GROUP EXAM 2 REVIEW

Needed for class....

- Download the *Exam2ReviewData.csv* file and ***10-Exam2Review_OneQuantitativeVariable_Inference.R*** from Canvas
  - *Modules -> Module 10 -> Files and links for activities... -> Group Exam 2 Review*
- Login to the RStudio server and upload both files

- Read through the study on pg. 246

Exam Information

Sunday, October 19th 4 - 5:30 pm in Gaines Hall 101

- Exam 2 Review Study Session

Monday, October 20th

- Group Exam 2 – during normal class time

Wednesday, October 22nd

- Individual Exam 2 – during normal class time

Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Coming up….	10/20 • Group Midterm Exam 2 in class	10/21	10/22 • Individual Midterm Exam 2 in class	10/23	10/24 • Read Sect 5.6 and Chapters 19 and 20 in the ***online textbook before class*** • Watch 5.6, 19.1, 19.2, 19.3TheoryTests, and 19.4TheoryIntervals videos and complete the video notes in the ***coursepack*** *(pgs. 252 – 260)* • *Optional Module 11 Video Quiz (Canvas)* • Complete Activity 17 (pgs. 261 - 266) in class	10/25
10/26	10/27 • Complete Activity 18 (pgs. 267 – 271) in class	10/28	10/29 • Complete Module 11 Lab (pgs. 272 – 276) in class and on Gradescope	10/30	10/31 • Read Chapter 6, 7, 8, 20, and 21 in the ***online textbook before class*** • Watch 6.1, 6.2, 6.3, Ch7, Ch21_Overview, 21.1, 21.3, 21.4TheoryTests, 21.4TheoryIntevals videos and complete the video notes in the ***coursepack*** *(pgs. 281 - 298)* • *Optional Module 12 Video Quiz (Canvas)* • Complete Activity 19 (pgs. 299 - 303) in class	11/1

Group Exam 2 Review

- Group exam covers simulation-based methods for a single quantitative variable (Module 6)
- *Write out the parameter of interest for the study on the whiteboard for instructors to check!*
- *Write out the null and alternative hypotheses in notation for the study on the whiteboard for instructors to check!*
- *Need to create a plot of the data (in R file) for part f*
- Complete the remaining questions using the provided R file

Exploratory Data Analysis for a Single Quantitative Variable (Module 6)

- Pages 113 - 135 in the Coursepack
  - *Review interpretations of Q_1, Q_3, Median, standard deviation*
  - *How do we identify outliers?*
  - *Review description of distribution of quantitative variables*
    - Shape, center, spread, outiers
- Chapters 5 in the online textbook (https://mtstateintrostats.github.io/IntroStatTextbook/index.html)
- Assignment 5

Simulation Methods for analyzing One Quantitative Variable (Module 6)

- Pages 113 – 127, 136 – 142 in the Coursepack
- Chapter 17 in the online textbook
- Assignment 5
- Golden Ticket (pg. 388)

Theory-based Methods for analyzing One Quantitative Variable (Module 7)

- Pages 143 – 157 in the Coursepack
- Chapter 17 in the online textbook
- Assignment 5
- Golden Ticket (pg. 388)

Review of Simulation Methods vs Theoretical Methods for a Single Quantitative Variable

	Simulation	Theory
Distribution	Create the simulation null distribution: Shift the data by adding $\mu_0 - \bar{x}$ to each value. Sample with replacement n times. Plot sample mean on the null distribution.	T-distribution with n – 1 degrees of freedom
Finding the p-value	Count the proportion of simulated samples at $\bar{x}$ and more extreme. The smaller the p-value the stronger the evidence against the null hypothesis.	Calculate the Standardized statistic $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ Find the area under the t-distribution with n- 1 df at the standardized statistic and more extreme.
Confidence Interval	Create the bootstrap distribution and use the percentile method to find the 90%, 95% or 99% CI.	$\bar{x} \pm t^* \times s/\sqrt{n}$ t* multiplier is found from the t-distribution with n – 1 df at a given percentile for the confidence level

Assumptions

Simulation Methods

- Independent observations

Theory-based Methods

- Independent observations
- Large enough sample size (Normality)
  - $n < 30$ – distribution of sample must be approximately symmetric with no outliers
  - $30 \leq n < 100$ – distribution must not have extreme outliers
  - $n \geq 100$ – sample size is large enough regardless of the shape of the sample

Standardized Statistic

$$t = \frac{\bar{x} - \mu_0}{SE(\bar{x})}$$

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

The standardized statistic measures the number of standard errors the statistic is from the null value.

Confidence Interval

$$\bar{x} \pm t^* \times SE(\bar{x})$$

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

The confidence interval is an interval estimate for the parameter of interest.

	R Code	Where to find example?

Boxplot:
• Plot title includes Type of plot, observational units, variable
	`object %>% # Data set piped into...`
	`  ggplot(aes(x = variable)) + # Name variable to plot`
	`  geom_boxplot() + # Create boxplot`
	`  labs(title = "Don't forget to title the plot!", # Title for plot`
	`  x = "x-axis label", # Label for x axis`
	`  y = "y-axis label") # Label for y axis`
		pg. 121, pg, 131, pg. 137, pg. 153

Null Distribution	`one_mean_test(object$variable,#Enter the object name and variable`
	`  null_value = xx, #Enter the null value for the study`
	`  summary_measure = "mean", #Can choose between mean or median`
	`  shift = xx, #Difference between the null value and the sample mean`
	`  as_extreme_as = xx, #Value of the summary statistic`
	`  direction = "xx", #Specify direction of alternative hypothesis`
	`  number_repetitions = 10000)`
		Pg. 124, pg. 139

Bootstrap Distribution	`one_mean_CI(object$variable, #Enter the name of the variable`
	`  summary_measure = "mean", #choose the mean or median`
	`  number_repetitions = 10000, # Number of simulations`
	`  confidence_level = xx)`
		Pg. 127, pg. 140

Theory p-value	– `pt` will give you a p-value using the t-distribution with a given degrees of freedom (enter for `yy`). For a single mean, `df` $= n - 1$. – Enter the value of the standardized statistic for `xx` – If a "greater than" alternative, change `lower.tail = TRUE` to `FALSE`. – If a two-sided test, multiply by 2. `pt(xx, df = yy, lower.tail=TRUE)`	pg. 147, pg. 155
t* multiplier for CI	– `qt` will give you the multiplier using a t-distribution with a given degrees of freedom (enter for `yy`). For a single mean, `df` $= n - 1$. – Enter the percentile for the given level of confidence (e.g., 0.975 for a 95% confidence level). `qt(percentile, df = yy, lower.tail=FALSE)`	pg. 148, pg. 156

WRITING GUIDES	What it includes	Where to find example?

Parameter of Interest	• Population word (true, long-run, population)
	• Summary measure (depends on the type of data)
	• Context
	  • Observational units
	  • Variable (success)
		pg. 137, 231

P-value interpretation	• Statement about probability/proportion of samples • Statistic • Direction of alternative • Null hypothesis (in context)	pg. 140, 232
Confidence interval interpretation	• How confident you are (90%, 95%, 99%..) • Parameter of interest • Calculated interval • Order of subtraction	Pg. 141, 156, 233
Conclusion	• Amount of evidence • Parameter of interest • Direction of alternative hypothesis	Pg. 140, 155, 233
Standardized Statistic	• Measures the number of standard errors the sample statistic is above (or below) the null value	pg. 147, 235

Study Design (Module 8)

- Pages 168 – 196 in the Coursepack
- Chapter 2 in the online textbook
- Assignment 6

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

	Study Type	
Selection of cases	Randomized Experiment	Observational Study
By random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

Exploratory Data Analysis for two categorical variables (Module 8)

- Pages 168 – 189, 196 – 201 in the Coursepack
- Chapter 4 in the online textbook
- Assignment 5
- Golden Ticket (pg. 388)

Simulation Methods for analyzing Two Categorical Variables (Module 8)

- Pages 168 – 189, 202 - 211 in the Coursepack
- Chapter 15 in the online textbook
- Assignment 6, 7
- Golden Ticket (pg. 388)

Theory-based Methods for analyzing Two Categorical Variable (Module 9)

- Pages 210 – 223 in the Coursepack
- Chapter 15 in the online textbook
- Golden Ticket (pg. 388)

Review of Simulation Methods vs Theoretical Methods for Testing

	Simulation	Theory
Distribution	Create the simulation null distribution: Label cards with the total number of successes and failures. Mix together, shuffle into two piles; n_1 for the total number in group 1 and n_2 for the total number in group 2. Plot the difference in proportion (group 1 – group 2) of success on null distribution.	Standard Normal Distribution
Finding the p-value	Count the proportion of simulated samples at $\widehat{p_1} - \widehat{p_2}$ and more extreme. The smaller the p-value the stronger the evidence	Calculate the Standardized statistic $$z = \frac{\widehat{p_1} - \widehat{p_2} - 0}{\sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ Find the probability under the standard normal distribution at the standardized statistic and more extreme
Confidence Interval	Create the bootstrap distribution and use the percentile method to find the 90%, 95% or 99% CI.	$$\widehat{p_1} - \widehat{p_2} \pm z^* \times \sqrt{\frac{\widehat{p_1} \times (1 - \widehat{p_1})}{n_1} + \frac{\widehat{p_2} \times (1 - \widehat{p_2})}{n_2}}$$

Assumptions

Simulation Methods

- Independent observations

Theory-based Methods

- Independent observations
- Large enough sample size (Success/Failure Condition)
    - More than 10 successes and more than 10 failures in each group

Standardized Statistic for Two Categorical Variables

$$z = \frac{statistic\ - null\ value}{SE_0(statistic)}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

Where, $SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ And $\hat{p}_{pool} = \frac{total\ successes}{n_1 + n_2}$

*Remember that the standardized statistic measures the number of standard errors the statistic is from the null value.

Confidence Intervals

$$\textbf{statistic} \pm \textbf{multiplier} \times (\textbf{standard error of statistic})$$

Margin of Error

To estimate a population difference in proportion, $\pi_1 - \pi_2$:

- **Two-sample z-interval** (*theory-based*):

$$\hat{p}_1 - \hat{p}_2 \pm (\text{multiplier}) \times \sqrt{\frac{\hat{p}_1 \times (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1-\hat{p}_2)}{n_2}}$$

$SE(\hat{p}_1 - \hat{p}_2)$

Note: The z* multiplier is the value found at a specific percentile of the standard normal distribution to match the confidence level.

	R Code	Where to find example?

Segmented Bar plot:
• Plot title includes Type of plot, observational units, variable
	`object %>%`
	`  ggplot(aes(x = explanatory, fill = response))+ #Enter the variables to plot`
	`  geom_bar(stat = "count", position = "fill") + #Creates a segmented bar plot`
	`  labs(title = "Don't forget to title a plot!", #Make sure to title your plot`
	`  y = "y-axis label", #y-axis label`
	`  x = "x-axis label")  #x-axis label`
		pg. 178, 198, 212,

Null Distribution	`two_proportion_test(formula = response~explanatory, # response ~ explanatory`
	`  data = object, # Name of data set`
	`  first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1`
	`  number_repetitions = 10000, # Always use a minimum of 10000 repetitions`
	`  response_value_numerator = "xx", # Define which outcome is a success`
	`  as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)`
	`  direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")`
		Pg. 184, 206

Bootstrap Distribution	`two_proportion_bootstrap_CI(formula = response~explanatory,`
	`  data=object, # Name of data set`
	`  first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1`
	`  response_value_numerator = "xx", # Define which outcome is a success`
	`  number_repetitions = 10000, # Always use a minimum of 10000 repetitions`
	`  confidence_level = xx) # Enter the level of confidence as a decimal`
		Pg. 186, pg. 208

Theory p-value	– Enter the value of the standardized statistic for xx. – If a "greater than" alternative, change `lower.tail = TRUE` to `FALSE`. – If a two-sided test, multiply by 2. `pnorm(xx, lower.tail=TRUE)`	pg. 215, 220
z* multiplier for CI	– `qnorm` will give you the multiplier using the standard normal distribution. – Enter the percentile for the given level of confidence (e.g., 0.975 for a 95% confidence level). `qnorm(percentile, lower.tail=TRUE)`	pg. 216, 221

WRITING GUIDES	What it includes	Where to find example?

Parameter of Interest	• Population word (true, long-run, population)
	• Summary measure (depends on the type of data)
	• Context
	  • Observational units
	  • Variable (success)
		pg. 203, 240

P-value interpretation	• Statement about probability/proportion of samples • Statistic • Direction of alternative • Null hypothesis (in context)	pg. 207, 220, 242
Confidence interval interpretation	• How confident you are (90%, 95%, 99%..) • Parameter of interest • Calculated interval • Order of subtraction	Pg. 208, 221, 243
Conclusion	• Amount of evidence • Parameter of interest • Direction of alternative hypothesis	Pg. 208, 221, 242
Standardized Statistic	• Measures the number of standard errors the sample statistic is above (or below) the null value	pg. 220, 244