Unit 5: Inference for categorical data
3. Chi-square testing

Sta 101 - Fall 2018

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at *https://www2.stat.duke.edu/courses/Fall18/sta101.001/*

Outline

Announcements

- PS5 due Friday November 2 11:55pm
- Peer evaluations due Friday November 2 11:55pm

- Midterm 2 Review (lab) session Thursday, November 1
- Midterm 2 Review (lecture) session Monday, November 5 ***(Dr. Abrahamsen giving lecture)***

- Midterm 2 Wednesday, November 7

Announcements

Midterm 2

- **Material Covered:**
  - Unit 3 (Hypothesis Testing of a Single mean with CLT Methods)
  - Unit 4
  - Unit 5
- **What to bring:**
  - Cheat sheet
    - 1 page (8.5" by 11")
    - Front/back ok
    - CAN be typed
  - Calculator (no phones)
- **Provided:**
  - Z-tables, t-tables, Chi-Squared-tables
- **Exam**
  - 4 written questions (like AEs)
  - 5 T/F
  - 10 MC

Midterm 2 Review Suggestions

- **Short answer *review*:**
  - Make sure you understand how to do the application exercises.
  - Review Problem Sets (graded)
- **Short answer *practice*:**
  - Practice test
  - Suggested practice problems in the book
  - Lab Review tomorrow

Midterm 2 Review Suggestions

- **Concept review:**
  - Video notes:
  - Lecture slides **(has material not in the videos/book)**:
  - Readiness Assessments+Performance Assessments:
    - Why are all the other options wrong?
  - What to think about (among other things):
    - Interpretations of analyses (WORDING IS IMPORTANT)
    - Conclusions we would make (WORDING IS IMPORTANT)
    - Relationships between different analyses
    - Know exact definitions
    - FOCUS ON THE WHY BEHIND ANALYSES
    - If there's an equation/analysis, make sure you know how to put that equation/analysis into words in the context of the problem.
    - Common misconceptions (lecture notes)
    - What test to use under certain conditions

1. Housekeeping
2. Main ideas
   1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow \chi^2$ 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 $\rightarrow \chi^2$ 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

Inference for categorical data

If sample size related conditions are met:

Inference for categorical data

If sample size related conditions are met:

- Categorical data with 2 levels → Z

Inference for categorical data

If sample size related conditions are met:

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

Inference for categorical data

If sample size related conditions are met:

- Categorical data with 2 levels → 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 → χ^2

Inference for categorical data

If sample size related conditions are met:

- Categorical data with 2 levels → 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 → χ^2
  - one variable: χ^2 *test of goodness of fit*, no CI
  - two variables: χ^2 *test of independence*, no CI

2

Inference for categorical data

If sample size related conditions are met:

- Categorical data with 2 levels → 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 → χ^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:

2

Inference for categorical data

If sample size related conditions are met:

- Categorical data with 2 levels → 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 → χ^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)

2

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

3

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

- **One Categorical Variable of Interest:** Powerball outcomes 59 levels.
- Want to compare it to a "specified distribution."

3

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?

χ^2 test of goodness of fit

Ho: The data follows the specified distribution
Ha: The data does not follow the specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	?	?	?		?	1

3

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?

χ^2 test of goodness of fit

Ho: The data follows the specified distribution
Ha: The data does not follow the specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

3

χ^2 test of goodness of fit

Ho: The data does resemble a specified distribution
Ha: The data does not resemble a specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590
Data Expected from the Specified Distribution: "Each number is equally likely to be drawn"	10 (ie: *590* *(1/59))	10 (ie: *590* *(1/59))	10 (ie: *590* *(1/59))	...	10 (ie: *590* *(1/59))	*590*

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

3

In a TOY Powerball game players select from a set of 2 green balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 2 green balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

Ho: $p_{ball1} = 1/2$
Ha: $p_{ball1} \neq 1/2$

When only two levels to our categorical variable, this is an equivalent test.

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/2	1/2	1

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Total
Observed Data (from historical data)	294	296	590

3

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

4

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

- Categorical variable with 2 levels
- Categorical variable with 2 levels

$H_0: p_{TPR} = p_{Other}$,
$H_a: p_{TPR} \neq p_{Other}$,
where p = probability of being motivated to vote

4

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?

$H_0: p_{TPR} = p_{Other}$,
$H_a: p_{TPR} \neq p_{Other}$,
where p = probability of being motivated to vote

4

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?

$H_0: p_{TPR} = p_{Other}$,
$H_a: p_{TPR} \neq p_{Other}$, where p= probability of being motivated to vote

Equivalent to saying

$H_0: P(\text{motivated to vote}|TPR) = P(\text{motivated to vote}|other)$,
$H_a: P(\text{motivated to vote}|TPR) \neq P(\text{motivated to vote}|other)$

4

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?

$H_0: p_{TPR} = p_{Other}$,
$H_a: p_{TPR} \neq p_{Other}$, where p= probability of being motivated to vote

Equivalent to saying

$H_0: P(\text{motivated to vote}|TPR) = P(\text{motivated to vote}|other)$,
$H_a: P(\text{motivated to vote}|TPR) \neq P(\text{motivated to vote}|other)$

Equivalent to saying

H_0: Tea Party affiliation is **not associated (or independent)** with being motivated to vote
H_a: Tea Party affiliation is **associated (or dependent)** with being motivated to vote

4

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

5

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

- Categorical variable with >2 levels
- Categorical variable with >2 levels

H_0: Party affiliation and motivation to vote are independent/not associated
H_a: Party affiliation and motivation to vote are dependent/associated

5

Outline

1. Housekeeping
2. Main ideas
   1. Categorical data: 2 levels → Z, >2 levels → 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

The χ^2 statistic

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

6

The χ^2 distribution

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

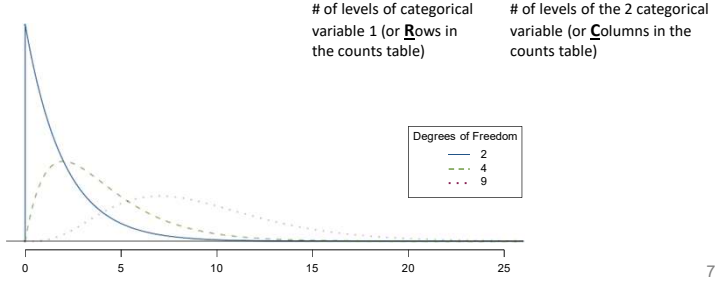
- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R-1) \times (C-1)$

7

The χ^2 distribution

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**: $df = k - 1$ (# of levels of the one categorical variable)
- **For χ^2 independence test**: $df = (R-1) \times (C-1)$
  - R: # of levels of categorical variable 1 (or **R**ows in the counts table)
  - C: # of levels of the 2 categorical variable (or **C**olumns in the counts table)



7

Finding areas under the chi-square curve

p-value = P($\chi^2 \geq \chi^2$- *statistic)*
=tail area under the chi-square distribution (as usual)

Finding areas under the chi-square curve

p-value = P($\chi^2 \geq \chi^2$- *statistic)*
=tail area under the chi-square distribution (as usual)

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

Finding areas under the chi-square curve

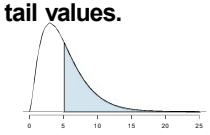
p-value = P($\chi^2 \geq \chi^2$- *statistic)*
=tail area under the chi-square distribution (as usual)

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

Finding areas under the chi-square curve

p-value = P($\chi^2 \geq \chi^2$- *statistic)*
=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
	…								

Outline

1. Housekeeping

2. Main ideas
   1. Categorical data: 2 levels → Z, >2 levels → 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

Conditions for χ^2 testing

Conditions for χ^2 Goodness of Fit Test

1. *Independence:*
   1. Random Sample/Assignment
   2. n<10% of Population
   3. Each case contributes to only one cell in the table.
2. *Sample size / distribution:*
   1. Each **expected cell** must have at least 5 *expected* cases.

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590
Data Expected from the Specified Distribution: "Each number is equally likely to be drawn"	10 (ie: 590 *(1/59))	10 (ie: 590 *(1/59))	10 (ie: 590 *(1/59))	...	10 (ie: 590 *(1/59))	590

9

Conditions for χ^2 testing

Conditions for χ^2 IndependenceTest

1. *Independence:*
   1. For each population: random sampling/assignment
   2. n_1<10% of population 1, n_2<10% of population 2,
   3. Each case contributes to only one cell in the table.
2. *Sample size / distribution:*
   1. Each **expected cell** must have at least 5 *expected* cases.

	dating	cohabiting	married	total
obese	81 (113)	103 (110)	147 (108)	331
not obese	359 (327)	326 (319)	277 (316)	962
total	440	429	424	1293

9

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

10

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*

Condition Violated: Each case MUST contribute to only one cell in the table.

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?

	Had Dog	Had Cat	Had Fish	Had Bird	Had Rodent	Had None	Total
High Socioeconomic Status							
Middle Socioeconomic Status	1	1					
Low Socioeconomic Status							
Total							

Ex: I had dogs and cats growing up. My case contributed twice.->
CONDITIONS FOR χ^2 TEST OF INDEPENDENCE VIOLATED

Outline

1. Housekeeping
2. Main ideas
   1. Categorical data: 2 levels $\rightarrow$ Z, $>$2 levels $\rightarrow \chi^2$ 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.

Outline

1. Housekeeping

2. Main ideas
   1. Categorical data: 2 levels → Z, >2 levels → 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

Summary of main ideas

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

12