Difference of Two Proportions

confidence interval

hypothesis testing

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

Results from the GSS

The GSS asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of introductory statistics 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

response	p_gss	p_duke
A great deal	0.67	0.66
Some	0.18	0.29
A little	0.08	0.04
Not at all	0.07	0.02

How do Duke students and the American public at large compare with respect to their concern on melting ice cap?

Parameter of interest

Difference between the proportions of *all* Duke student 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.

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

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
$\hat{p}$	0.66	0.67

Inference for comparing proportions

- The details are the same as before…
- CI: *point estimate ± margin of error*
- HT: Use *Z = (point estimate - 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

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

Sample proportions are also nearly normally distributed ?

Conditions for inference for comparing two independent proportions:

1. Independence:

- **within gorups** : sampled obervations must be independent within each group
  - random sample/assignment
  - if sampling without replacement, n<10% of population
- **between groups**: two groups must be independent of each other (non-paired)

2. Sample size/skew: Each sample should meet the success-failure condition:

$$n_1 p_1 \geq 10 \; and \; n_1(1-p_1) \geq 10$$

$$n_2 p_2 \geq 10 \; and \; n_2(1-p_2) \geq 10$$

Conditions for 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.
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.

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
$\hat{p}$	0.657	0.668

CI: *point estimate ± margin of error*

$$\hat{p}_1 - \hat{p}_2 \pm z^* SE_{\hat{p}_1 - \hat{p}_2}$$

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm 1.96 \times 0.0497$$

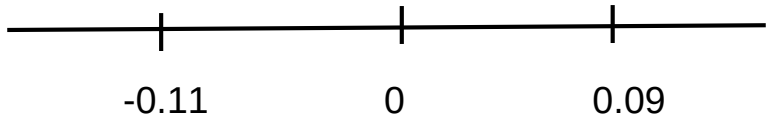
$$= -0.011 \pm 0.097$$

$$= (-0.108, 0.086)$$

Based on the confidence interval we calculated , should we expect to find a significant difference (at the equivalent significance level) between the population proportions of Duke students and the American public at large who would be bothered a great deal by the melting of the northern ice cap?

95% CI for $p_{Duke} - p_{US}$ =(-.11, 0.09).

$H_0 : p_{Duke} - p_{US} = 0$ $H_1 : p_{Duke} - p_{US} \neq 0$



cannot reject H_0

Hypothesis test for comparing two proportions

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

Both (a) and (c) are correct.

1. Set the hypotheses :

2. Calculate the point estimate

3. Check conditions:

Hypothesis test for comparaing two proportions

- The details are the same as before…
- CI: *point estimate ± margin of error*
- HT: Use *Z = (point estimate - 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

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

flashback to working with one proportion

	confidence interval	hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np_0 \geq 10$ $n(1-p_0) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

working with two proportions

	observed	*expected*
	confidence interval	hypothesis test
success-failure condition	$n_1 \widehat{p_1} \geq 10$ $\quad n_2 \widehat{p_2} \geq 10$ $n_1(1 - \widehat{p_1}) \geq 10$ $\quad n_2(1 - \widehat{p_2}) \geq 10$	$H_0: p_1 = p_2$
standard error	$SE = \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$	

pooled proportion

- In the case of comparing two proportions where H_0: $p_1 = p_2$, there isn't a given null value we can use to calculated 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.
- This simply means finding the proportion of total successes among the total number of observations.

Pooled proportion

$$\hat{p}_{pool} = \frac{total\ number\ of\ successes}{total\ n}$$
$$= \frac{\sharp\ of\ successes_1 + \sharp\ of\ successes_2}{n_1 + n_2}$$

Calculate the estimated **pooled proportion** 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
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

$$= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666$$

working with two proportions

	observed	*expected*
	confidence interval	hypothesis test
success-failure condition	$n_1\,\hat{p}_1 \geq 10$ $\quad n_2\hat{p}_2 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $\quad n_2(1-\hat{p}_2) \geq 10$	$n_1\,\hat{p}_{pool} \geq 10$ $\quad n_2\hat{p}_{pool} \geq 10$ $n_1(1-\hat{p}_{pool}) \geq 10$ $\quad n_2(1-\hat{p}_{pool}) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$

Hypothesis test for comparing two proportions

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.

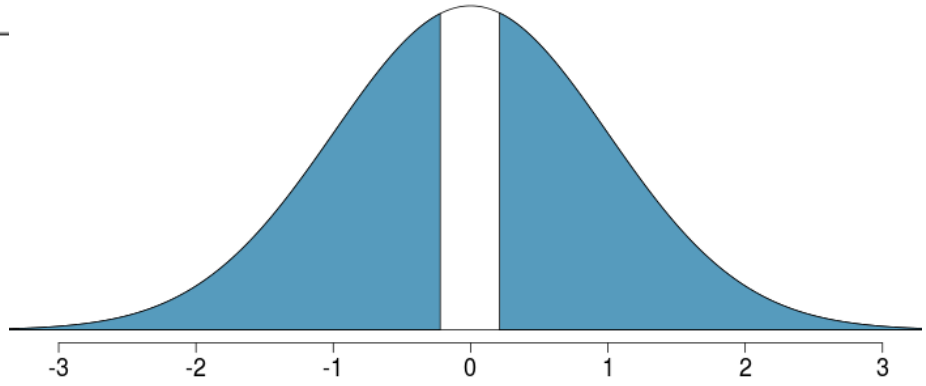
Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$H_0 : p_{Duke} - p_{US} = 0$ $H_1 : p_{Duke} - p_{US} \neq 0$

$$Z = \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}}$$

$$= \frac{-0.011}{0.0495} = -0.22$$



$$p - value = 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82$$

Recap - comparing two proportions

- Population parameter: (p_1 - p_2), point estimate: ($\hat{p}_1$ - $\hat{p}_2$)

- 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 → randomization (Section 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

Assignment

A SurveyUSA poll asked respondents whether any of their children have ever been the victim of bullying. Also recorded on this survey was the gender of the respondent (the parent). Below is the distribution of reponses by gender of the respondent.

	Male	Female
Yes	34	61
No	52	61
Not sure	4	0
Total	90	122

a) Calculate a 95% confidence interval for the gender difference in the response “Yes” to the question about whether any of their children have ever been the victim of bullying.

b) Conduct a hypothesis test, at 5% significance level evaulating if males and females are equally likely to answer “Yes” ” to the question about whether any of their children have ever been the victim of bullying.