Unit 3: Inference for Categorical and Numerical Data

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

10/30/2017

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

Questions Statisticians Answer...

	Means	**Proportions**
	When we measure a **number** for each individual (e.g. height, weight, SAT score). We use the **mean** to summarize the data for a group. *-Example: The average weight of a sample of 100 granny smith apples from my orchard was 72 grams.*	When we measure which of **one of two options** is true for each individual (e.g. organ donor or not?, lived or died? red or blue?) we use the **proportion** to summarise the data for a group. ***-**Example: The proportion of people at this fertility clinic who have live births using ART is 0.4 (i.e. 40%).*
One Sample	Is the mean in my group different from some value I care about? *-Example: Are the apples in my orchard heavier than typical granny smith apples (which have a known average weight of 70 grams)?* *-Null: Mean in my orchard = 70g* *-Alternative: Mean in my orchard > 70g*	Is the proportion in my group different from some value I care about? *-Example: Is the probability of a having a live birth at my clinic higher than the known US probability of a live birth using ART (which is 0.3)?* *-Null: Probability in my clinic = 0.3* *-Alternative: Probability in my clinic > 0.3*
Two Samples	Are the means of two groups different? *-Example: Does my new drug lower cholesterol more than a placebo?* *-Null: Mean cholesterol of drug takers = Mean cholesterol of placebo takers* *-Alternative: Mean cholesterol of drug takers < Mean cholesterol of placebo takers*	Are the proportions in two groups different? *-Example: Is Curry's probability of making a shot higher for "Hot Shots" than for "Not Shots"* *-Null: Probability for hot shots = Probability for not shots* *-Alternative: Probability for hot shots > Probability for not shots*

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 (~10 success, ~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) Not

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:

	SEEI2012	**PSYC 201-17**	**PSYC 201-18**
Yes	578	43	
No	104	0	
Total	680	43	

Estimating the population difference

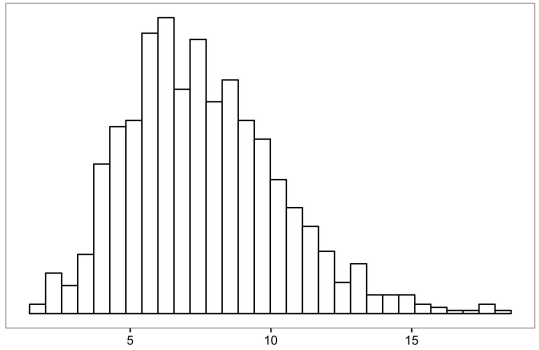
Parameter of interest: Difference between the proportions of all students and all Americans who would be bothered a great deal by the northern ice cap completely melting.

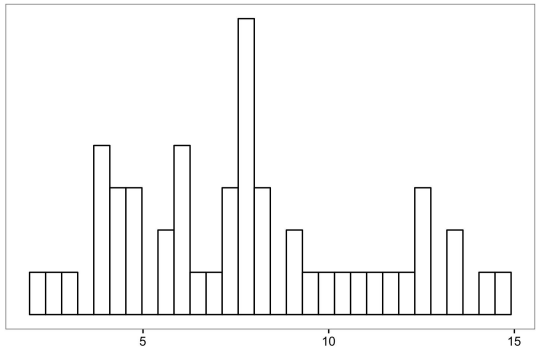
$p_{class} - p_{US}$

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

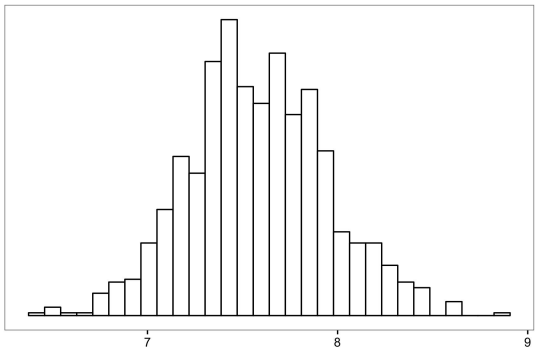
A reminder about the Central Limit Theoreom





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 ± margin of error

HT: Use $Z = \dfrac{point\ estimate - null\ vaue}{Standard\ Error}$

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

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

Practice Question 1: Why the new SE estimate?

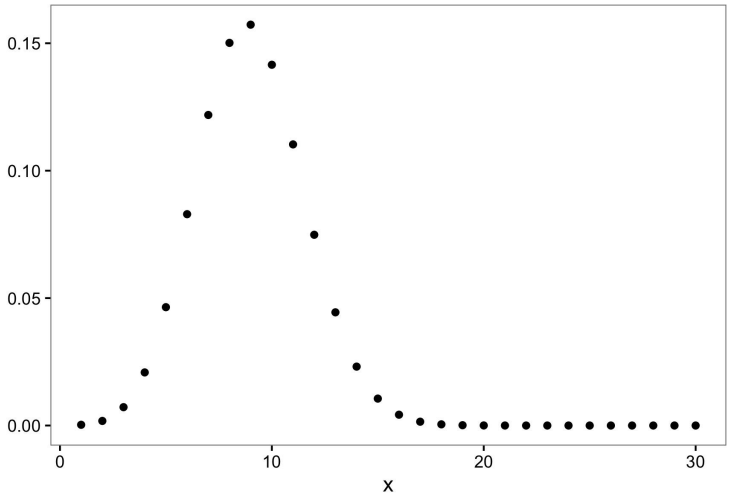
Naïve intuition: Find the SE for the class data, find the SE for the US data.

Add them up

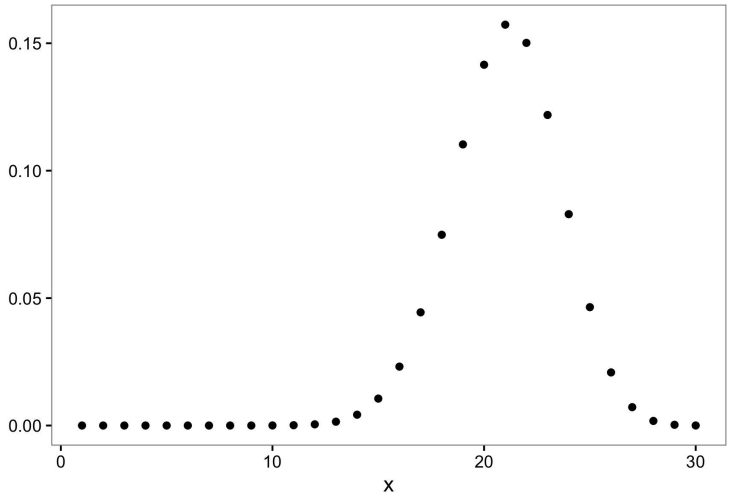
$$SE_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}} \qquad SE_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

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

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 a great deal by the melting of the northern ice cap ($p_{class} - p_{US}$).

$$(\hat{p}_1 - \hat{p}_2) \pm Z^* \times \sqrt{\frac{\hat{p}_{class}(1-\hat{p}_{class})}{n_{class}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

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

$$= -0.011 \pm 0.097$$

$$= (-0.108, 0.086)$$

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) $H_0: p_{class} = p_{US}$
$H_A: p_{class} \neq p_{US}$

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

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

(d) $H_0: p_{class} = p_{US}$
$H_A: p_{class} < p_{US}$

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) $\mathbf{H_0: p_{class} = p_{US}}$
$\mathbf{H_A: p_{class} \neq p_{US}}$

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

(c) $\mathbf{H_0: p_{class} - p_{US} = 0}$
$\mathbf{H_A: p_{class} - p_{US} \neq 0}$

(d) $H_0: p_{class} = p_{US}$
$H_A: p_{class} < p_{US}$

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 $\hat{p}$.

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

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

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