Unit 5: Inference for categorical data

3. Chi-square testing

Sta 101 - Fall 2015

Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

Slides posted at *http://bit.ly/sta101_f15*

1. Housekeeping

2. Main ideas
    1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
    2. The χ^2 statistic is always positive and right skewed
    3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

- MT 2 next week
  - Bring a calculator + cheat sheet + writing utensil
  - Tables will be provided
- MT 2 review session: Sat, Nov 7, 4-5pm, Old Chem 116
  - + office hours as usual: *https://stat.duke.edu/courses/Fall15/sta101.002/info/#oh*
  - + extra office hours from Dr. Monod: Friday, 1:30-3pm (Old Chem 122A)
- MT 2 review materials posted on the course website
- Project 1 due Friday evening (+ work on it in lab on Thursday)
- PS 5 due Friday evening, RA 5 due Saturday evening (note day change to allow for review before midterm)

1. Housekeeping

2. Main ideas
    1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
    2. The χ^2 statistic is always positive and right skewed
    3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

1. Housekeeping

2. Main ideas
   1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
   2. The χ^2 statistic is always positive and right skewed
   3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

If sample size related conditions are met:

If sample size related conditions are met:

- Categorical data with 2 levels $\rightarrow$ Z

If sample size related conditions are met:

- Categorical data with 2 levels $\rightarrow$ Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions

If sample size related conditions are met:

- Categorical data with 2 levels $\rightarrow$ Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions
- Categorical data with more than 2 levels $\rightarrow \chi^2$

If sample size related conditions are met:

- Categorical data with 2 levels $\rightarrow$ Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions

- Categorical data with more than 2 levels $\rightarrow \chi^2$
  - one variable: χ^2 *test of goodness of fit*, no CI
  - two variables: χ^2 *test of independence*, no CI

If sample size related conditions are met:

- Categorical data with 2 levels $\rightarrow$ Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions
- Categorical data with more than 2 levels $\rightarrow \chi^2$
  - one variable: χ^2 *test of goodness of fit*, no CI
  - two variables: χ^2 *test of independence*, no CI

If sample size related conditions are not met:

If sample size related conditions are met:

- Categorical data with 2 levels $\rightarrow$ Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions
- Categorical data with more than 2 levels $\rightarrow \chi^2$
  - one variable: χ^2 *test of goodness of fit*, no CI
  - two variables: χ^2 *test of independence*, no CI

If sample size related conditions are not met: Simulation based inference (randomization for HT / bootstrapping for CI, when appropriate)

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

Clicker question

Suppose the Gallup poll instead asked about

- party affiliation (Tea Party Republican, Other Republican, and Non-Republican), and
- motivation to vote (extremely unmotivated, very unmotivated, unmotivated, motivated, very motivated, extremely motivated)

We want to find out whether party affiliation is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence

Clicker question

Suppose the Gallup poll instead asked about

- party affiliation (Tea Party Republican, Other Republican, and Non-Republican), and
- motivation to vote (extremely unmotivated, very unmotivated, unmotivated, motivated, very motivated, extremely motivated)

We want to find out whether party affiliation is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence

1. Housekeeping

2. Main ideas
    1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
    2. The χ^2 statistic is always positive and right skewed
    3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

χ^2 statistic: When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E} \qquad \text{where } k = \text{total number of cells}$$

Important points:

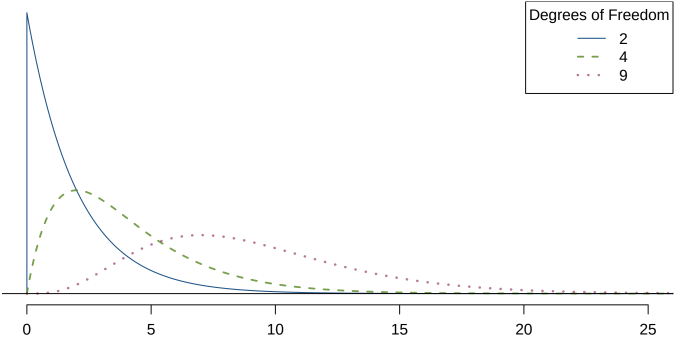
- Use **counts** (not **proportions**) in the calculation of the text statistic, even though we're truly interested in the proportions for inference
- Expected counts are calculated assuming the null hypothesis is true

The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R - 1) \times (C - 1)$

The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R - 1) \times (C - 1)$



p-value = tail area under the chi-square distribution (as usual)

p-value = tail area under the chi-square distribution (as usual)

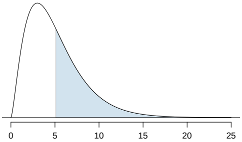
- Using the applet: *https://gallery.shinyapps.io/dist_calc/*

p-value = tail area under the chi-square distribution (as usual)

- Using the applet: *https://gallery.shinyapps.io/dist_calc/*
- Using R: `pchisq()`

p-value = tail area under the chi-square distribution (as usual)

- Using the applet: *https://gallery.shinyapps.io/dist_calc/*
- Using R: `pchisq()`
- Using the table: works a lot like the t table, but only provides upper tail values.



Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	...								

1. Housekeeping

2. Main ideas
    1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
    2. The χ^2 statistic is always positive and right skewed
    3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

1. *Independence:* In addition to what we previously discussed for independence, each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size / distribution:* Each cell must have at least 5 *expected* cases.

Clicker question

Suppose a poll asked the following questions:

- How would you identify your socio-economic status: low, middle, high?
- What type of pet did you have growing up, select all that apply: cat, dog, fish, bird, rodent, none of the above?

What test is most appropriate for evaluating the relationship between these two variables?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence

(e) none of the above

1. Housekeeping

2. Main ideas
    1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
    2. The χ^2 statistic is always positive and right skewed
    3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

Application exercise: 5.3 Chi-square tests

See course website for details.

1. Housekeeping

2. Main ideas
   1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
   2. The χ^2 statistic is always positive and right skewed
   3. At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow$ χ^2 square
2. The χ^2 statistic is always positive and right skewed
3. At least 5 expected successes for χ^2 testing