MA256 Lesson 12 - Two Groups, Two Means (6.1-6.3)

Review for Single Mean:

Hypotheses (in symbols):

$H_0 : \mu = \mu_0$

$H_a : \mu \neq \mu_0$

Strength of Evidence: Calculate t statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

reject H_0 for t "more extreme" using the guidlines for appropriate significance level.

Confidence Interval: $\bar{x} \pm t^*_{(1-\alpha/2,n-1)} \times s/\sqrt{n}$

reject H_0 if μ_0 is NOT in CI.

Two means:

Hypotheses (in symbols):

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

Strength of Evidence: Calculate t statistic: $t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

Confidence Interval: $\bar{x}_1 - \bar{x}_2 \pm t^*_{(1-\alpha/2,n-1)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Validity conditions: Symmetric distribution in both groups OR $\geq$ 20 observations in each group and sample distribtuion not strongly skewed.

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.

Observational study; there is no randomization of treatment.

b. Identify the observational units.

Each student

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

Explanatory: gender (categorical); response: number of Facebook friends (quantitative)

d. What is an appropriate null and alternative hypothesis?

H_0: There is no association between a person's sex and the number of Facebook friends.
H_a: There is an association between a person's sex and the number of Facebook friends.

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

Means: 594.30 - 405.60 = 188.70. Medians: 532 - 485 = 47

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.

The difference in the population mean number of hours spent online daily, in particular how many more hours are spent by females then males

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. no change

ii. Standard deviation of the simulated statistics under the null hypothesis. no change

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

iv. Approximate p-value from simulation. no change

v. Value of t-test statistic. Will change in sign

vi. p-value from t-test. no change

vii. Midpoint of confidence interval. Will change in sign

viii. Endpoints of confidence interval. both will change in sign

ix. Width of confidence interval. no change

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.

Yes, the exact same conclusion will be reached about strength of evidence against the null hypothesis. The CI will be different because it is estimating a different parameter but will mean the same thing in terms of whether males or females are, on average, spending more hours online each day.

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.

IOCT Times: Quantitative Response Variable
Sex: Categorical Explanatory Variable

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

Null Hypothesis: The difference in the long run average IOCT completion times for males and females is 57 sec.
Alt Hypothesis: The difference in the long run average IOCT completion times for males and females is not 57 sec.

3. Define the parameters of interest and assign symbols.

μ_F: The long-run average IOCT Completion Times for Females
μ_M: The long-run average IOCT Completion Times for Males

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

$H_0 : \mu_F - \mu_M = 57$ seconds
$H_a : \mu_F - \mu_M \neq 57$ seconds

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 = min(IOCT_Time),
            LowerQuartile = quantile(prob =.25, IOCT_Time),
            Median = median(IOCT_Time),
            UpperQuartile = quantile(prob=.75, IOCT_Time),
            Maximum = max(IOCT_Time))
```

```
## # A tibble: 2 x 6
##   sex   Minimum LowerQuartile Median UpperQuartile Maximum
##   <fct>   <int>         <dbl>  <dbl>         <dbl>   <int>
## 1 F         184          224.    251          288.     629
## 2 M         139          166     178          191      335
```

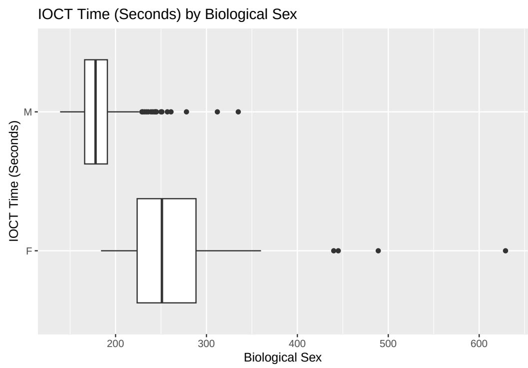
```
288.5-223.8 #IQR-Females
```

```
## [1] 64.7
```

```r
191-166 # IQR-Males
```

```
## [1] 25
```

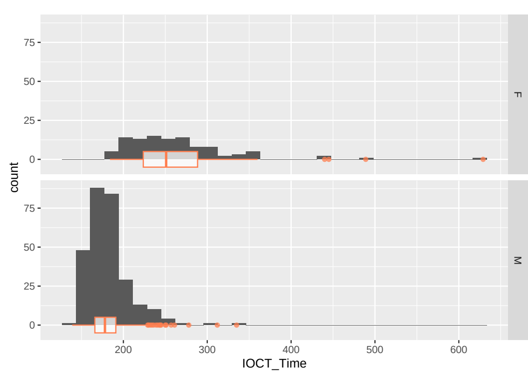
```r
ioct.data %>%
  ggplot(aes(x = IOCT_Time, y = as.factor(sex))) + geom_boxplot()+
  labs(x = "Biological Sex", y = "IOCT Time (Seconds)",
       title = "IOCT Time (Seconds) by Biological Sex")
```



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

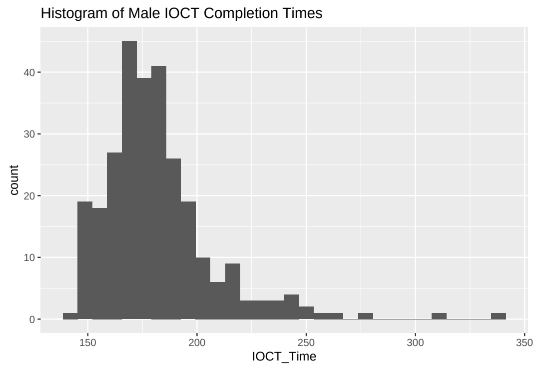
```r
ioct.data %>%
  ggplot(aes(x=IOCT_Time)) + geom_histogram() + geom_boxplot(width = 10, color = "coral", alpha = 0.75) +
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



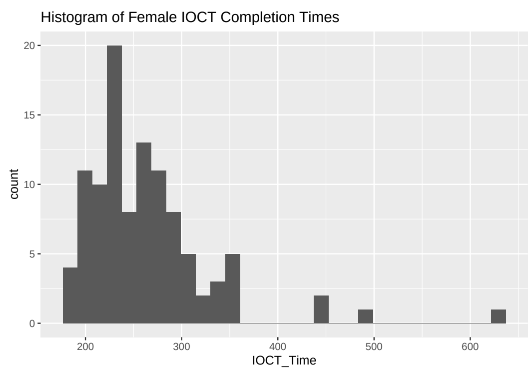

```r
#Male Histogram
ioct.data %>%
  filter(sex=="M") %>% ggplot(aes(x=IOCT_Time)) + geom_histogram() +
  labs(title="Histogram of Male IOCT Completion Times")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Female Histogram
ioct.data %>%
  filter(sex=="F")%>%
  ggplot(aes(x=IOCT_Time))+geom_histogram()+
  labs(title="Histogram of Female IOCT Completion Times")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Yes. There are 104 females and 280 males (at least 20 observations each) in the sample and neither distribution for IOCT Times is strongly skewed.

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

(a) What is the standardized statistic?

```
ioct.data %>%
  group_by(sex) %>%
  summarize(avgtime=mean(IOCT_Time),
            sdtime=sd(IOCT_Time),
            size=n())
```

```
## # A tibble: 2 x 4
##   sex   avgtime sdtime  size
##   <fct>   <dbl>  <dbl> <int>
## 1 F        264.   66.1   104
## 2 M        182.   25.8   280
```

```
# Calculate the Standardized Statistic
xbar_M <- 181.8750
xbar_F <- 264.1346
s_M <- 25.84629
```

```r
s_F <- 66.12272
n_M <- 280
n_F <- 104
sd <- sqrt(s_M^2/n_M+s_F^2/n_F)
null <- 57
statistic <- xbar_F-xbar_M
t <- (statistic-null)/sd
c(statistic, t)
```

```
## [1] 82.259600  3.789712
```

```r
## using built-in t-test
Fioct <- ioct.data %>% filter(sex=="F") %>% select(IOCT_Time)
Mioct <- ioct.data %>% filter(sex=="M") %>% select(IOCT_Time)

t.test(Fioct, Mioct, mu = 57, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Fioct and Mioct
## t = 3.7897, df = 114.89, p-value = 0.0002418
## alternative hypothesis: true difference in means is not equal to 57
## 95 percent confidence interval:
##  69.05678 95.46245
## sample estimates:
## mean of x mean of y
##  264.1346  181.8750
```

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

With a t statistic of 3.7523, I have very strong evidence against the null hypothesis so I would expect a very small p-value (< 0.01).

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

```r
pval <- 2*(1-pt(abs(t), (n_F+n_M) - 2))
pval
```

```
## [1] 0.0001751845
```

The p-value of 0.000175 is the probability of observing a long run differene in average IOCT times of 82 seconds assuming the null hypothesis is true.

(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.

Although possible, it is very unlikely that the observed difference in average IOCT completion times for males and females occurred by random chance due to very strong evidence against the null hypothesis . I would reject the null in favor of there being a difference in average IOCT completion times for males and females that is not equal to 57 seconds.

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

(a) What is the 95% confidence interval?

```r
multiplier <- qt(0.975, (n_F+n_M) - 2)
sd <- sqrt(s_M^2 / n_M + s_F^2 / n_F)
c(statistic - multiplier * sd, statistic + multiplier * sd)
```

```
## [1] 69.15431 95.36489
```

(b) Does the 95% confidence interval agree with your conclusion in #5?

Yes. Since the null hypothesis value for the difference in average IOCT times of females and males (57 seconds) is not in my 95% confidence interval range of plausible values, 57 seconds is not plausible and I reject the Null hypothesis. Since my confidence interval shows the long run average female IOCT time is between 69.15 and 95.36 seconds larger than the average male IOCT time, I would revise my conclusion to reflect the appropriate direction rather than say "not equal to."

(c) Interpret the 95% confidence interval.

We are 95% confident that the true long run average male IOCT time is less than the true long run average female IOCT time by between 69.15 and 95.36 seconds.

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.

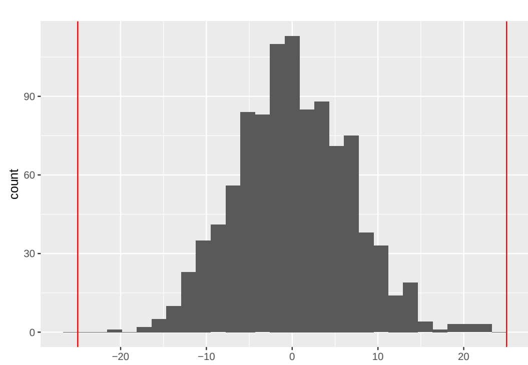
```r
set.seed(256)

M <- 1000
xbar <- 82
mu0 <- 57
sexs <- ioct.data$sex
iocts <- ioct.data$IOCT_Time

RES <- data.frame(res = rep(NA, M))

for(rep in 1:M){
  sex.shuf <- sample(sexs)
  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")
```



```r
# estimate p-value
sum(RES$res >= mu0) / M
```

```
## [1] 0
```