Unit 5: Inference for categorical data

2. Inference for comparing two proportions

STA 104 - Summer 2017

Duke University, Department of Statistical Science

Prof. van den Boom

Slides posted at

http://www2.stat.duke.edu/courses/Summer17/sta104.001-1/

- Example MT2 posted on the course website

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(mean = (p_1 - p_2), SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Conditions:

- Independence: Random sample/assignment + 10% rule
- Sample size / skew: At least 10 successes and failures

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

Expected proportion of success for both groups when $H_0 : p_1 = p_2$ is defined as the *pooled proportion*:

$$\hat{p}_{pool} = \frac{\text{total successes}}{\text{total sample size}} = \frac{suc_1 + suc_2}{n_1 + n_2}$$

Clicker question

Suppose in group 1, 30 out of 50 observations are successes, and in group 2, 20 out of 60 observations are successes. What is the pooled proportion?

(a) $\frac{30}{50}$
(b) $\frac{20}{60}$
(c) $\frac{30}{50} + \frac{20}{60}$
(d) $\frac{30+20}{50+60}$
(e) $\frac{\frac{30}{50} + \frac{20}{60}}{2}$

When S-F fails, simulate!

- If the S-F condition is met, can do theoretical inference: Z test, Z interval
- If the S-F condition is not met, must use simulation based methods: randomization test, bootstrap interval

On the perils of living dangerously in the slasher horror film

Abstract: *The slasher horror film has been deplored based on claims that it depicts eroticized violence against predominately female characters as punishment for sexual activities. To test this assertion, a quantitative content analysis was conducted to examine the extent to which gender differences are evident in the association between character survival and engagement in sexual activities. Information pertaining to gender, engagement in sexual activities, and survival was coded for film characters from a simple random sample of 50 English-language, North American slasher films released between 1960 and 2009.*

Welsh, Andrew. "On the perils of living dangerously in the slasher horror film: Gender differences in the association between sexual activity and survival." Sex Roles 62.11-12 (2010): 762-773.

Males...

Is survival for **male** characters in slasher films associated with sexual activity?

Gender	Sexual activity	Outcome of physical aggression		n
		Survival	Death	
Female				
	Present	13.3% (n=11)	86.7% (n=72)	83
	Absent	28.1% (n=39)	71.9% (n=100)	139
Male				
	Present	9.5% (n=7)	90.5% (n=67)	74
	Absent	14.8% (n=28)	85.2% (n=161)	189

$H_0 : p_{\text{sex present}} = p_{\text{sex absent}}$ $\quad H_A : p_{\text{sex present}} \neq p_{\text{sex absent}}$

1. Independence: The movies are randomly selected, but the characters are not. Characters featured in the same movie may not be independent, but we'll make a simplifying assumption and ignore this potential.
2. Success-failure: ?

Clicker question

Assuming that the null hypothesis ($H_0 : p_{sex\ present} = p_{sex\ absent}$) is true, which of the following is the pooled proportion of characters who survived?

(a) $\frac{7}{74} = 0.095$
(b) $\frac{28}{189} = 0.148$
(c) $\frac{7}{74+189} = 0.027$
(d) $\frac{7+28}{74+189} = 0.133$
(e) $\frac{67+161}{74+189} = 0.867$

Gender	Sexual activity	Outcome of physical aggression		*n*
		Survival	Death	
Female				
	Present	13.3% (*n*=11)	86.7% (*n*=72)	83
	Absent	28.1% (*n*=39)	71.9% (*n*=100)	139
Male				
	Present	9.5% (*n*=7)	90.5% (*n*=67)	74
	Absent	14.8% (*n*=28)	85.2% (*n*=161)	189

Clicker question

Assuming that the null hypothesis ($H_0 : p_{sex\ present} = p_{sex\ absent}$) is true, how many males characters involved in sexual activity are expected to survive?

(a) $0.133 \times (28 + 7) = 4.655$
(b) $0.133 \times 74 = 9.842$
(c) $0.133 \times (74 + 189) = 34.979$
(d) 7
(e) $7 + 28 = 35$

Gender	Sexual activity	Outcome of physical aggression		*n*
		Survival	Death	
Female				
	Present	13.3% (*n*=11)	86.7% (*n*=72)	83
	Absent	28.1% (*n*=39)	71.9% (*n*=100)	139
Male				
	Present	9.5% (*n*=7)	90.5% (*n*=67)	74
	Absent	14.8% (*n*=28)	85.2% (*n*=161)	189

Simulation scheme

1. Use 263 index cards, where each card represents a male character in a slasher film in the sample.
2. Mark 35 of the cards as “survival” and the remaining 228 as “death”.
3. Shuffle the cards and split into two groups of size 74 and 189, for sexual activity present and absent, respectively.
4. Calculate the difference between the proportions of “survival” in the sexual activity present and absent groups.
5. Repeat steps (3) and (4) many times to build a randomization distribution of differences in simulated proportions.

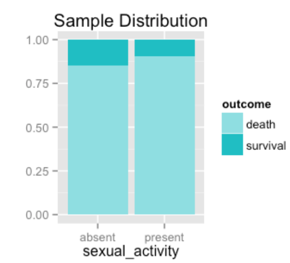
Simulate in R

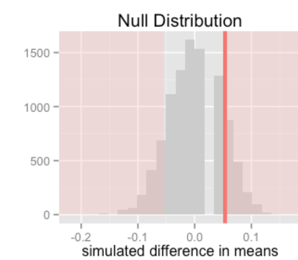
```
# read and subset data for males
slasher <- read.csv("https://stat.duke.edu/~mc301/data/slasher.csv")
slasher_m <- slasher %>%
  filter(gender == "male")

# load inference function
load(url("https://stat.duke.edu/~mc301/R/fun/inference.RData"))

# run the hypothesis test
inference(y = outcome, x = sexual_activity, data = slasher_m,
          success = "survival", statistic = "proportion",
          type = "ht", null = 0, alternative = "twosided",
          method = "simulation", seed = 66613)
```

```
Response variable: categorical (2 levels), Explanatory variable: categorical (2 levels)
n_absent = 189, p_hat_absent = 0.1481
n_absent = 74, p_hat_absent = 0.0946
H0: p_absent =  p_absent
HA: p_absent != p_absent
p_value = 0.3465
```





Application exercise: App Ex 5.2

See course website for details.

1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
2. For HT where $H_0 : p_1 = p_2$, pool!
3. When S-F fails, simulate!