MA206, Lesson 16 - Two Groups - Two Means

Review: What types of variables do we compare when testing two proportions?
What types of variables do we compare when testing two means?

What is the parameter of interest when comparing two means?

What is the statistic of interest used to infer about the parameter?

What is the null hypothesis if we assume no association between two groups?

What makes up the Five-Number Summary used to build boxplots?

What makes up the Inter-Quartile Range?

How do we define outliers using the Five-Number Summary and Boxplots?

What are the validity conditions to use theoretical methods for two groups, two means?

How do we calculate SD and SE for two groups, two means?

How do we compute the standardized statistic for two means?
t =

How do we compute the p-value for the standardized t statistic?

How do we calculate the confidence interval for the difference in two means?

1) Many students pull "all-nighters" when they have an important exam or a pressing assignment. Concerns that may arise include *Can you really function well the next day after a sleepless night? What about several days later? Can you recover from a sleepless night by getting a full night's sleep on the following nights?* Researchers Stickgold, James, and Hobson investigated the delayed effects of sleep deprivation in a study published in 2000 in *Nature Neuroscience*. They had 21 volunteers, aged 18-25 years old, who were trained on a visual discrimination task that involved watching stimuli appear on a computer screen and reported what was seen afterwards. After their training period, they were tested.

The volunteers were randomly separated into two groups. The control group (10 individuals) were given no limitations on their sleep for three days before being tested. The test group (11 individuals) were deprived of sleep for 30 hours, followed by two nights of unrestricted sleep before being tested. After the third night, both groups were retested on the task and assessed on their *improvement*, in milliseconds, for the response task. That is, if someone got better, they had a positive value, and if they got worse, they had a negative value. If the performance was the same, the score would be 0. The goal of the study was to see if the improvement scores would be higher for those without sleep deprivation than for the sleep deprived group. The dataset is available on Teams (sleep.csv).

a) Identify and classify the explanatory and response variables in this study.

b) Was this an experiment or an observational study?

c) In words and symbols, state the null and alternate hypotheses to investigate whether sleep deprivation has a negative effect on the improvement on performance in visual discrimination tasks.

d) Create a histogram of the results as a whole. Create a second histogram of the results broken up by group. Compare the two graphs and comment on the results.

e) Create a 5 Number Summary Table for both groups.

	Minimum	**Lower Quartile**	**Median**	**Upper Quartile**	**Maximum**
Deprived					
Not Deprived					

```
See Course Guide for code.

Sleep %>%
 group_by(Group)%>%
 summarize(Minimum = min(improvement),
           LowerQuartile = quantile(prob =.25, improvement),
           Median = median(improvement),
           UpperQuartile = quantile(prob=.75, improvement),
           Maximum = max(improvement))
```

f) Calculate the Inter-Quartile Range for both groups.

g) Using the IQR, calculate the cutoffs to classify an observation as an outlier for each group.

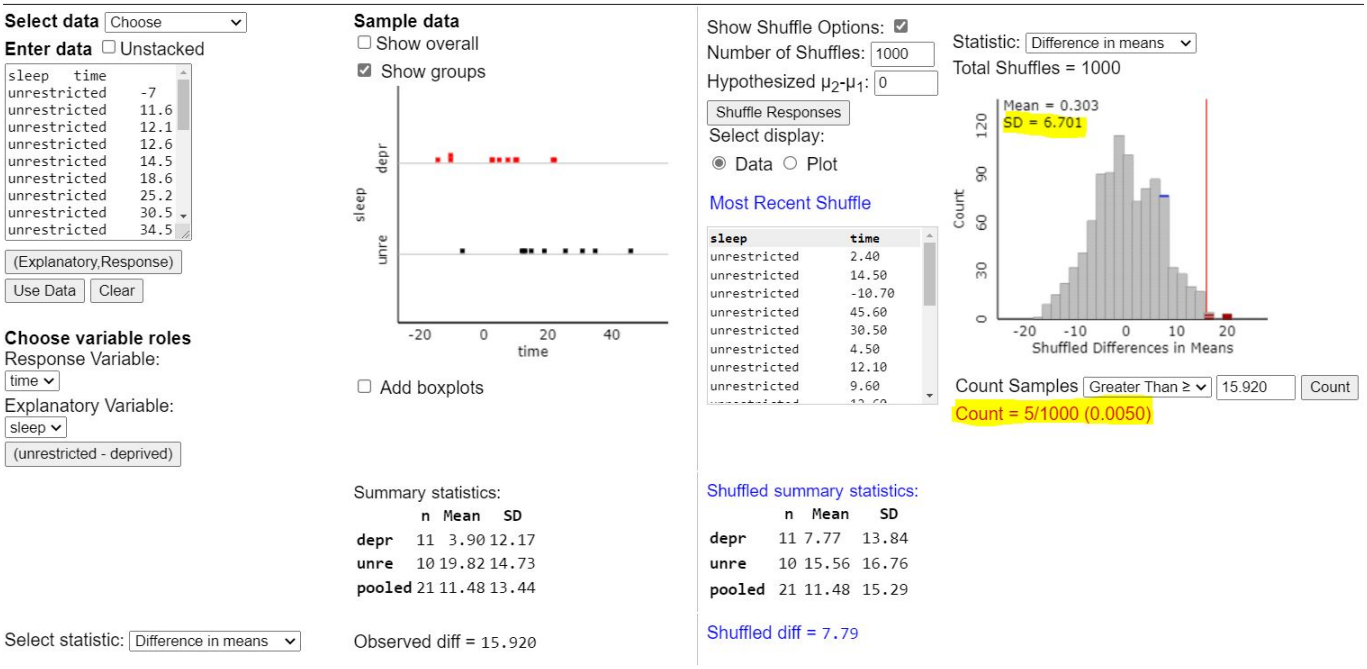
h) Calculate the observed statistic (The observed difference in means between the two groups).

```
See Course Guide

Sleep %>%
 group_by(Group) %>%
 summarise(xbar = mean(improvement),
           s = sd(improvement),
           n = n())
```

i) Do we meet validity conditions to use theoretical methods?

j) Using the simulation results in the figure below, report the SD and calculate the standardized statistic.




```
Sample.Stat <- 19.8-3.9
simsd <- 6.701
Sample.Stat/simsd
```

j) Using the applet results, report your p-value and interpret what it means.

k) Usingthe 2SD method and the standard deviation found in **j** above, what is your estimated 95% confidence interval? Interpret what this means.

l) Can you generalize these results? Can you infer a cause-and-effect relationship with these results? Explain.

2) You and your roommate want to solve one of the biggest debates of all time: Which comic movies are better, DC or Marvel? To solve it, you collected data from *https://www.the-numbers.com/movies/keywords/DC-Comics* and pulled movies dating from 1978 through 2021, removing any incomplete information, and wanted to compare the "Worldwide Box Office" proceeds with the "Production Budget" costs to calculate total profits. You want to see if there is a difference between Marvel and DC movie profits. The data can be found in the Movies.csv file in Teams.

a) What is your null and alternate hypothesis?

b) Update your data set, using the TidyVerse tutorial as a guide to generate a new variable, "**Profit**", by calculating Worldwide Box Office - Production Budget.

```
Movies2 <- Movies %>%
  mutate(Profit = Worldwide.Box.Office - Production.Budget)
```

c) Calculate your 5 Number Summary for the comic profits as a whole and the accompanying Histogram with Boxplot.

```
Movies2 %>%
 ggplot(aes(x=Profit))+
 geom_histogram(color="black", fill="gray")+
 geom_boxplot(color="blue", fill="lightblue", lwd=2)+
 theme_classic()+
 labs(title="Histogram", x="Profits", y="Count")+
 scale_x_continuous(labels=scales::dollar_format())
```

```
Movies2 %>%
 summarize(Minimum = min(Profit),
           LowerQuartile = quantile(prob =.25, Profit),
           Median = median(Profit),
           UpperQuartile = quantile(prob=.75, Profit),
           Maximum = max(Profit))
```

	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Profits					

d) Generate a Split Histogram with accompanying boxplots to compare DC and Marvel profits. What are your initial thoughts?

```
Movies2 %>%
 ggplot(aes(x=Profit))+
 geom_histogram(color="black", fill="gray")+
 geom_boxplot(color="blue", fill="lightblue", lwd=2)+
 facet_grid(Brand ~.)+
 theme_classic()+
 labs(title="Histogram", x="Profits", y="Count")+
 scale_x_continuous(labels=scales::dollar_format())
```

e) Complete the 5-Number Summary Table for both groups.

	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
D.C.					
Marvel					

f) What is our observed statistic (Difference in means)?

g) Do we meet validity conditions?

```
Movies2 %>%
 group_by(Brand) %>%
 summarise(xbar = mean(Profit),
           s = sd(Profit),
           n = n())

Movies2 %>%
 filter(Brand=="DC") %>%
 ggplot(aes(x=Profit))+
 geom_histogram()

Movies2 %>%
 filter(Brand=="Marvel") %>%
 ggplot(aes(x=Profit))+
 geom_histogram()
```

h) Using theoretical methods, calculate the standardized statistic and p-value in accordance with the research question. Interpret your results in terms of the problem.

```
xbar1 <- 245511367
xbar2 <- 460595147
s1 <- 281608950
s2 <- 436568409
n1 <- 29
n2 <- 60

sd <- sqrt(s1^2/n1 + s2^2/n2)
t <- (xbar1 - xbar2)/sd
pvalue <- 2*(1 - pt(abs(t), n1+n2-2))
```

i) Calculate a 95% confidence interval for the true difference in profits between Marvel and DC movies.

```
se <- sd
m <- qt(1 - 0.05/2, n1+n2-2)
CI <- c(xbar1 - xbar2 - m*se, xbar1-xbar2 + m*se); CI
```