Lecture 14 - Inference for proportions

Sta 102

June 7th, 2016

Colin Rundel & Mine Çetinkaya-Rundel

Inference

Testing in Context

	Independent Variable: None	Categorical (2 levels)	Categorical (>2 levels)	Numerical
Dependent Variable: Numerical	Test of One Mean	Test of Two Means	ANOVA	Regression
Categorical (2 levels)	Test of One Proportion	Test of Two Proportions	χ^2 - Test of Independence	Logistic Regression
Categorical (>2 levels)	χ^2 - GoF	χ^2 - Test of Independence	χ^2 - Test of Independence	~~Multinomial Regression~~

Inference for a single proportion

Example - Experimental Design

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 GSS 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

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

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

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

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

Inference on a proportion

What percent of all Americans have a 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 has the form

$$\textit{point estimate} \pm \textit{critical value} \times \textit{standard error}$$

Inference on a proportion

What percent of all Americans have a 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 has the form

$$\textit{point estimate} \pm \textit{critical value} \times \textit{standard error}$$

- What we need to know then is

$$SE_{\hat{p}} = ? \qquad CV = ?$$

Proportions and the CLT

What kind of probability model can we use for $\hat{p}$?

Proportions and the CLT

What kind of probability model can we use for $\hat{p}$?

It may be useful to instead think about $K = n\hat{p}$, what distribution will that have?

Proportions and the CLT

What kind of probability model can we use for $\hat{p}$?

It may be useful to instead think about $K = n\hat{p}$, what distribution will that have?

$$K \sim Binom(n, p)$$

Proportions and the CLT

What kind of probability model can we use for $\hat{p}$?

It may be useful to instead think about $K = n\hat{p}$, what distribution will that have?

$$K \sim Binom(n,\, p)$$

$$n\hat{p} \approx X \sim N\left(\mu = np,\, \sigma = \sqrt{np(1-p)}\right)$$

Proportions and the CLT

What kind of probability model can we use for $\hat{p}$?

It may be useful to instead think about $K = n\hat{p}$, what distribution will that have?

$$K \sim Binom(n,\ p)$$

$$n\hat{p} \approx X \sim N\left(\mu = np,\, \sigma = \sqrt{np(1-p)}\right)$$

We can then find the distribution of $\hat{p}$ by dividing by n,

$$\hat{p} \approx \frac{X}{n} \sim N\left(\mu = p,\, \sigma = \sqrt{\frac{p(1-p)}{n}}\right)$$

Central limit theorem (as applied to proportions)

A sample proportion will have a sampling distribution that is approximately normal with,

$$\hat{p} \sim N\left(\mu = p,\ \sigma = SE = \sqrt{\frac{p\,(1-p)}{n}}\right).$$

Central limit theorem (as applied to proportions)

A sample proportion will have a sampling distribution that is approximately normal with,

$$\hat{p} \sim N\left(\mu = p,\ \sigma = SE = \sqrt{\frac{p\,(1-p)}{n}}\right).$$

But of course this is true only under certain conditions ... any guesses?

Central limit theorem (as applied to proportions)

A sample proportion will have a sampling distribution that is approximately normal with,

$$\hat{p} \sim N\left(\mu = p,\ \sigma = SE = \sqrt{\frac{p\,(1-p)}{n}}\right).$$

But of course this is true only under certain conditions ... any guesses?

Assumptions/conditions:

1. *Independence*:
   - *Random sample*
   - *10% condition*: If sampling without replacement, $n < 10\%$ of the population.
2. *Normality*: At least 10 successes ($np \geq 10$) and 10 failures ($n(1-p) \geq 10$).

Back to experimental design...

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

Back to experimental design...

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

Given: $n = 670$, $\hat{p} = \frac{571}{670} = 0.85$.

Back to experimental design...

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

Given: $n = 670$, $\hat{p} = \frac{571}{670} = 0.85$.

Are CLT conditions met?

Back to experimental design...

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

Given: $n = 670$, $\hat{p} = \frac{571}{670} = 0.85$.

Are CLT conditions met?

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...

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

Given: $n = 670$, $\hat{p} = \frac{571}{670} = 0.85$.

Are CLT conditions met?

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.

Calculating the Confidence Interval

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$.
Calculate the 95% confidence interval for this proportion, and interpret it in context of the data.

$$CI = \text{point estimate} \pm \text{margin of error}$$

Calculating the Confidence Interval

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$.
Calculate the 95% confidence interval for this proportion, and interpret it in context of the data.

$$\begin{aligned} CI &= \text{point estimate} \pm \text{margin of error} \\ &= \text{point estimate} \pm \text{critical value} \times SE \end{aligned}$$

Calculating the Confidence Interval

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$.
Calculate the 95% confidence interval for this proportion, and interpret it in context of the data.

$$\begin{aligned} CI &= \text{point estimate} \pm \text{margin of error} \\ &= \text{point estimate} \pm \text{critical value} \times SE \\ &= \hat{p} \pm Z^{\star} \times SE \end{aligned}$$

Calculating the Confidence Interval

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Calculate the 95% confidence interval for this proportion, and interpret it in context of the data.

$$
\begin{aligned}
CI &= \text{point estimate} \pm \text{margin of error} \\
&= \text{point estimate} \pm \text{critical value} \times SE \\
&= \hat{p} \pm Z^\star \times SE \\
&= 0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}} = (0.82, 0.88)
\end{aligned}
$$

Choosing a sample size

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

Choosing a sample size

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

$$ME = Z^{\star} \times SE$$

Choosing a sample size

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

$$ME = Z^\star \times SE$$

$$0.01 \quad \geq \quad 1.96 \times \sqrt{\frac{p \times (1-p)}{n}}$$

Choosing a sample size

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

$$ME = Z^{\star} \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{p \times (1-p)}{n}} \\
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Using } \hat{p} \textit{ from previous study}
\end{aligned}$$

Choosing a sample size

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

$$ME = Z^{\star} \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{p \times (1-p)}{n}} \\
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Using } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n}
\end{aligned}$$

Choosing a sample size

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

$$ME = Z^{\star} \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{p \times (1-p)}{n}} \\
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Using } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\
n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2}
\end{aligned}$$

Choosing a sample size

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

$$ME = Z^{\star} \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{p \times (1-p)}{n}} \\
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Using } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\
n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} \\
n &\geq 4898.04
\end{aligned}$$

Choosing a sample size

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

$$ME = Z^\star \times SE$$

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{p \times (1-p)}{n}} \\
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Using } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\
n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} \\
n &\geq 4898.04 \rightarrow n \ \textit{ should be at least 4,899}
\end{aligned}$$

What if there isn't a previous study?

What if there isn't a previous study?

… use $\hat{p} = 0.5$. Why?

What if there isn't a previous study?

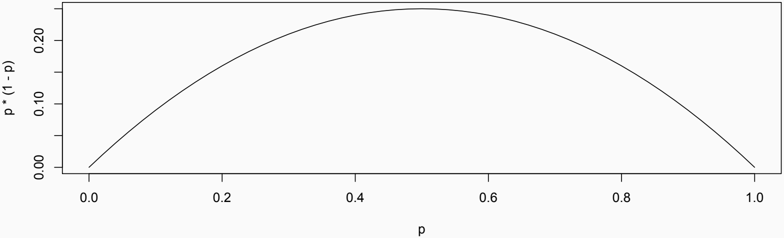
… use $\hat{p} = 0.5$. Why?

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

What if there isn't a previous study?

… use $\hat{p} = 0.5$. Why?

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



HT for proportions

Given what we know so far, how should we set up a hypothesis test for evaluating if more than 80% of all Americans have good intuition about experimental design?

HT for proportions

Given what we know so far, how should we set up a hypothesis test for evaluating if more than 80% of all Americans have good intuition about experimental design?

H_A is what we are interested in and H_0 represents the status quo, both *must* be about the population parameter of interest.

HT for proportions

Given what we know so far, how should we set up a hypothesis test for evaluating if more than 80% of all Americans have good intuition about experimental design?

H_A is what we are interested in and H_0 represents the status quo, both *must* be about the population parameter of interest.

Parameter of interest: p, point estimate: $\hat{p}$

HT for proportions

Given what we know so far, how should we set up a hypothesis test for evaluating if more than 80% of all Americans have good intuition about experimental design?

H_A is what we are interested in and H_0 represents the status quo, both *must* be about the population parameter of interest.

Parameter of interest: p, point estimate: $\hat{p}$

Hypotheses:

$$H_0 : p = 0.8$$
$$H_A : p > 0.8$$

CI vs. HT for proportions

For a test of one proportion our null and alternative hypotheses are about p, therefore when we assume H_0 is true we fix $p = p_0$. Hence,

- Standard error:
  - CI: calculate using observed sample proportion:

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - HT: calculate using the null value:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

CI vs. HT for proportions

For a test of one proportion our null and alternative hypotheses are about p, therefore when we assume H_0 is true we fix $p = p_0$. Hence,

- Standard error:
  - CI: calculate using observed sample proportion:

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - HT: calculate using the null value:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

- Success-failure condition:
  - CI: At least 10 *observed* successes and failures, calculated using the sample proportion, $\hat{p}$
  - HT: At least 10 *expected* successes and failures, calculated using the null value, p_0

Back to the GSS

The GSS found that 571 out of 670 (85%) of 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?

Back to the GSS

The GSS found that 571 out of 670 (85%) of 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?

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

Back to the GSS

The GSS found that 571 out of 670 (85%) of 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?

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

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

Back to the GSS

The GSS found that 571 out of 670 (85%) of 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?

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

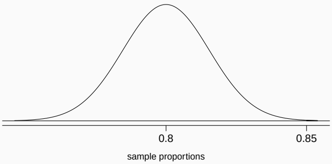
$$\begin{aligned} SE &= \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154 \\ Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \end{aligned}$$

Back to the GSS

The GSS found that 571 out of 670 (85%) of 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?

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

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

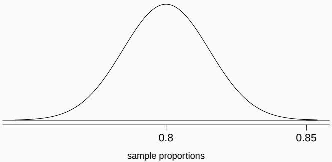


Back to the GSS

The GSS found that 571 out of 670 (85%) of 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 p-value is small we reject H_0.

Common Misinterpretations

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 a 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."

Is this statement justified?

Inference for difference of two proportion

Example - Melting ice cap survey

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 & Duke

The GSS asks this question, below is the distribution of responses from the 2010 survey:

A great deal	454
Not a great deal	226
Total	680

Results from the GSS & Duke

The GSS asks this question, below is the distribution of responses from the 2010 survey:

A great deal	454
Not a great deal	226
Total	680

The same question was asked of 88 Duke students, of which 56 said it would bother them a great deal.

We will collapse the data such that we consider only the responses of a great deal and its compliment, not a great deal.

Collapsed Results

	US	Duke	Total
A great deal	454	56	510
Not a great deal	226	32	258
Total	680	88	768

This is an example of a contingency table (specifically a 2 x 2 contingency table).

Collapsed Results

	US	Duke	Total
A great deal	454	56	510
Not a great deal	226	32	258
Total	680	88	768

This is an example of a contingency table (specifically a 2 x 2 contingency table).

We are interested in comparing proportion of Duke students who say it would both them a gread deal ($p_{GD|Duke} = 56/88$) to the proportion of all Americans who say it would both them a gread deal ($p_{GD|US} = 454/680$).

Condition on what?

Knowing which of the two variables to condition on can be tricky some times.

Ask yourself - which of the two variables is most likely the dependent variable (y) and which is most likely the independent variable (x). In other words, changes in x should *cause* changes in y (not the other way around).

Once we know this then the two proportions of interest are:

$$p_{y_1|x_1} \quad \text{and} \quad p_{y_1|x_2}$$

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 melting.

$$p_{GD|Duke} - p_{GD|US}$$

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 melting.

$$p_{GD|Duke} - p_{GD|US}$$

- *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 melting.

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

Inference for comparing proportions

The details for inference are the same as what we've seen previously,

Inference for comparing proportions

The details for inference are the same as what we've seen previously,

- CI: *point estimate* $\pm$ *critical value* $\times$ *std error*

Inference for comparing proportions

The details for inference are the same as what we've seen previously,

- CI: $\textit{point estimate} \pm \textit{critical value} \times \textit{std error}$
- HT: $\textit{Test Statistic} = \frac{\textit{point estimate} - \textit{null value}}{\textit{std error}}$, find appropriate p-value using sampling distribution.

Inference for comparing proportions

The details for inference are the same as what we've seen previously,

- CI: $\textit{point estimate} \pm \textit{critical value} \times \textit{std error}$
- HT: $\textit{Test Statistic} = \frac{\textit{point estimate} - \textit{null value}}{\textit{std error}}$, find appropriate p-value using sampling distribution.
- We just need to figure out the appropriate sampling distribution and its parameters..

Sampling Distribution

Last time we saw that the sampling distribution for $\hat{p}$ is a normal with mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$.

Sampling Distribution

Last time we saw that the sampling distribution for $\hat{p}$ is a normal with mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$.

We can combine that result with the approach we used for the test of two means to find the distribution of $\hat{p}_1 - \hat{p}_2$

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(\mu = E(\hat{p}_1 - \hat{p}_2),\ \sigma^2 = Var(\hat{p}_1 - \hat{p}_2)\right)$$

Sampling Distribution

Last time we saw that the sampling distribution for $\hat{p}$ is a normal with mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$.

We can combine that result with the approach we used for the test of two means to find the distribution of $\hat{p}_1 - \hat{p}_2$

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(\mu = E(\hat{p}_1 - \hat{p}_2),\ \sigma^2 = Var(\hat{p}_1 - \hat{p}_2)\right)$$

$$\begin{aligned} E(\hat{p}_1 - \hat{p}_2) &= E(\hat{p}_1) - E(\hat{p}_2) \\ &= p_1 - p_2 \end{aligned}$$

$$\begin{aligned} Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n} \end{aligned}$$

Sampling Distribution

Last time we saw that the sampling distribution for $\hat{p}$ is a normal with mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$.

We can combine that result with the approach we used for the test of two means to find the distribution of $\hat{p}_1 - \hat{p}_2$

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(\mu = E(\hat{p}_1 - \hat{p}_2),\ \sigma^2 = Var(\hat{p}_1 - \hat{p}_2)\right)$$

$$\begin{aligned} E(\hat{p}_1 - \hat{p}_2) &= E(\hat{p}_1) - E(\hat{p}_2) \\ &= p_1 - p_2 \end{aligned} \qquad \begin{aligned} Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n} \end{aligned}$$

Note - as with the test of two means, this result requires that $\hat{p}_1$ and $\hat{p}_2$ are independent.

Conditions for CI for the difference of two 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 the difference of two 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.

Conditions for CI for the difference of two 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.

   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 the difference of two 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.

   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 the difference of two 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.

   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 *both* groups.

CI for difference of proportions

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 $(p_{GD|Duke} - p_{GD|US})$.

	Duke	US
A great deal	56	454
Not a great deal	32	226
Total	88	680

CI for difference of proportions

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 $(p_{GD|Duke} - p_{GD|US})$.

	Duke	US
A great deal	56	454
Not a great deal	32	226
Total	88	680

$\hat{p}_{GD|Duke} = 56/88 = 0.636$
$\hat{p}_{GD|US} = 454/680 = 0.668$

CI for difference of proportions

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 $(p_{GD|Duke} - p_{GD|US})$.

	Duke	US
A great deal	56	454
Not a great deal	32	226
Total	88	680

$$\hat{p}_{GD|Duke} = 56/88 = 0.636$$
$$\hat{p}_{GD|US} = 454/680 = 0.668$$

$$SE \approx \sqrt{\frac{\hat{p}_{GD|Duke}(1 - \hat{p}_{GD|Duke})}{n_{Duke}} + \frac{\hat{p}_{GD|US}(1 - \hat{p}_{GD|US})}{n_{US}}}$$
$$= \sqrt{\frac{0.636(1 - 0.636)}{88} + \frac{0.668(1 - 0.668)}{680}} = 0.0537$$

CI for difference of proportions, cont.

$$\begin{aligned}\hat{p}_{GD|Duke} &= 0.636 \\ \hat{p}_{GD|US} &= 0.668 \\ SE &= 0.0537\end{aligned}$$

CI for difference of proportions, cont.

$$\begin{aligned} \hat{p}_{GD|Duke} &= 0.636 \\ \hat{p}_{GD|US} &= 0.668 \\ SE &= 0.0537 \end{aligned}$$

$$\begin{aligned} CI &= PE \pm CV \times SE \\ &= (\hat{p}_{GD|Duke} - \hat{p}_{GD|US}) \pm Z^{\star} \times \sqrt{\frac{\hat{p}_{GD|Duke}(1 - \hat{p}_{GD|Duke})}{n_{Duke}}} \\ &= (0.636 - 0.668) \pm 1.96 \times 0.0537 \\ &= (-0.138, 0.074) \end{aligned}$$

CI for difference of proportions, cont.

$$\begin{aligned} \hat{p}_{GD|Duke} &= 0.636 \\ \hat{p}_{GD|US} &= 0.668 \\ SE &= 0.0537 \end{aligned}$$

$$\begin{aligned} CI &= PE \pm CV \times SE \\ &= (\hat{p}_{GD|Duke} - \hat{p}_{GD|US}) \pm Z^{\star} \times \sqrt{\frac{\hat{p}_{GD|Duke}(1 - \hat{p}_{GD|Duke})}{n_{Duke}}} \\ &= (0.636 - 0.668) \pm 1.96 \times 0.0537 \\ &= (-0.138, 0.074) \end{aligned}$$

What conclusion should we draw here?

Hypotheses for testing the difference of two proportions

Just like the other hypothesis tests we have seen thus far, we formulate our null and alternative 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 as follows,

$$H_0 : p_{GD|Duke} = p_{GD|US} \quad \Rightarrow \quad p_{GD|Duke} - p_{GD|US} = 0$$
$$H_A : p_{GD|Duke} \neq p_{GD|US} \quad \Rightarrow \quad p_{GD|Duke} - p_{GD|US} \neq 0$$

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.

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

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.

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

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

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

A slight wrinkle ...

In setting the null hypothesis for comparing two proportions we haven't fixed either $p_{GD|Duke}$ or $p_{GD|US}$ - instead we have fixed their difference.

A slight wrinkle …

In setting the null hypothesis for comparing two proportions we haven't fixed either $p_{GD|Duke}$ or $p_{GD|US}$ - instead we have fixed their difference.

As such, we don't have a specific null value we can use to calculated the *expected* number of successes and failures in each group *or* the standard error. So, we know the following

A slight wrinkle ...

In setting the null hypothesis for comparing two proportions we haven't fixed either $p_{GD|Duke}$ or $p_{GD|US}$ - instead we have fixed their difference.

As such, we don't have a specific null value we can use to calculated the *expected* number of successes and failures in each group *or* the standard error. So, we know the following

$$
\begin{aligned}
p_{GD|Duke} &= p_{GD|US} \\
p_{GD|Duke} &= ? \\
p_{GD|US} &= ?
\end{aligned}
$$

A slight wrinkle ...

In setting the null hypothesis for comparing two proportions we haven't fixed either $p_{GD|Duke}$ or $p_{GD|US}$ - instead we have fixed their difference.

As such, we don't have a specific null value we can use to calculated the *expected* number of successes and failures in each group *or* the standard error. So, we know the following

$$
\begin{aligned}
p_{GD|Duke} &= p_{GD|US} \\
p_{GD|Duke} &= ? \\
p_{GD|US} &= ?
\end{aligned}
$$

Does this null give us any additional useful information?

Proportions and Probabilities

Think about the sample proportions as probabilities, what does it mean if

$$P(GD|Duke) = P(GD|US)$$

Proportions and Probabilities

Think about the sample proportions as probabilities, what does it mean if

$$P(GD|Duke) = P(GD|US)$$

If these two probabilities are equal then global warming concern is *independent* of the Duke vs. US grouping. Which means that,

$$P(GD|Duke) = P(GD|US) = P(GD)$$

Pooling

As such, our null hypothesis is equivalent to claiming that our two categorical variables are independent. So when conducting the hypothesis test we assume the null hypothesis to be true, which means we must also assume that the two variables are independent.

Under the assumption of independence our best guess for both $p_{GD|Duke}$ and $p_{GD|US}$ will be $\hat{p}_{GD}$, which is the sample proportion of *all* respondents (from Duke or US) who answered "A great deal".

We call this value $\hat{p}_{pooled}$,

$$\hat{p}_{pooled} = \frac{\#\text{ of successes in } 1 + \#\text{ of successes in } 2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Pooled estimate of a proportion

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.

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$$\hat{p}_{pooled} = \frac{56 + 454}{88 + 680} = \frac{510}{788} = 0.664$$

Pooled estimate of a proportion

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.

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$$\hat{p}_{pooled} = \frac{56 + 454}{88 + 680} = \frac{510}{788} = 0.664$$

Which sample proportion ($\hat{p}_{GD|Duke}$ or $\hat{p}_{GD|US}$) is closer to the pooled estimate? Why?

Implications for the SE

Under the null hypothesis we are stating that $p_1 = p_2$ which does not uniquely identify either p_1 or p_2. Therefore we are using the pooled proportion ($\hat{p}$) as our best guess for p_1 and p_2 under the null hypothesis.

For a *confidence interval* we use $\hat{p}_1$ and $\hat{p}_2$ to approximate for p_1 and p_2

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

While for a *hypothesis test* we use $\hat{p}_{pooled}$ to approximate for p_1 and p_2

$$\begin{aligned} SE &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{\hat{p}_p(1-\hat{p}_p)}{n_1} + \frac{\hat{p}_p(1-\hat{p}_p)}{n_2}} \\ &= \sqrt{\hat{p}_p(1-\hat{p}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \end{aligned}$$

HT for comparing 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?

$$\hat{p}_{pooled} = 0.664, \quad n_1 = 88, \quad n_2 = 680$$

HT for comparing 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?

$$\hat{p}_{pooled} = 0.664, \quad n_1 = 88, \quad n_2 = 680$$

$$SE = \sqrt{\hat{p}_p(1 - \hat{p}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.664(1 - 0.664)\left(\frac{1}{88} + \frac{1}{680}\right)} = 0.0535$$

HT for comparing 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?

$$\hat{p}_{pooled} = 0.664, \quad n_1 = 88, \quad n_2 = 680$$

$$SE = \sqrt{\hat{p}_p(1 - \hat{p}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.664(1 - 0.664)\left(\frac{1}{88} + \frac{1}{680}\right)} = 0.0535$$

$$Z = \frac{(56/88 - 454/680) - 0}{0.0535} = -0.59$$

HT for comparing 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?

$$\hat{p}_{pooled} = 0.664, \quad n_1 = 88, \quad n_2 = 680$$

$$SE = \sqrt{\hat{p}_p(1 - \hat{p}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.664(1 - 0.664)\left(\frac{1}{88} + \frac{1}{680}\right)} = 0.0535$$

$$Z = \frac{(56/88 - 454/680) - 0}{0.0535} = -0.59$$

$$\begin{aligned} \text{p-value} &= P(Z < -0.59 \text{ or } Z > 0.59) \\ &= 0.277 + 0.277 = 0.555 \end{aligned}$$

HT and CI agreement

Confidence interval:

$$CI = (-0.138, 0.074)$$

Hypothesis test:

$$H_0 : p_{GD|Duke} = p_{GD|US} \qquad Z = -0.59$$

$$H_A : p_{GD|Duke} \neq p_{GD|US} \qquad \text{p-value} = 0.555$$

Do the results of the Confidence interval and hypothesis test agree? Do the necessarily have to agree?

Picking successes?

What would happen to our analysis if we had picked "Not a great deal"?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

Picking successes?

What would happen to our analysis if we had picked "Not a great deal"?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$H_0 : p_{NGD|Duke} = p_{NGD|US}$

$H_0 : p_{NGD|Duke} \neq p_{NGD|US}$

$$\hat{p}_{pooled} = \frac{32 + 226}{88 + 680} = \frac{258}{788} = 0.336$$

Picking successes?

What would happen to our analysis if we had picked "Not a great deal"?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$H_0 : p_{NGD|Duke} = p_{NGD|US}$

$H_0 : p_{NGD|Duke} \neq p_{NGD|US}$

$$\hat{p}_{pooled} = \frac{32 + 226}{88 + 680} = \frac{258}{788} = 0.336$$

$$SE = \sqrt{0.336(1 - 0.336)\left(\frac{1}{88} + \frac{1}{680}\right)} = 0.0535$$

Picking successes?

What would happen to our analysis if we had picked "Not a great deal"?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$H_0 : p_{NGD|Duke} = p_{NGD|US}$

$H_0 : p_{NGD|Duke} \neq p_{NGD|US}$

$$\hat{p}_{pooled} = \frac{32 + 226}{88 + 680} = \frac{258}{788} = 0.336$$

$$SE = \sqrt{0.336(1 - 0.336)\left(\frac{1}{88} + \frac{1}{680}\right)} = 0.0535$$

$$Z = \frac{(32/88 - 226/680) - 0}{0.0535} = 0.585$$

$$\text{p-value} = P(Z < -0.59 \text{ or } Z > 0.59) = 0.277 + 0.277 = 0.555$$

Swapping dependent and independent variables?

What would happen to our analysis if we had swapped our independent and dependent variable?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

Swapping dependent and independent variables?

What would happen to our analysis if we had swapped our independent and dependent variable?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$$H_0 : p_{Duke|GD} = p_{Duke|NGD}$$

$$H_0 : p_{Duke|GD} \neq p_{Duke|NGD}$$

$$\hat{p}_{pooled} = \frac{56 + 32}{510 + 258} = \frac{88}{788} = 0.115$$

Swapping dependent and independent variables?

What would happen to our analysis if we had swapped our independent and dependent variable?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$H_0 : p_{Duke|GD} = p_{Duke|NGD}$

$H_0 : p_{Duke|GD} \neq p_{Duke|NGD}$

$$\hat{p}_{pooled} = \frac{56 + 32}{510 + 258} = \frac{88}{788} = 0.115$$

$$SE = \sqrt{0.115(1 - 0.115)\left(\frac{1}{510} + \frac{1}{258}\right)} = 0.0241$$

Swapping dependent and independent variables?

What would happen to our analysis if we had swapped our independent and dependent variable?

	Duke	US	Total
A great deal	56	454	510
Not a great deal	32	226	258
Total	88	680	788

$H_0 : p_{Duke|GD} = p_{Duke|NGD}$

$H_0 : p_{Duke|GD} \neq p_{Duke|NGD}$

$$\hat{p}_{pooled} = \frac{56+32}{510+258} = \frac{88}{788} = 0.115$$

$$SE = \sqrt{0.115(1-0.115)\left(\frac{1}{510}+\frac{1}{258}\right)} = 0.0241$$

$$Z = \frac{(56/510 - 32/258) - 0}{0.0241} = 0.59$$

$$\begin{aligned} \text{p-value} &= P(Z < -0.59 \text{ or } Z > 0.59) \\ &= 0.2775 + 0.2775 = 0.555 \end{aligned}$$

Recap

Recap - inference for one proportion

- Population parameter: p, point estimate: $\hat{p}$

Recap - inference for one proportion

- Population parameter: p, point estimate: $\hat{p}$
- Conditions:
  - independence
    - random sample and 10% condition
  - at least 10 successes and failures
    - *observed* for CI
    - *expected* for HT

Recap - inference for one proportion

- Population parameter: p, point estimate: $\hat{p}$
- Conditions:
  - independence
    - random sample and 10% condition
  - at least 10 successes and failures
    - *observed* for CI
    - *expected* for HT
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0
  - for Power:
    - Step 1 - use p_0
    - Step 2 - use $p_A = p_0 + \delta$

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
    - *observed* for CI
    - *expected* for HT

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
    - *observed* for CI
    - *expected* for HT
- $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - 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
  - for Power:
    - Step 1 - use $\hat{p}_{pool}$
    - Step 2 - use $\hat{p}_1$ and $\hat{p}_2$

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{\sigma}{\sqrt{n}}$	$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
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	$SE = \frac{\sigma}{\sqrt{n}}$	$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
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 as an approximation.

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{\sigma}{\sqrt{n}}$	$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
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 as an approximation.
- When working with proportions, we will not know p therefore
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead