Lecture 21: Difference of two proportions

Chapter 6.2

Previously... Proportions

If

- The sample observations are independent
- The success-failure condition holds:
  - $np \geq 10$
  - $n(1-p) \geq 10$

then the sampling distribution of $\widehat{p}$ is nearly normal and hence we can:

- construct confidence intervals using z^*
- conduct hypothesis tests and compute p-values using z-tables

Previously... Proportions

Both

- to verify success/failure conditions
- in estimate of SE

we require an estimate of the true proportion p. For

- confidence intervals: use sample proportion $\widehat{p}$ in place of p
- hypothesis tests: use null value p_0 in place of p

Question for today

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

Confidence Interval: Example from Text

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

As you may know, 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

Build a 90% confidence interval for the difference in proportions.

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

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

When

- both sample proportions $\widehat{p}_1$ and $\widehat{p}_2$ are approximately normal:
  - are independent
  - satisfies the success/failure condition from Lecture 9.1: $np \geq 10$ and $n(1-p) \geq 10$
- the samples are independent from each other...

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

...the sampling distribution for the difference of sample proportions $\widehat{p}_1 - \widehat{p}_2$ is approximately normal with

- mean $p_1 - p_2$
- standard error

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

where p_1, p_2 are the true population proportions, and n_1, n_2 are the sample sizes.

Standard Error

Recall when looking at numerical data, 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

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

Confidence Intervals

Check the conditions:

- Normality for each group
  - Each group is a sample random sample from less than 10% of the population
  - The success/failure condition holds for both samples separately:
    - $n_1\widehat{p}_1 \geq 10$ and $n_1(1-\widehat{p}_1) \geq 10$
    - $n_2\widehat{p}_2 \geq 10$ and $n_2(1-\widehat{p}_2) \geq 10$
- We assume both groups were sampled independently from each other.

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:

$$\begin{aligned} SE_{\widehat{p}_1-\widehat{p}_2} &= \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \\ &= \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025 \end{aligned}$$

Confidence Intervals

A 90% confidence interval using the normal model is, as always:

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

Since the confidence interval does not contain 0,

Hypothesis Tests of $H_0 : p_1 = p_2$

We are typically interested in differences in proportions being 0 vs something else. For example

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

Note the null hypothesis can be re-expressed as $H_0 : p_1 = p_2$.

Thus, under the null hypothesis the two proportions are equal. i.e. $p_1 = p_2 = p$

Hypothesis Tests of $H_0 : p_1 = p_2$

So to

- verify the success-failure conditions
- compute the standard SE

we use a pooled estimate $\widehat{p}$ of the proportion p

$$\widehat{p} = \frac{\text{number of successes}}{\text{number of cases}} = \frac{\widehat{p}_1 n_1 + \widehat{p}_2 n_2}{n_1 + n_2}$$

The estimate of the SE to use in for this hypothesis test is:

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

Exercise 6.31

A 2010 survey asked 827 randomly sample voters in California "Do you support/oppose/don't know drilling for oil and natural gas off the coast of California?" The responses were:

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

Conduct a hypothesis test at the $\alpha = 0.01$ significance level to determine if the proportion of college graduates who support off-shore drilling in California is different than that of non-college graduates.

Next Time

Chi-square tests for

- Goodness-of-fit
- Independence of two variables

Jury Selection

In both sensational trials a big issue was the racial makeup of the jury.

The question we ask is: is there a way to figure out if there is a racial bias in jury selection?

Jury Selection

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

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

Jury Selection

Say we had $n = 100$ people picked as jurors, we expect the breakdown 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 breakdown. Fairly obvious bias in juror selection!

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

Jury Selection

But what about the following? We expected 72 whites, but observe 75. 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	$n = 100$