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?
One categorical variable and one quantitative variable

What is the parameter of interest when comparing two means?
The long-run or true difference in means; The true difference between two means from two groups.
$\mu_1 - \mu_2$

What is the statistic of interest used to infer about the parameter?
The difference between the two observed means for each of the two groups.
$\bar{x}_1 - \bar{x}_2$

What is the null hypothesis if we assume no association between two groups?
$H_0 : \mu_1 - \mu_2 = 0$

What makes up the Five-Number Summary used to build boxplots?
Minimum, Lower Quartile, Median, Upper Quartile, Maximum

What makes up the Inter-Quartile Range?
The range of values between the Lower Quartile and the Upper Quartile

How do we define outliers using the Five-Number Summary and Boxplots?
Outliers are those observations more than 1.5 × the IQR away from the Lower Quartile or Upper Quartile.

What are the validity conditions to use theoretical methods for two groups, two means?
More than 20 observations in both groups and neither group is strongly skewed.

How do we calculate SD and SE for two groups, two means?
$\text{SD} = \text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

How do we compute the standardized statistic for two means?
$\text{t} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

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

For a less than alternative hypothesis test, pt(t, n-2)
For a greater than alternative hypothesis test, 1 - pt(t, n-2)
For a not equal to alternative hypothesis test, 2 * (1 - pt(abs(t), n-2))

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

$(\bar{x}_1 - \bar{x}_2) \pm qt(1 - \frac{\alpha}{2}, n - 2) \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Where the multiplier M is $M = qt(1 - \frac{\alpha}{2}, n - 2)$

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.

The Explanatory Variable is Group (deprived vs. not deprived) and it is categorical.
The Response Variable is Improvement (in ms) and it is quantitative.

b) Was this an experiment or an observational study?

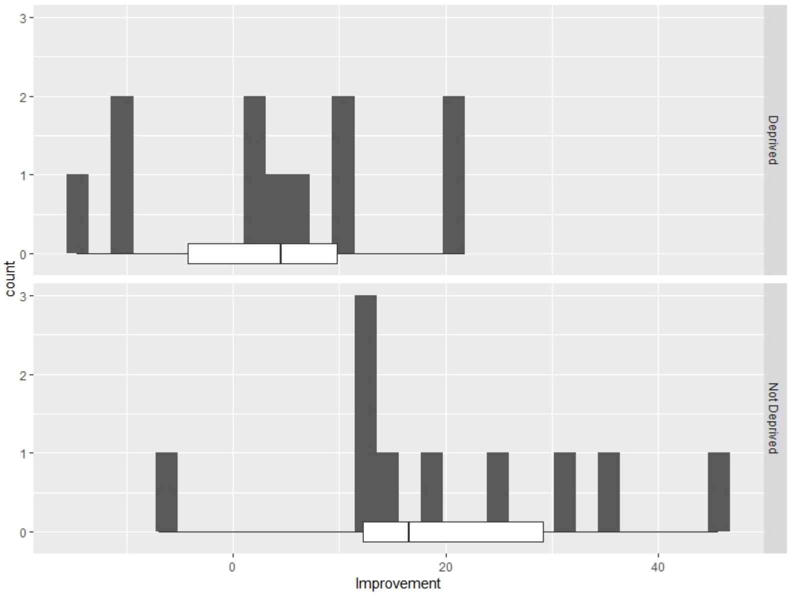
This was an experiment. The subjects were randomly assigned to the control and treatment groups.

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.

$H_0 : \mu_{unrestricted} - \mu_{restricted} = 0$.The long-run average improvement score in the unrestricted sleep group is the same as the long-run average improvement score in the sleep deprived group

$H_a : \mu_{unrestricted} - \mu_{restricted} > 0$.The long-run average improvement score in the unrestricted sleep group is greater than the long-run average improvement score in the sleep deprived group

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.



There is a clear divide of the two groups, it looks like the sleep deprived group had consistently lower improvement times.

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

	Minimum	**Lower Quartile**	**Median**	**Upper Quartile**	**Maximum**
Deprived	-14.7	-4.25	4.5	9.8	21.8
Not Deprived	-7	12.2	16.6	29.2	45.6

`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.

Deprived IQR = 9.8 - (-4.25) = 14.05

Not Deprived IQR = 29.2 - 12.2 = 17

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

Deprived Outlier Lower Bound = -4.25 - 1.5 × 14.05 = -25.325

Deprived Outlier Upper Bound = 9.8 + 1.5 × 14.05 = 30.875

Not Deprived Outlier Lower Bound = 12.2 - 1.5 × 17 = -13.3

Not Deprived Outlier Upper Bound $= 29.2 + 1.5 \times 17 = 54.7$

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())
```

$$\bar{x}_{unrestricted} = 19.8$$

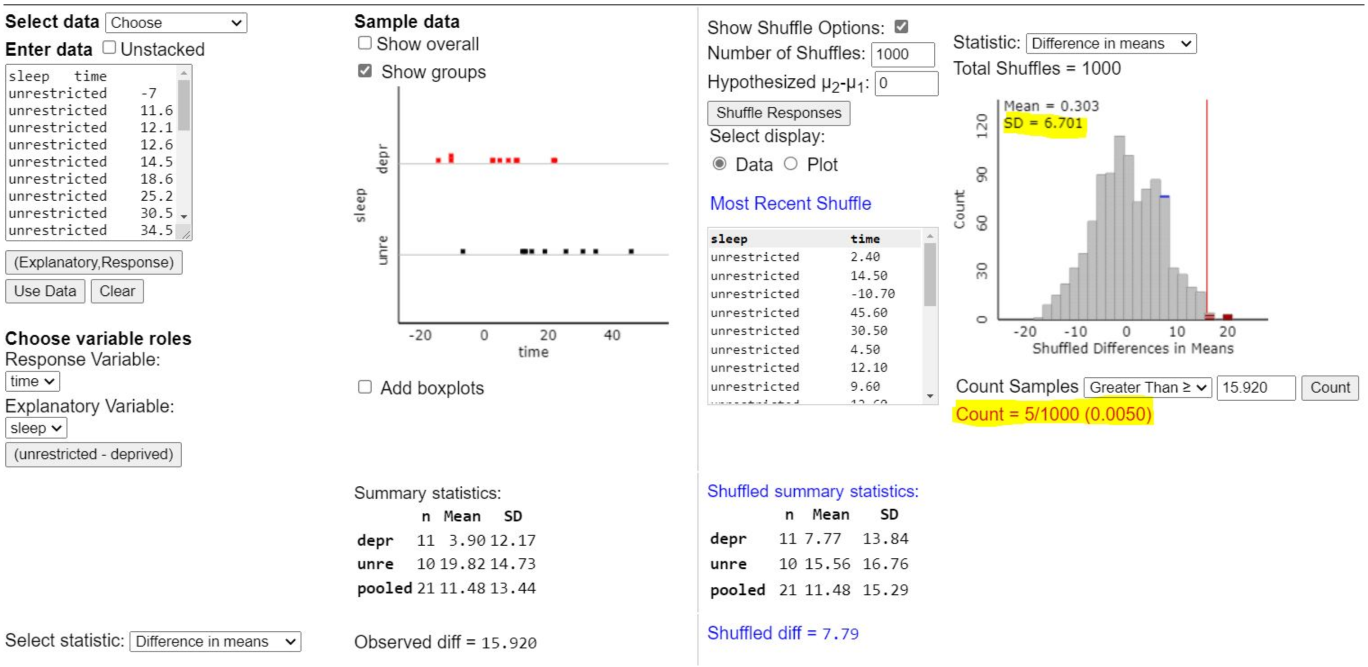
$$\bar{x}_{restricted} = 3.9$$

$$\bar{x}_{unrestricted} - \bar{x}_{restricted} = 19.8 - 3.9 = 15.9$$

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

No, here we only have 10 and 11 observations, so neither group has at least 20 observations.

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



Select data Choose

Enter data □ Unstacked

```
sleep   time
unrestricted    -7
unrestricted    11.6
unrestricted    12.1
unrestricted    12.6
unrestricted    14.5
unrestricted    18.6
unrestricted    25.2
unrestricted    30.5
unrestricted    34.5
```

Choose variable roles

Response Variable: time

Explanatory Variable: sleep

(unrestricted - deprived)

Select statistic: Difference in means

Sample data

□ Show overall

☑ Show groups

□ Add boxplots

Summary statistics:

```
        n  Mean   SD
depr    11  3.90 12.17
unre    10 19.82 14.73
pooled  21 11.48 13.44
```

Observed diff = 15.920

Show Shuffle Options: ☑

Number of Shuffles: 1000

Hypothesized μ_2-μ_1: 0

Select display: ● Data ○ Plot

Most Recent Shuffle

```
sleep          time
unrestricted   2.40
unrestricted   14.50
unrestricted   -10.70
unrestricted   45.60
unrestricted   30.50
unrestricted   4.50
unrestricted   12.10
unrestricted   9.60
```

Shuffled summary statistics:

```
        n  Mean   SD
depr    11 7.77   13.84
unre    10 15.56  16.76
pooled  21 11.48  15.29
```

Shuffled diff = 7.79

Statistic: Difference in means

Total Shuffles = 1000

Mean = 0.303

SD = 6.701

Count Samples Greater Than ≥ 15.920

Count = 5/1000 (0.0050)

SD = 6.701

$t = \frac{observed - null}{SD} = \frac{15.9-0}{6.701} = 2.37278$

```
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.

Our p-value is 0.005, which is very strong evidence against the null hypothesis.
The p-value of 0.007 is the probability of observing a difference of 15.9 ms or larger assuming that sleep deprivation has no effect on improvement scores.

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

$CI = statistic \pm M \times se = 15.9 \pm 2 * 6.701 = (2.198, 29.302)$

I am 95% confident that the long-run mean improvement score is 2.198ms to 29.302ms higher with the unrestricted sleep treatment than with the restricted sleep treatment.

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

You cannot generalize these results, as the subjects were a convenience sample of volunteers. However, as it was a randomized experiment, you can infer a cause-and-effect relationship between the treatment group (sleep deprived vs. not sleep deprived) and the improvement score.

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?

Despite some potentially strong and biased feelings,

$H_0 : \mu_{Marvel} - \mu_{DC} = 0$

$H_a : \mu_{Marvel} - \mu_{DC} \neq 0$

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.

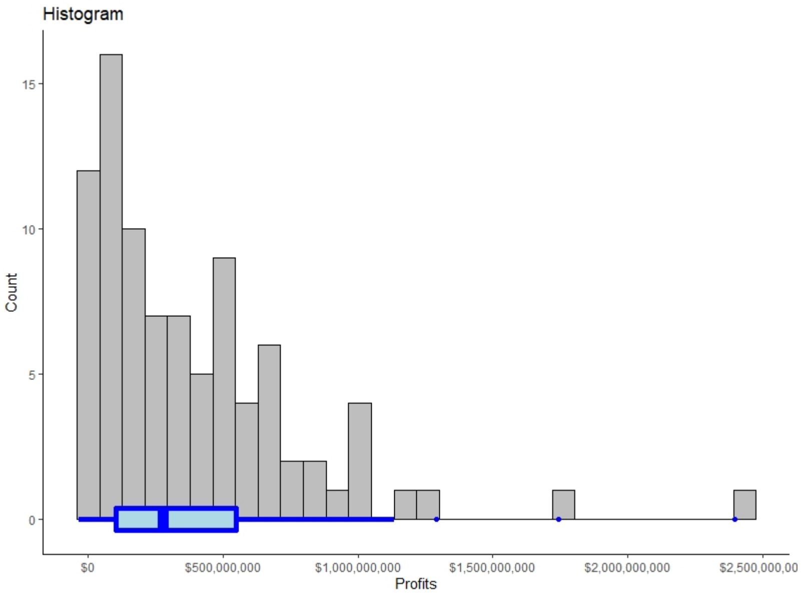
```
Movies2 <- Movies %>%
  mutate(Profit = Worldwide.Box.Office - Production.Budget)
```

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))
```

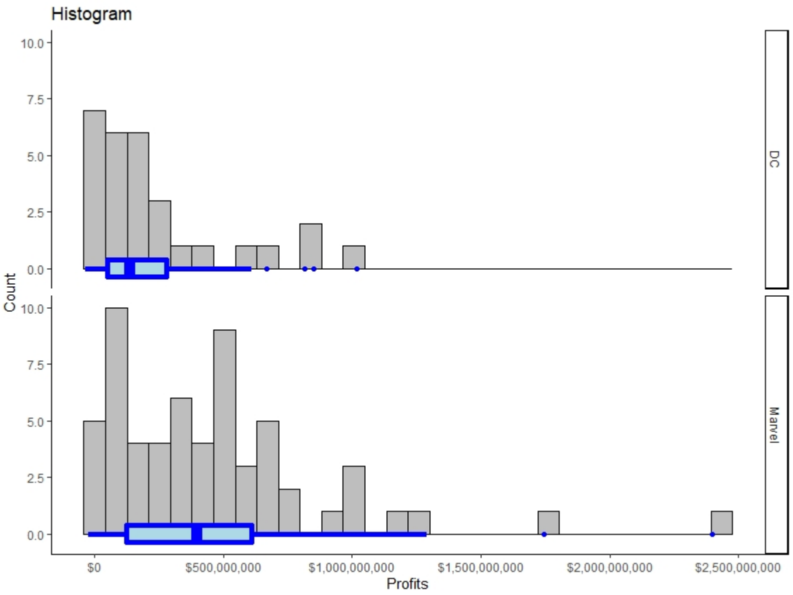
	Minimum	**Lower Quartile**	**Median**	**Upper Quartile**	**Maximum**
Profits	-35,977,304	102,782,518	278,664,533	547,862,775	2,397,800,564



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())
```

Initial observations looks like Marvel has some outliers above what D.C. has done. Additionally, we see that D.C. outliers are within the upper quartile range of Marvel movies and wouldn't be considered outliers there, leading credence to the alternate hypothesis that perhaps there is a difference between Brands and movie profits.



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

	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
D.C.	-35,977,304	48,976,250	136,439,693	278,664,533	1,017,507,517
Marvel	-22,387,720	123,441,792	394,014,998	609,876,854	2,397,800,564

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

$\bar{x}_{Marvel} = 460,595,147$

$\bar{x}_{DC} = 245,511,367$

$\bar{x}_{Marvel} - \bar{x}_{DC} = 460,595,147 - 245,511,367 = 215,083,780$

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()
```

Yes, we meet validity conditions. Our dataset has 29 DC movies and 60 Marvel movies. It could be argued there is too much skew to give accurate results.

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.

SD $= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 76883937.53$
$t = \frac{observed-null}{SD} = \frac{245511367-460595147}{76883937.5} = -2.7975$
p-value $= 2 \times (1 - pt(abs(t), n-2)) = 0.006339$
With a p-value of 0.006, we have very strong evidence against the null hypothesis that there is no difference between the mean revenue between Marvel and DC movies.

```
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.

observed $\pm M \times SE$ = 245511367 - 460595147 ± qt(1 - $\frac{0.05}{2}$, n-2) * sd = (-367,898,931 , -62,268,629).

I am 95% confident that the long-run mean difference in profits is between 62,268,629 and 367,898,931 higher for Marvel movies than for D.C. Movies.

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