Chapter 6: Inference for Categorical Data

Math 140 · Fall '21

Based on content in OpenIntro Stats, 4th Ed

Hitchman

October 18, 2021

Section 6.1
Inference for one proportion

(mostly a nice review of Chapter 5)

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

(a) All 1000 get the drug
(b) 500 get the drug, 500 don't

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

(a) All 1000 get the drug
(b) *500 get the drug, 500 don't*

Results from the GSS

The General Social Survey asks the same question, below is the distribution of responses from the 2010 survey:

All 1000 get the drug	99
500 get the drug 500 don't	571
Total	670

Parameter and point estimate

Q: *We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"? What are the parameter of interest and the point estimate?*

Parameter and point estimate

Q: *We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"? What are the parameter of interest and the point estimate?*

- Parameter of interest: *Proportion of all Americans who have good intuition about experimental design.*

$$p \text{ (a population proportion)}$$

Parameter and point estimate

Q: *We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"? What are the parameter of interest and the point estimate?*

- Parameter of interest: *Proportion of all Americans who have good intuition about experimental design.*

$$p \text{ (a population proportion)}$$

- Point estimate: *Proportion of sampled Americans who have good intuition about experimental design.*

$$\hat{p} \text{ (a sample proportion)}$$

Inference on a proportion

Q: *What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"?*

Inference on a proportion

Q: *What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"?*

- *We can answer this research question using a confidence interval, which we know is always of the form*

$$\textit{point estimate} \pm \textit{MOE}$$

Inference on a proportion

Q: *What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"?*

- *We can answer this research question using a confidence interval, which we know is always of the form*

$$\textit{point estimate} \pm \textit{MOE}$$

- *And we also know that* $\textit{MOE} = \textit{critical value} \times \textit{standard error}$ *of the point estimate.*

Inference on a proportion

Q: *What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"?*

- *We can answer this research question using a confidence interval, which we know is always of the form*

$$\text{point estimate} \pm MOE$$

- *And we also know that* $MOE = \text{critical value} \times \text{standard error}$ *of the point estimate.*

Standard error of a sample proportion

$$SE_{\hat{p}} = \sqrt{\frac{p\,(1-p)}{n}}$$

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

If certain conditions are met, sample proportions will be nearly normally distributed with mean equal to the pop'n proportion, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

$$\hat{p} \sim N\left(p, \sqrt{\frac{p\ (1-p)}{n}}\right)$$

Conditions:

1. *Independence:* Sampled observations must be independent.
2. *Sample size:* There should be at least 10 successes and 10 failures in the observed sample.

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

If certain conditions are met, sample proportions will be nearly normally distributed with mean equal to the pop'n proportion, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

$$\hat{p} \sim N\left(p, \sqrt{\frac{p\ (1-p)}{n}}\right)$$

Conditions:

1. *Independence:* Sampled observations must be independent.
2. *Sample size:* There should be at least 10 successes and 10 failures in the observed sample.

Note: If p is unknown (most cases), we use $\hat{p}$ in the calculation of the standard error.

Back to experimental design...

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?*

Back to experimental design...

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?*

Given: $n = 670, \hat{p} = 0.85$. *First, check conditions.*

Back to experimental design...

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?*

Given: $n = 670, \hat{p} = 0.85$. *First, check conditions.*

1. Independence: *The sample is random, and 670 < 10% of all Americans, therefore we can assume that one respondent's response is independent of another.*

Back to experimental design...

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?*

Given: $n = 670, \hat{p} = 0.85$. *First, check conditions.*

1. Independence: *The sample is random, and 670 < 10% of all Americans, therefore we can assume that one respondent's response is independent of another.*
2. Success-failure: *571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.*

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

(a) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

(b) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

(c) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

(d) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

(a) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}} \rightarrow (0.82, 0.88)$

(b) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

(c) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

(d) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

$$MOE = z^{\star} \times SE$$

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

$$MOE = z^\star \times SE$$

$$0.01 \quad \geq \quad 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \quad \rightarrow \textit{Use } \hat{p} \textit{ from previous study}$$

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

$$MOE = z^{\star} \times SE$$

$$\begin{aligned} 0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \quad \rightarrow \textit{Use } \hat{p} \textit{ from previous study} \\ 0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \end{aligned}$$

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

$$MOE = z^\star \times SE$$

$$\begin{aligned} 0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \quad \rightarrow \text{Use } \hat{p} \text{ from previous study} \\ 0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\ n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} \end{aligned}$$

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

$$MOE = z^\star \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \quad \rightarrow \textit{Use } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\
n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} \\
n &\geq 4898.04
\end{aligned}$$

Choosing a sample size

Q: *How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.*

$$MOE = z^\star \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} && \rightarrow \textit{Use } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\
n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} \\
n &\geq 4898.04 && \rightarrow \textit{n should be at least 4,899}
\end{aligned}$$

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Q: *why?*

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Q: *why?*

- *if you don't know any better, 50-50 is a good guess*

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Q: *why?*

- *if you don't know any better, 50-50 is a good guess*
- $\hat{p} = 0.5$ *gives the most conservative estimate – highest possible sample size*

Confidence Intervals (CI) vs. Hypothesis Tests (HT) for proportions

- Success-failure condition:
  - CI: At least 10 observed successes and failures
  - HT: At least 10 expected successes and failures, calculated using the null value
- Standard error:
  - CI: calculate using observed sample proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?*

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?*

$$H_0 : p = 0.80 \qquad H_A : p > 0.80$$

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?*

$$H_0 : p = 0.80 \qquad H_A : p > 0.80$$

$$SE = \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154$$

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?*

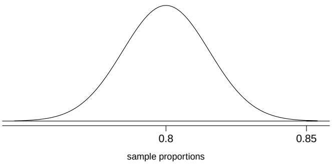
$$H_0 : p = 0.80 \qquad H_A : p > 0.80$$

$$\begin{aligned} SE &= \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154 \\ Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \end{aligned}$$

Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?*

$$H_0 : p = 0.80 \qquad H_A : p > 0.80$$

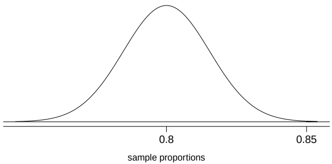
$$
\begin{aligned}
SE &= \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154 \\
Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \\
p - value &= 1 - 0.9994 = 0.0006
\end{aligned}
$$



Q: *The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?*

$$H_0 : p = 0.80 \qquad H_A : p > 0.80$$

$$\begin{aligned} SE &= \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154 \\ Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \\ p - value &= 1 - 0.9994 = 0.0006 \end{aligned}$$



Since the p-value is low, we reject H_0. The data provide convincing evidence that more than 80% of Americans have a good intuition on experimental design.

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

(a) Yes
(b) No
(c) Cannot tell

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

(a) Yes
(b) *No*
(c) Cannot tell

Section 6.2
Inference for two proportions

Melting ice cap

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

Parameter and point estimate

- *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.

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

Parameter and point estimate

- *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.

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

Inference for comparing proportions

- The details are the same as before...

Inference for comparing proportions

- The details are the same as before...
- CI: *point estimate* $\pm$ *margin of error*

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ to find appropriate p-value.

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ 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.

Inference for comparing proportions

- The details are the same as before...
- CI: *point estimate* $\pm$ *margin of error*
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ 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

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

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.

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.
   - $105 < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.

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.
   - $105 < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.

   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.

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.
   - $105 < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.

   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.

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.
   - $105 < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.

   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.

Q: *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

Q: *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

Q: *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

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

Q: *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

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

$$= \quad (0.657 - 0.668)$$

Q: *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

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

$$= \quad (0.657 - 0.668) \pm 1.96$$

Q: *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

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

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

Q: *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

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

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

$$= -0.011 \pm$$

Q: *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

$$
\begin{aligned}
& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497
\end{aligned}
$$

Q: *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

$$
\begin{aligned}
& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097
\end{aligned}
$$

Q: *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

$$
\begin{aligned}
& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097 \\
= \quad & (-0.108, 0.086)
\end{aligned}
$$

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

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.

Flashback to working with one proportion

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

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

Flashback to working with one proportion

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

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

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

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

Pooled estimate of a 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.

Pooled estimate of a 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.

Pooled estimate of a 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 estimate of a proportion

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

Q: *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*

Q: *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

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

Q: *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*

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

$$= \frac{69 + 454}{105 + 680}$$

Q: *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

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2} = \frac{69 + 454}{105 + 680} = \frac{523}{785}$$

Q: *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

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

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

Q: *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

Q: *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

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

Q: *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*

$$
\begin{aligned}
Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\
&= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} =
\end{aligned}
$$

Q: *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*

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

Q: *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*

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

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22$$

Q: *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*

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

Q: *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*

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

Recap - comparing two proportions

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

Recap - comparing two proportions

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

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 $\rightarrow$ randomization (Section 6.4)

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 $\rightarrow$ 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 $H_0 : p_1 - p_2 = \textit{(some value other than 0)}$: use $\hat{p}_1$ and $\hat{p}_2$
      - this is pretty rare

Reference - 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	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Reference - 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.

Reference - 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