

# Unit 3: Inference for Categorical and Numerical Data

## 2. Inference for the difference of two proportions (3.2)

2/25/2022

# Quiz 5 - Confidence Intervals

# Recap from last time

1. We can use the CLT to make inferences about proportions
2. Confidence intervals can be used to make inferences about a population proportion
3. Confidence intervals can be used to do hypothesis tests

# Key ideas

1. You can use the Normal approximation for the difference of two proportions
2. The margin of error is not just the sum of the margin of errors for each proportion
3. If you think two proportions come from the same population, you can use a pooled estimate

# Results from the NSF SEEI2012



National Science Board  
SCIENCE & ENGINEERING INDICATORS 2016

The National Science Foundation asked this question as part of a survey on general scientific literacy in 2010. Here are the results:

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

# Estimating the population parameter

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.

$p$ : a population proportion

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

$\hat{p}$ : a sample proportion

# Sample proportions are also nearly Normally Distributed

The Central Limit theorem for proportions says that the sample proportion will be nearly normal with mean equal to the population mean  $p$  and standard error  $\sqrt{\frac{p(1-p)}{n}}$

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

Only holds under the assumptions of the Central Limit Theorem:

- Independent samples
- N large enough ( $\sim 10$  success,  $\sim 10$  failures)

# Melting ice caps

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 it bother you if this actually happened?**

- (a) Yes
- (b) No

# Results from the NSF SEEI2012



## National Science Board SCIENCE & ENGINEERING INDICATORS 2016

The National Science Foundation asked this question as part of a survey on general scientific literacy in 2016. Here are the results:

	<b>SEEI2016</b>	<b>85309-21</b>	<b>85309-2</b>
Yes	578	15	10
No	104	1	1
Total	680	16	11

# Estimating the population difference

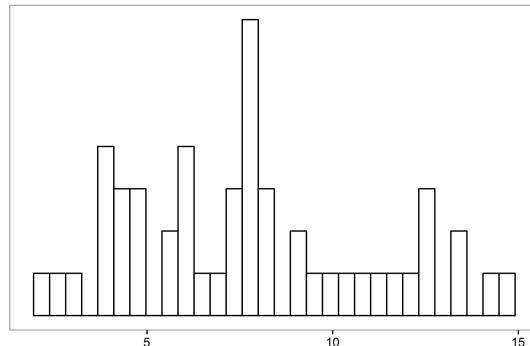
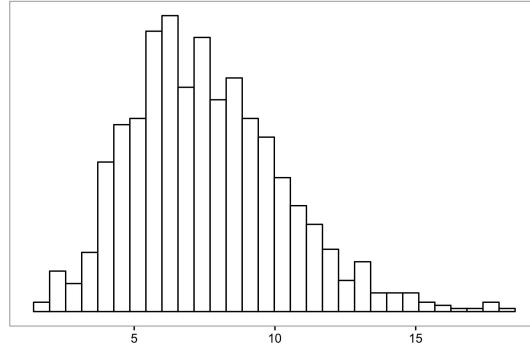
**Parameter of interest:** Difference between the proportions of all students and all Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{class} - p_{US}$$

**Point estimate:** Difference between the proportions of sampled students and sampled Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{class} - \hat{p}_{US}$$
: a sample proportion

# A reminder about the Central Limit Theorem

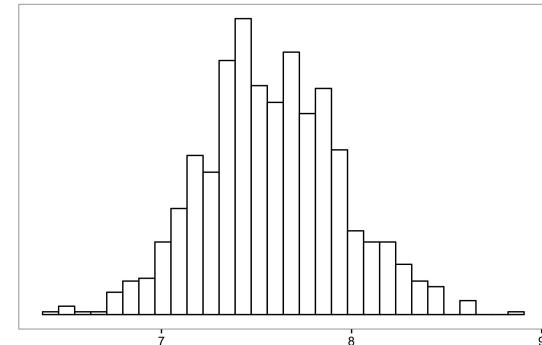


Take the mean,  
Repeat many times...



When I draw **independent samples** from the population, as sample size **approaches infinity**, the distribution of means approaches normality

Many statistical methods we use leverage this relationship  
(t-test, linear regression, ANOVA, etc)



# Inference for comparing proportions

Details almost the same as before...

*CI: point estimate  $\pm$  margin of error*

*HT: Use  $Z = \frac{\text{point estimate} - \text{null value}}{\text{Standard Error}}$*

We just need the appropriate standard error for the point estimate  
( $SE_{\text{class-US}}$ )

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

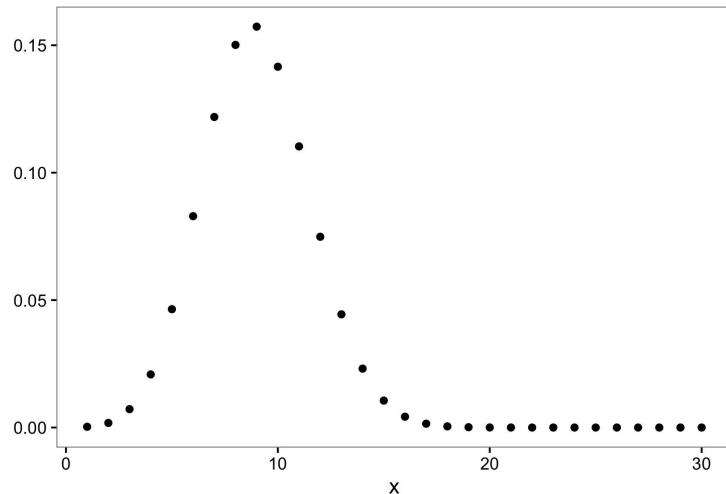
# Practice Question 1: Why the new SE estimate?

Naïve intuition: Find the SE for the class data, find the SE for the US data.  
Add them up

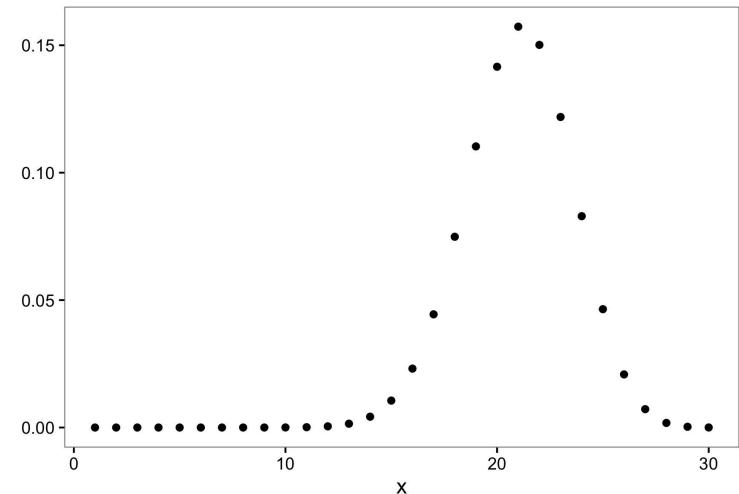
$$SE_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}} \quad SE_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

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

# Why is the correct SE estimate smaller?



$p=.3$



$p=.7$

# Conditions for CI for difference of proportions (Normal approx)

## Independence within groups

The people in the US group are sampled independently of each-other.

The people in the class group are sampled independently of each-other.

## Independence between groups

The sampled students and US residents are independent of each-other

## Success-failure

At least 10 observed successes and 10 observed failures in each group.

# Difference of proportions are also nearly-normally distributed

Construct a 95% confidence interval for the difference between the proportions of students and Americans who would be bothered by the melting of the northern ice cap ( $p_{\text{class}} - p_{\text{US}}$ ).

$$\hat{p}_1 - \hat{p}_2 \pm Z^* \times \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$
$$.938 - .85 \pm 1.96 \times \sqrt{\frac{.938 \times .062}{16} + \frac{.85 \times .15}{680}}$$
$$.089 \pm .12$$
$$(-.031, .209)$$

## Practice Question 2

Which of the following is the correct set of hypotheses for testing if the proportion of 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) $H_0: p_{\text{class}} = p_{\text{US}}$<br>$H_A: p_{\text{class}} \neq p_{\text{US}}$                                 | (c) $H_0: p_{\text{class}} - p_{\text{US}} = 0$<br>$H_A: p_{\text{class}} - p_{\text{US}} \neq 0$ |
| (b) $H_0: \hat{p}_{\text{class}} - \hat{p}_{\text{US}} = 0$<br>$H_A: \hat{p}_{\text{class}} - \hat{p}_{\text{US}} \neq 0$ | (d) $H_0: p_{\text{class}} = p_{\text{US}}$<br>$H_A: p_{\text{class}} < p_{\text{US}}$            |

## Practice Question 2

Which of the following is the correct set of hypotheses for testing if the proportion of 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)  $H_0: p_{\text{class}} = p_{\text{US}}$

$$H_A: p_{\text{class}} \neq p_{\text{US}}$$

(c)  $H_0: p_{\text{class}} - p_{\text{US}} = 0$

$$H_A: p_{\text{class}} - p_{\text{US}} \neq 0$$

(b)  $H_0: \hat{p}_{\text{class}} - \hat{p}_{\text{US}} = 0$

$$H_A: \hat{p}_{\text{class}} - \hat{p}_{\text{US}} \neq 0$$

(d)  $H_0: p_{\text{class}} = p_{\text{US}}$

$$H_A: p_{\text{class}} < p_{\text{US}}$$

# A pooled estimate of the population proportion

If you think that two samples come from the same population ( $p$ ). Or you want to test whether they do, you used a *pooled estimate* of  $\hat{p}$ .

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

$$\hat{p}_1 - \hat{p}_2 \sim N \left( 0, \sqrt{\frac{\hat{p}_{pool}(1 - \hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1 - \hat{p}_{pool})}{n_2}} \right)$$

# Testing $H_0$ : No difference between the samples

$$\hat{p}_{pool} = \frac{578 + 15}{680 + 16} = .852$$

$$SE_{pool} = \sqrt{\frac{.852 \times .148}{680} + \frac{.852 \times .148}{16}}$$
$$= .09$$

$$.938 - .85 \sim N(0, .09)$$

$$.088 \sim N(0, .09)$$

`pnorm(.088, mean=0, sd=.09) = .84`

# Key ideas

1. You can use the Normal approximation for the difference of two proportions
2. The margin of error is not just the sum of the margin of errors for each proportion
3. If you think two proportions come from the same population, you can use a pooled estimate