Announcements

- Confidence Intervals Lab due Monday after break
- I'm not assigning other HW over break
- Extra Credit opportunity relating to March Madness posted on Campuswire
    - Initial component needs to be sent to me by Sunday, rest can be submitted later
- We'll have an in-class lab on the Monday we return after break
- I realize I'm behind in grading in here. I'm hoping to get caught up over break
- Polling: `rembold-class.ddns.net`

Remember this?

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

Confidence Interval v Hypothesis Testing for Proportions

- Success-failure conditions:
  - Confidence interval: At least 10 observed successes and failures
  - Hypothesis test: At least 10 expected successes and failures, calculated using the null value
- Standard error:
  - Confidence interval: Calculate using the observed samples proportion:

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

  - Hypothesis test: Calculate using the null value proportion:

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

Hypothesis Test Example

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

Hypothesis Test Example

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

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

Hypothesis Test Example

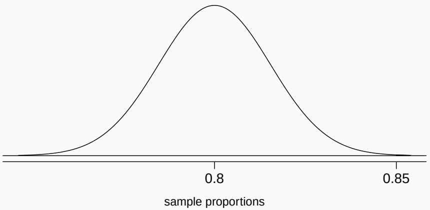
The GSS found that 571 out of 670 (85%) of Americans answer the question on the experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have good intuition about experimental design?

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

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

$$Z = \frac{0.85 - 0.80}{0.0154} = 3.25$$

$$\text{p-value} = 1 - 0.9994 = 0.0006$$



Hypothesis Test Example

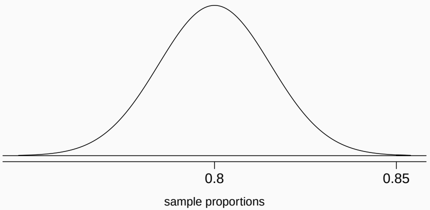
The GSS found that 571 out of 670 (85%) of Americans answer the question on the experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have good intuition about experimental design?

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

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

$$Z = \frac{0.85 - 0.80}{0.0154} = 3.25$$

$$\text{p-value} = 1 - 0.9994 = 0.0006$$



P-value is well below the 5% threshold, so we reject the null hypothesis in favor of the alternative hypothesis that more than 80% of Americans have good intuition about experimental design.

Understanding Check

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

Understanding Check

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

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
    - if not $\rightarrow$ randomization

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
    - if not $\rightarrow$ randomization
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0

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: $\textit{point estimate} \pm \textit{margin of error}$
- HT: Use $Z = \frac{\textit{point estimate} - \textit{null value}}{SE}$ 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.
   - $n_{Duke} < 10\%$ 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.
   - $n_{Duke} < 10\%$ 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 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.

   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 sampled 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.
   - $n_{Duke} < 10\%$ 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 sampled US residents are independent of each other.
3. ***Success-failure:***
   At least 10 observed successes and 10 observed failures in the two groups.

Example

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

Example

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

Example

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

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

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

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

$$= -0.011 \pm 0.097$$

$$= (-0.108, 0.086)$$

Understanding Check

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) $H_0 : p_{Duke} - p_{US} > 0$
$H_A : p_{Duke} - p_{US} \leq 0$

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

Understanding Check

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) $H_0 : p_{Duke} - p_{US} > 0$
$H_A : p_{Duke} - p_{US} \leq 0$

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

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.
- 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

$$\hat{p} = \frac{(\#\ of\ successes)_1 + (\#\ of\ successes)_2}{n_1 + n_2}$$

Example

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 $\hat{p}_{US}$) is 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

Example

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 $\hat{p}_{US}$) is 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

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

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

Hypothesis Example

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.

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

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

Hypothesis Example

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

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

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times0.334}{105} + \frac{0.666\times0.334}{680}}} = \frac{-0.011}{0.0495}$$

Hypothesis Example

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

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

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

$$= -0.22$$

Hypothesis Example

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

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

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times0.334}{105} + \frac{0.666\times0.334}{680}}} = \frac{-0.011}{0.0495}$$

$$= -0.22$$

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

Recap - comparing two proportions

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

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

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