Sta 111 - Summer II 2017
Probability and Statistical Inference

14. Difference of two proportions

Lu Wang

Duke University, Department of Statistical Science

July 25, 2017

Outline

1. Inference for $p_1 - p_2$
   1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
   2. Confidence intervals for $p_1 - p_2$
   3. Hypothesis testing for $p_1 - p_2$

      For HT where $H_0 : p_1 = p_2$, pool!

2. When S-F fails, simulate!

3. Summary

4. Homework 4

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

(a) A great deal
(b) Some
(c) A little
(d) Not at all

The GSS asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of students at Duke University:

	GSS	Duke
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

- *Parameter of interest:* Difference between the proportions of *all* Duke students and *all* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

- *Point estimate:* Difference between the proportions of *sampled* Duke students and *sampled* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{Duke} - \hat{p}_{US}$$

- The details are the same as before...
- CI: *point estimate* $\pm$ *margin of error*
- HT: Use $Z = \frac{\text{point estimate} - \text{null value}}{SE}$ to find appropriate p-value.
- We just need the appropriate standard error of the point estimate ($SE_{\hat{p}_{Duke} - \hat{p}_{US}}$), which is the only new concept.

Standard error of the difference between two sample proportions

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

CLT conditions for constructing CI for difference of proportions

1. *Independence within groups:*
   - The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well.
   - $n_{Duke} <$ 10% of all Duke students and 680 $<$ 10% of all Americans.
   - Hence we can assume that the attitudes of Duke students in the sample are independent of each other, and attitudes of US residents in the sample are independent of each other as well.
2. *Independence between groups:* The sampled Duke students and the US residents are independent of each other.
3. *Success-failure:*
   At least 10 observed successes and 10 observed failures in the two groups.

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

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap, $p_{Duke} - p_{US}$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
proportion	0.657	0.668

Practice

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

(a) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} \neq p_{US}$

(b) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
$H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$

(c) $H_0 : p_{Duke} - p_{US} = 0$
$H_A : p_{Duke} - p_{US} \neq 0$

(d) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} < p_{US}$

- When constructing a confidence interval for a population proportion, we check if the *observed* number of successes and failures are at least 10.

$$n\hat{p} \geq 10 \qquad\qquad n(1 - \hat{p}) \geq 10$$

- When conducting a hypothesis test for a population proportion, we check if the *expected* number of successes and failures are at least 10.

$$np_0 \geq 10 \qquad\qquad n(1 - p_0) \geq 10$$

For HT where $H_0 : p_1 = p_2$, pool! - pooled estimate of a proportion

- In the case of comparing two proportions where $H_0 : p_1 = p_2$ (almost always for HT), there isn't a given null value for p_1 or p_2 that we can use to calculate the *expected* number of successes and failures in each sample.
- Therefore, we need to first find a common (*pooled*) proportion for the two groups, and use that in our analysis.
  - verify the success-failure condition
  - estimate the standard error
- This simply means finding the proportion of total successes among the total number of observations.

Pooled estimate of a proportion

$$\hat{p} = \frac{\#\ \textit{of successes}_1 + \#\ \textit{of successes}_2}{n_1 + n_2}$$

Practice

Calculate the estimated <u>pooled proportion</u> of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{Duke}$ or $\hat{p}_{US}$) the pooled estimate is closer to? Why?

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
proportion	0.657	0.668

$$\begin{aligned}\hat{p} &= \frac{\#\ \textit{of successes}_1 + \#\ \textit{of successes}_2}{n_1 + n_2} \\ &= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666\end{aligned}$$

HT for difference of proportions - p value

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\
&= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22 \\
p - value &= 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82
\end{aligned}
$$

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

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- CLT conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 10 successes and failures in each group
    - if not $\rightarrow$ randomization (Ch 6.4)
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - for CI: use $\hat{p}_1$ and $\hat{p}_2$
  - for HT:
    - when $H_0 : p_1 = p_2$: use $\hat{p}_{pool} = \frac{\#\ suc_1 + \# suc_2}{n_1 + n_2}$
    - when $H_0 : p_1 - p_2 =$ *(some value other than 0)*: use $\hat{p}_1$ and $\hat{p}_2$
      - this is pretty rare

Recap - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use s.
- When working with proportions,
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

Graded questions:

- Ch 6: 6.12, 6.20, 6.28, 6.30, 6.44, 6.48

Practice questions:

- Single proportion: 6.1, 6.3, 6.5, 6.11, 6.15, 6.19, 6.21
- Comparing two proportions: 6.23, 6.25, 6.29, 6.33, 6.35
- Inference for proportions via simulation: 6.51, 6.53, 6.55
- Comparing three or more proportions (Chi-square): 6.39, 6.41, 6.43, 6.45, 6.47