MA256 Lesson 11 - Two Groups, Two Means (6.1-6.3)

Review for Single Mean:

Hypotheses (in symbols):

H_0 :

H_a :

Strength of Evidence:

Confidence Interval:

Two means:

Hypotheses (in symbols):

H_0 :

H_a :

Strength of Evidence:

Confidence Interval:

Validity conditions:

Q1) In a survey of introductory statistics students, an instructor asked her students to report how many Facebook friends they had. Suppose that the intent is to study whether there is an association between number of Facebook friends and a person's sex.

a. Identify whether the study is an experiment or an observational study. Explain.

b. Identify the observational units.

c. Identify the explanatory and response variables. Also, for each variable identify whether it is categorical or quantitative.

d. What is an appropriate null and alternative hypothesis?

e. Below are a few summary statistics about the data. In this context, calculate (an) appropriate statistic value(s) to compare men to women?

	Sample size	Sample mean	Sample Median	Sample SD
Women	35	594.30	532	309.80
Men	13	405.60	485	228.40

Q2) Suppose that randomly sampled college students are asked how many hours they typically spend online each day. You conduct a two-sided test of the null hypothesis that $\mu_{females} - \mu_{males} = 0$, and you also calculate a 95% confidence interval for $\mu_{females} - \mu_{males}$.

a). Describe (in words) what the parameter $\mu_{females} - \mu_{males}$ means here.

b). Now suppose that your friend analyzes the same data but with the order of subtraction reversed (males - females, rather than females - males). Describe the impact (if any) on each of the following. In other words, describe how your friend's findings will compare to yours with regard to each of the following.

i. Distribution of simulated statistics under the null hypothesis.

ii. Standard deviation of the simulated statistics under the null hypothesis.

iii. Observed value of the statistic (difference in sample means).

iv. Approximate p-value from simulation.

v. Value of t-test statistic.

vi. p-value from t-test.

vii. Midpoint of confidence interval.

viii. Endpoints of confidence interval.

ix. Width of confidence interval.

c). What is the bottom line: Will you and your friend reach the same conclusions even though you disagreed about the order in which to perform the subtraction (males - females or females - males)? Explain.

IOCT Tabbing

How fair are the "tabbing" IOCT times? Currently to tab the IOCT a male must have a time faster than 2:38 (158 seconds) and a female must have a time faster than 3:35 (215 seconds). This is a difference of 57 seconds. Is this a fair difference? How much do male and female's times differ on the IOCT? You think that the difference in average male and female times is not 57 seconds. The data we have is the IOCT times for all cadets who took MA206 in AY21-1 and have a valid IOCT time.

```
# library(tidyverse)
# ioct.data <- read.csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/IOCT_tab_data.csv",
#                       stringsAsFactors = TRUE)
```

1. Identify the explanatory and response variables recorded and classify them as either categorical or quantitative.

2. In words, state the null and the alternative hypotheses to t est whether male or female's IOCT times differ from 57 seconds.

3. Define the parameters of interest and assign symbols.

4. State the null and the alternative hypotheses in symbols.

5. Calculate the five-number summary of IOCT time by group, calculate the IQR for each group, and create a graphical representation of the five-number summary in R Studio using example code from the course guide.

```
# ioct.data %>%
#   group_by(sex) %>%
#   summarize(Minimum = XXXX,
#             LowerQuartile = XXXX,
#             Median = XXXX,
#             UpperQuartile = XXXX,
#             Maximum = XXXX)
#
#   #IQR-Females
#   # IQR-Males
#
# ioct.data %>%
#   ggplot(......
#   labs(x = "IOCT Time (Seconds)", y = "Biological Sex",
#        title = "IOCT Time (Seconds) by Biological Sex")
```

6. Do the validity conditions appear to be satisfied for these data? Justify your answer.

```
# Put some useful plots here to help you decide.
```

7. Conduct the theory-based two-sample t-test:

(a) What is the standardized statistic?

```
# ioct.data %>%
#   group_by(sex) %>%
#   summarize(avgtime=XXXXX,
#             sdtime=XXXXX,
#             size=XXXXX)
#
# # Calculate the Standardized Statistic
# xbar_M <- XXXXX
# xbar_F <- XXXXX
# s_M <- XXXXX
# s_F <- XXXXX
# n_M <- XXXXX
# n_F <- XXXXX
# sd <- XXXXX
# null <- XXXXX
# statistic <- XXXXX
# t <- XXXXX
#
# ## using built-in t-test (this is not covered in the course guide... see if you can figure it out)
# Fioct <- XXXXX
# Mioct <- XXXXX
#
# t.test(Fioct, Mioct, mu = 57, var.equal = FALSE)
```

(b) In light of your standardized statistic, should you expect the p-value to be large or small? How are you deciding?

(c) What is your p-value? Provide the value (number) and an explanation in context of the problem.

```
# pval <- XXXXX
# pval
```

(d) Based on the p-value, evaluate the strength of evidence provided by the data against the null hypothesis. Do you reject or fail to reject your null hypothesis.

8. Determine the 95% confidence interval for the difference in means of male and female IOCT Times.

(a) What is the 95% confidence interval?

```
# multiplier <- XXXXX
# sd <- XXXXX
# c(statistic - multiplier * sd, statistic + multiplier * sd)
```

(b) Does the 95% confidence interval agree with your conclusion in #7?

(c) Interpret the 95% confidence interval.

9. Operating under the null hypothesis mentioned in # 3 above, create a simulation to simulate males and females taking the IOCT. Estimate the p-value. Use 1000 replications and plot your results.

```
# set.seed(256)
#
# M <- 1000
# xbar <- XXXXX
# mu0 <- XXXXX
# sexs <- ioct.data$sex
# iocts <- ioct.data$IOCT_Time
#
# RES <- data.frame(res = rep(NA, M))
#
# for(rep in 1:M){
#   sex.shuf <- sample(XXXXX)
#   m.mean <- mean(iocts[sex.shuf == "M"])
#   f.mean <- mean(iocts[sex.shuf == "F"])
#   RES$res[rep] <- f.mean - m.mean
# }
#
# RES %>% ggplot(aes(x=res)) +
#   geom_histogram() +
#   geom_vline(xintercept = 82-57, color="red") +
#   geom_vline(xintercept = -(82-57), color="red")
#
# # estimate p-value
# sum(RES$res >= mu0) / M
```