Lecture 21: Difference of two proportions

Chapter 6.2

Question for today

How do we infer about a difference in proportions $p_1 - p_2$?

Surveys

The way a question is phrased in survey can influence a person's response. Ex on p.269: the Pew Research Center conducted a survey with the following question:

By 2014 all Americans will be required to have health insurance. X while Y. Do you approve of disapprove of this policy?

where X and Y were randomly ordered between

- People who do not buy insurance will pay a penalty
- People who cannot afford it will receive financial help from the government

Let's infer about the difference in proportion of people who approve. Any guesses which is higher?

Example from Text

	Sample size n_i	Approve (%)	Disapprove (%)	Other (%)
people who do not buy it will pay a penalty given first	771	47	49	3
people who cannot afford it will receive financial help from the gov't given first	732	34	63	3

Example from Text

	Sample size n_i	Approve (%)	Don't Approve (%)
people who do not buy it will pay a penalty given first	771	47	53
people who cannot afford it will receive financial help from the gov't given first	732	34	66

So $\widehat{p}_1 - \widehat{p}_2 = 0.47 - 0.34 > 0$: people are more likely to support Obamacare in the first scenario.

Conditions...

When

- Both sample proportions $\widehat{p}_1$ and $\widehat{p}_2$ are approximately normal:
  - independence
  - success/failure condition: at least 10 successes and failures
- the two samples are independent from each other

... for Sampling Dist'n of $\widehat{p}_1 - \widehat{p}_2$ Being Normal

The sampling distribution of $\widehat{p}_1 - \widehat{p}_2$ is approximately Normal with

- mean $p_1 - p_2$
- standard error

$$SE_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{SE^2_{\widehat{p}_1} + SE^2_{\widehat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Standard Error

Recall we showed that the SE for $\overline{x}_1 - \overline{x}_2$ was

$$SE_{\overline{x}_1 - \overline{x}_2} = \sqrt{SE^2_{\overline{x}_1} + SE^2_{\overline{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Compare this to

$$SE_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{SE^2_{\widehat{p}_1} + SE^2_{\widehat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

What p_1 & p_2?

What p_1 & p_2 do we

- Use to check success/failure condition?
- Use in $SE_{\widehat{p}_1 - \widehat{p}_2}$?

For

- Confidence intervals: plug in $\widehat{p}_1$ and $\widehat{p}_2$
- Hypothesis tests: plug in pooled estimate $\widehat{p}$

Confidence Intervals

What is a 90% confidence interval for the difference in proportions?

Check the conditions:

- Normality for each group
  - Independence: both groups $\leq$ 10% of respective populations
  - The success/failure condition for both groups:
    - Group 1: 362 successes and $771 - 362 = 409$ failures
    - Group 2: 249 successes and 483 failures
- We assume both groups were sampled independently.

Confidence Intervals

- Point estimate is $\widehat{p}_1 - \widehat{p}_2 = 0.47 - 0.34 = 0.13$
- Plug in $\widehat{p}_1$ and $\widehat{p}_2$ into SE:

$$SE_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} = \ldots = 0.025$$

- A 90% confidence interval for $p_1 - p_2$ is:

$$\text{point estimate} \pm z^* \times SE = 0.13 \pm 1.65 \times 0.025 = (0.09, 0.17)$$

Interpretation

Two key observations:

- $(9\%, 17\%)$ does not contain 0, suggestive of a true difference.
- The sign of the difference: $\widehat{p}_1 - \widehat{p}_2 = 0.13 > 0$

More support Obamacare if stated as follows:

People who do not buy it will pay a penalty while people who cannot afford it will receive financial help from the government.

Hypothesis Tests

Now we are interested in testing the difference of two proportions:

$$\begin{aligned} & H_0 : p_1 - p_2 = 0 \\ \text{vs} \quad & H_1 : p_1 - p_2 \neq 0 \end{aligned}$$

Note this can be re-expressed as:

$$\begin{aligned} & H_0 : p_1 = p_2 \\ \text{vs} \quad & H_1 : p_1 \neq p_2 \end{aligned}$$

i.e. under H_0 the two proportions are both equal to some value p:

$$p_1 = p_2 = p$$

Hypothesis Tests

So to

- Verify the success-failure condition
- Compute the standard SE

we use a pooled estimate $\widehat{p}$ of the proportion p. i.e. as if there were no difference between them, so we can combine them:

$$\widehat{p} = \frac{\text{total } \# \text{ of successes}}{\text{total } \# \text{ of cases}}$$

The SE to use is:

$$SE_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2}} = \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Exercise 6.31 on Page 305

A 2010 survey asked 827 randomly sample voters in California "How do you feel about drilling for oil and natural gas off the coast of California?"

	College Grad Yes	College Grad No
Support	154	132
Oppose	180	126
Don't Know	104	131
Total	438	389

Test at the $\alpha = 0.10$ significance level if the proportion of college graduates who support off-shore drilling is different than that of non-college graduates.

Exercise 6.31 on Page 305

The pooled estimate is $\widehat{p} = \frac{154+132}{438+389} = 0.346$. Check the conditions:

1. Normality of both point estimates
   - Independence
     - $n_1 = 438 \leq 10\%$ of pop. of CA college grads
     - $n_2 = 389 \leq 10\%$ of pop. of CA non college grads
   - Success/failure: both groups have at least 10 successes and 10 failures.
2. We assume that both groups are sampled independently.

Exercise 6.31 on Page 305

- Point estimate $\widehat{p}_1 - \widehat{p}_2 = 0.352 - 0.339 = 0.013$
- $SE_{\widehat{p}_1 - \widehat{p}_2} \approx \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.033$
- Test statistic: z-score of $\widehat{p}_1 - \widehat{p}_2$ under $H_0 : p_1 - p_2 = 0$

$$z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.013 - 0}{0.033} = 0.392$$

- p-value: 0.6922. i.e. we fail to reject H_0. We don't have strong evidence of a difference in support.

Jury Selection

Preview of next lecture: In many trials a big issue is the racial makeup of the jury.

Question: is there a way to figure out if there is a racial bias in jury selection?

Jury Selection

Say we have a juror pool (registered voters) where the racial breakdown is:

Race	White	Black	Hispanic	Other	Total
Registered Voters	72%	7%	12%	9%	100%

Jury Selection

If we pick $n = 100$ jurors at random (i.e. unbiasedly), we expect the breakdown of counts to be:

Race	White	Black	Hispanic	Other	Total
Registered Voters	72%	7%	12%	9%	100%
Representation	72	7	12	9	$n = 100$

Jury Selection

Say we observe the following counts:

Race	White	Black	Hispanic	Other	Total
Registered Voters	72%	7%	12%	9%	100%
Representation	0	0	100	0	$n = 100$

Fairly obvious bias in juror selection!

Jury Selection

But what about the following? Is there a bias? i.e. a non-random mechanism at play?

Race	White	Black	Hispanic	Other	Total
Registered Voters	72%	7%	12%	9%	100%
Representation	75	6	11	8	$n = 100$

Next Two Lectures

Chi-square tests are used to compare expected counts with observed counts.

Two tests we'll see:

- Goodness-of-fit tests: for frequency tables
- Tests for independence: for contingency/two-way tables