assumption: $np_k > 5, k = 1, \cdots, K$

Notes

χ^2-Distribution

chi-square



Notes

Example: Testing Mendel's Theories (pp 22–23, "Categorical Data Analysis" 2_{nd} Ed by Alan Agresti)

"Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain (dominant strain) plants of pure green strain. He predicted that second generation hybrid seeds would be 75% yellow and 25% green. One experiment produced $n = 8023$ seeds, of which $X_1 = 6022$ were yellow and $X_2 = 2001$ were green."

Use Pearson's χ^2 test to assess Mendel's hypothesis.

Notes

Color Preference Example

In Child Psychology, color preference by young children is used as an indicator of emotional state. In a study of 112 children, each was asked to choose "favorite" color from the 7 colors indicated below. Test if there is evidence of a preference at the 5% level.

Color	Blue	Red	Green	White	Purple	Black	Other
Frequency	13	14	8	17	25	15	20

Notes

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

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

- Grand total: 100
- Marginal total for males: 52
- Marginal total for females: 48
- Marginal total for right-handed: 87
- Marginal total for left-handed: 13

This is an example of a contingency table

Notes

Contingency Tables

- Bivariate categorical data is typically displayed in a contingency table
- The number in each cell is the frequency for each category level combination
- Contingency table for the previous example:

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

For a given contingency table, we want to test **if two variables have a relationship or not?** $\Rightarrow$ χ^2-Test

Notes

χ^2-Test for Independence

1. Define the null and alternative hypotheses:

   H_0 : there is no relationship between the 2 variables

   H_a : there is a relationship between the 2 variables

2. (If necessary) Calculate the marginal totals, and the grand total
3. Calculate the expected cell frequencies:

   $$\text{Expected cell frequency} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

4. Calculate the partial χ^2 values (χ^2 value for each cell of the table):

   $$\text{Partial } \chi^2 \text{ value } = \frac{(\text{observed - expected})^2}{\text{expected}}$$

Notes

χ^2-Test for Independence Cont'd

5. Calculate the χ^2 statistic:

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

6. Calculate the degrees of freedom (df)

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

7. Find the χ^2 critical value with respect to α
8. 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

Notes

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?

Notes

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 $\alpha = .10$

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

Notes

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

Notes

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.

Categorical Data Analysis II

CLEMSON UNIVERSITY

Notes

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

Categorical Data Analysis II

CLEMSON UNIVERSITY

Notes

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.

Categorical Data Analysis II

CLEMSON UNIVERSITY

Notes

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
```


Notes

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)
Total	.38	.62	1

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

Notes

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.

Notes

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
```


Notes

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

Notes

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
```


Notes