Inference for Comparing Two Proportions

Coming up…

- Lab Assignment 7 is due **Thursday just before your lab section time.**
- Problem Set 5 due Wednesday 3/27

What's new in Unit 5?

45 secs.

		Analyses and Ways to Conduct Them (that we know so far)	
Types of Variable(s) Involved	**Population Parameter**	**Confidence Interval for the Population Parameter**	**Hypothesis Test for the Population Parameter** *Ho: Pop. Param = # Ha: Pop. Param (≠ or < or >) #*
Single Numerical Variable	μ	CLT Confidence Interval (Unit 3+4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 3+4) Bootstrap Hypothesis Test (Unit 4)
	μdiff	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 4) Bootstrap Hypothesis Test (Unit 4)
	Median	Bootstrap Confidence Interval (Unit 4)	Bootstrap Hypothesis Test (Unit 4)
Single Categorical Variable (2 levels)	**p**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 5) Bootstrap Hypothesis Test (Unit 4) Randomization Testing (Unit 5-Selecting ball/chips out of bag, rolling dice)
Numerical Response Variable Categorical Explanatory Variable (2 levels)	μ1-μ2	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 4) Randomization Testing (Unit 1-Shuffling Cards)
	Median1-Median2	Bootstrap Confidence Interval (Unit 5)	Randomization Testing (Unit 1-Shuffling Cards)
Categorical Response Variable Categorical Explanatory Variable (both have 2 levels)	**p1-p2**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 5) Randomization Testing (Unit 1-Shuffling Cards)

What's new in Unit 5?

Types of Variables	Analysis	Hypotheses
Numerical Response Variable Categorical Explanatory Variable (>2 levels)	**ANOVA**	Ho: μ1=μ2=…=μk Ha: at least one pair of groups has means that are different
Single Categorical Variable (>2 levels)	**Chi-Squared Goodness of Fit Test**	Ho: The data follows the specified distribution. Ha: The data does not follow the specified distribution.
Categorical Response Variable Categorical Explanatory Variable (at least one has >2 levels)	**Chi-Squared Independence Test**	Ho: The two variables are independent/not associated. Ha: The two variables are dependent/associated.

Would we use the same analysis for the following research questions?

- "Is there an association between voting in 2018 and gender?"
- "Is there an association between political affiliation and gender?"

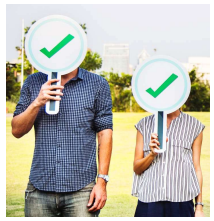
Sample Data (Registered Voters)

Gender *(Categorical with 2 Levels)*	Political Affiliation *(Categorical with 3 Levels)*	Voted in 2018? *(Categorical with 2 Levels)*
Male	Democrat	Yes
Female	Republican	Yes
Male	Independent	Yes
Male	Republican	Yes
Female	Republican	No
Female	Democrat	No
Female	Democrat	Yes
Male	Independent	Yes
...	...	...

"Is there an association between voting in 2018 and gender?"

*Categorical explanatory variable **with 2 levels***

*Categorical response variable **with 2 levels***



Comparing 2 Proportions Hypothesis Test
***provided necessary conditions are met**

$Ho: p_{male} - p_{female} = 0$ → No association

$Ha: p_{male} - p_{female} \neq 0$ → Association

p_{male} = *proportion of all registered US males that voted in* 2018
p_{female} = *proportion of all registered US females that voted in* 2018

"Is there an association between political affiliation and gender?"

*Categorical explanatory variable **with 2 levels***

*Categorical response variable **with >2 levels***

Chi-Squared Independence Test
***provided necessary conditions are met**

Ho: *political affiliation and gender are independent (not associated)*
Ha: *political affiliation and gender are dependent (associated)*

Would we use the same analysis for the following research questions?

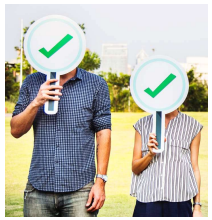
- "Are the proportions of men and women currently registered to vote equal?"
- "Are the proportions of republicans, democrats, and independents currently registered to vote equal?"

Sample Data (Registered Voters)

Gender *(Categorical with 2 Levels)*	Political Affiliation *(Categorical with 3 Levels)*
Male	Democrat
Female	Republican
Male	Independent
Male	Republican
Female	Republican
Female	Democrat
Female	Democrat
Male	Independent
...	...

"Are the proportions of men and women currently registered to vote equal?"

Categorical variable
with 2 levels



p= *proportion of all registered US citizens that are male*

Single Proportion Hypothesis Test
*provided necessary conditions are met

$Ho: p = 0.5$
$Ha: p \neq 0.5$

"Are the proportions of republicans, democrats, and independents currently registered to vote equal?"

Categorical variable
with >2 levels

Chi-Squared Goodness of Fit Test
*provided necessary conditions are met

Ho: *the registered voter data follows this distribution*
Ha: *the registered voter data doesn't follow this distribution*

Event	Probability
Registered Citizen is Republican	1/3
Registered Citizen is Democrat	1/3
Registered Citizen is Independent	1/3

Figuring out when an "observation" follows a normal distribution when you're not explicitly told.

If $X \sim Bin(n,p)$ and when certain conditions are met...
$X \sim N\left(mean = np, standard\ dev. = \sqrt{np(1-p)}\right)$

When other certain conditions are met...
$\bar{x} \sim N(mean = \mu, standard\ dev./error = \frac{\sigma}{\sqrt{n}})$

When other certain conditions are met...
$\bar{x}_1 - \bar{x}_2 \sim N(mean = \mu_1 - \mu_2, standard\ dev./error = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

When other certain conditions are met...
$\hat{p} \sim N(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

When other certain conditions are met...
$\hat{p}_1 - \hat{p}_2 \sim N(mean = p_1 - p_2, standard\ dev./error = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$

What are these conditions for each type of "observation"?

If $X \sim Bin(n,p)$ and when certain conditions are met...
$X \sim \boldsymbol{N}\left(mean = np, standard\ dev. = \sqrt{np(1-p)}\right)$

When other certain conditions are met...
$\bar{x} \sim \boldsymbol{N}(mean = \mu, standard\ dev./error = \frac{\sigma}{\sqrt{n}})$

When other certain conditions are met...
$\bar{x}_1 - \bar{x}_2 \sim \boldsymbol{N}(mean = \mu_1 - \mu_2, standard\ dev./error = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

When other certain conditions are met...
$\hat{p} \sim \boldsymbol{N}(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

When other certain conditions are met...
$\hat{p}_1 - \hat{p}_2 \sim \boldsymbol{N}(mean = p_1 - p_2, standard\ dev./error = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$

Why is this important?

Use z-tables

For instance…

$P\ (\hat{p} < obs) = P(Z < \frac{obs - mean}{stand.dev})$=probability

When other certain conditions are met…

$\hat{p} \sim N(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$