Unit 5: Inference for categorical data

3. Chi-square testing

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Abrahamsen

Slides posted at *https://stat.duke.edu/courses/Spring19/sta101.002*

Announcements

- MT 2 on Thursday, Mar 28
  - Bring a calculator + cheat sheet + writing utensil
  - Tables will be provided
- Practice Exam on Sakai
- PS 5 Due Friday
- PA 5 Due Sunday

Inference for categorical data

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

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

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*:

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

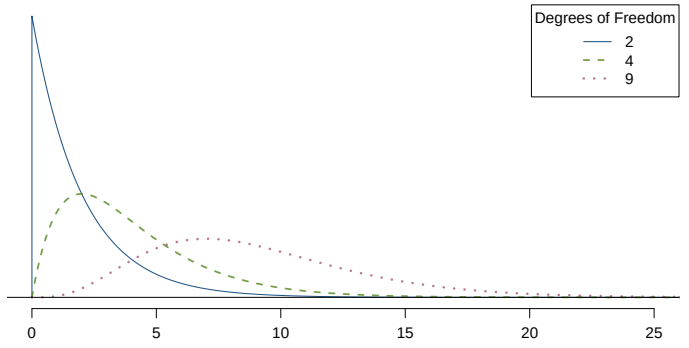
Important points:

- 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

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.

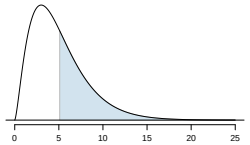
- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R-1) \times (C-1)$



Finding areas under the chi-square curve

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
	...								

Conditions for χ^2 testing

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

Example: χ^2 Tests for Independence

Example: Does money make people happy? (Data from GSS)

Family Income	Not too Happy	Pretty Happy	Very Happy	Total
Above average	26	233	164	423
Average	117	473	293	883
Below average	172	383	132	687
Total	315	1089	589	1993

We want to test if there is an association between money and happiness.

Assumptions:

- SRS (OK since GSS is considered a SRS)
- The **expected** cell count ≥ 5 for all cells.

Hypothesis Testing:

H_0 : Happiness is independent of family income

H_A : Happiness is associated with family income

Test Statistic:

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

p-value: Computed from χ^2 table

df = (# rows - 1)(# columns - 1)

Income & Happiness Example: $df = (3-1) \times (3-1) = 4$.

Computing the Test Statistic:

- Observed counts - given
- Expected counts for each cell:

$$\text{Expected} = \frac{\text{Row total} \times \text{Column total}}{\text{Total}}$$

Expected Counts:

Income	Not too Happy	Pretty Happy	Very Happy	Total
Above Avg	$\frac{423\times315}{1993}$	$\frac{423\times1089}{1993}$	$\frac{423\times589}{1993}$	423
Average	$\frac{883\times315}{1993}$	$\frac{883\times1089}{1993}$	$\frac{883\times589}{1993}$	883
Below Avg	$\frac{687\times315}{1993}$	$\frac{687\times1089}{1993}$	$\frac{687\times589}{1993}$	687
Total	315	1089	589	1993

Computing the Test Statistic:

- Observed counts - given
- Expected counts for each cell:

$$\text{Expected} = \frac{\text{Row total} \times \text{Column total}}{\text{Total}}$$

Expected Counts:

Income	Not too Happy	Pretty Happy	Very Happy	Total
Above Avg	66.86	231.13	125.01	423
Average	139.56	482.48	260.96	883
Below Avg	108.58	375.39	203.03	687
Total	315	1089	589	1993

Contribution to Test Statistic for Each Cell:

$$\frac{(\text{observed - expected})^2}{\text{expected}}$$

$\frac{(26-66.86)^2}{66.86} = 24.97$	$\frac{(223-231.13)^2}{231.13} = 0.02$	$\frac{(164-125.01)^2}{125.01} = 12.16$
$\frac{(117-139.56)^2}{139.56} = 3.65$	$\frac{(473-482.48)^2}{482.48} = 0.186$	$\frac{(293-260.96)^2}{260.96} = 3.93$
$\frac{(172-108.58)^2}{108.58} = 37.04$	$\frac{(383-375.39)^2}{375.39} = 0.15$	$\frac{(132-203.02)^2}{203.02} = 24.85$

Test Statistic: Add up all values in the table

$$\chi^2_{calc} = 24.97 + 0.02 + 12.16 + 3.65 + 0.186 + 3.93 + 37.04 + 0.15 + 24.85 = 106.96$$

Example: χ^2 Tests for Independence

p-value: From χ^2 table with $(3-1) \times (3-1) = 4$, p-value ≈ 0

Conclusion: Reject H_0 at all α-levels and conclude that there is an association between Happiness and Income.

Example: χ^2 Tests for Independence

To see what type of relationship there is between Happiness and Income, **compute the residuals**.

residuals = observed - expected

Residuals

Income	Not too Happy	Pretty Happy	Very Happy
Above Avg	$26 - 66.86$	$233 - 231.13$	$164 - 125.01$
Average	$117 - 139.56$	$473 - 482.48$	$293 - 260.96$
Below Avg	$172 - 108.58$	$383 - 375.39$	$132 - 203.03$

Above Average Income: We observe fewer than expected Not too Happy people and higher than expected Very Happy people.

Below Average Income: We observe higher than expected Not too Happy people and fewer than expected Very Happy people.

Example: χ^2 Tests for Independence

Conclusion from Residuals: We see that less income is associated with lower levels of happiness, higher income with greater happiness. **HOWEVER**, we can **NOT** say money makes you happy (no causal effect).

Application exercise: 5.3 Chi-square tests

See course website for details.

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