MA206, Lesson 17 - Two Groups - Two Means

What types of variables do we compare when testing two means?
One categorical variable and one quantitative variable.

What are the validity conditions to use theoretical methods for two means?
Both groups are symmetric, OR
At least 20 observations in both groups and neither group is strongly skewed.

How do we compute the standardized statistic for two means?
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?
$(\bar{x}_1 - \bar{x}_2) \pm qt(1 - \frac{\alpha}{2}, n - 2) \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

What types of variables do we compare when testing two proportions?
Two categorical variables.

What are the validity conditions to use theoretical methods for two proportions?
At least 10 successes and 10 failures in each group.

How do we compute the standardized statistic for two proportions?
z = $\frac{(\hat{p}_1-\hat{p}_2)-(\pi_1-\pi_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$

How do we compute the p-value for the standardized z statistic?
For a less than alternative hypothesis test, pnorm(z)
For a greater than alternative hypothesis test, 1 - pnorm(z)
For a not equal to alternative hypothesis test, 2 * (1 - pnorm(abs(z)))

How do we calculate the confidence interval for the difference in two means?
$(\hat{p}_1 - \hat{p}_2) \pm qnorm(1 - \frac{\alpha}{2}) \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

1) A waitress wanted to see if introducing herself to customers by name increased her tips. She collected data on two-person parties that she waited on during Sunday brunch, which had a fixed price of $23.21. For each party, the waitress flipped a coin to determine if she would give her name as part of her greeting or not. She then kept track of her tip at the end of the meal. Her results are shown in the table below. There was no skew that would prevent theoretical methods from being used.

	Average Tip	**Standard Deviation**	**Count**
Gave Name	$5.44	$1.75	20
Did Not Give Name	$3.49	$1.13	20

a) In your own words, what is the research question?

Does the waitress introducing herself to customers lead to an increase in tips?

b) How many variables are we looking at? Are they categorical or quantitative?

We are looking at two variables.
One categorical variable and one quantitative variable. This is a difference in means problem.

c) Using words and symbols, what is the null and alternate hypothesis?

$H_0 : \mu_1 - \mu_2 = 0$. There is no difference in average tips between those visits that the waitress introduces herself by name and those visits where she does not.
$H_a : \mu_1 - \mu_2 > 0$. The average tips for those visits that the waitress introduces herself by name is greater than those visits where she does not.

d) Report the standardized statistic, p-value, and 95% confidence interval.

t = 4.186343
p-value = 0.0000809 ($8.0914e^{-5}$)
CI = (1.0070, 2.8930)

e) Interpret your results in the context of the original problem.

With a p-value of 0.00008, we have very strong evidence against the null hypothesis that there is no association between introductions and tips at this restaurant.
In fact, we are 95% confident that, on average, she makes between $1.01 and $2.89 more in tips by introducing herself than she makes if she does not introduce herself first.

f) Can you generalize your results? Can you infer causation?

As customers were randomly assigned to a treatment, we can consider this experiment as being randomly assigned and thus infer causation.
For generalization, our results are limited to patrons of this restaurant who were assigned to this particular waitress.

2) Following the release of Star Trek: Into Darkness (2013), YouGov was interested in preferences of the people of London. Inspired by London's presence in the future film, they conducted a random telephone survey of voting-aged residents of London and asked the respondents if they preferred Star Trek or Star Wars. They also collected other questions from the respondents, to include the political party they favored (Labour or Tory).

	Labour	**Tory**
Star Wars	229	86
Star Trek	184	101

a) How many variables are we concerned with? Are they categorical or quantitative?

We are looking at two variables, both are categorical.

b) What is the null and alternate hypothesis?

$H_0 : \pi_1 - \pi_2 = 0$. There is no association between political party and movie franchise preference.
$H_a : \pi_1 - \pi_2 \neq 0$. There is an association between political party and movie franchise preference.

c) Report your standardized statistic, p-value, and 90% confidence interval.

Here, we look at two proportions, so we will use a z-score and applicable formulas.
z = 2.14893
p-value = 0.03163993
CI = (0.0224, 0.1668)

d) Interpret your results in the context of the original problem.

Our p-value and standardized statistic give strong evidence that there is indeed a difference between political party and preference between Star Wars and Star Trek.
In fact, we are 90% confident that the proportion of Labor Party members who prefer Star Wars is between 0.0224 and 0.1668 higher than the proportion of Tory Party members who prefer Star Wars.

3) A student wanted to see if there was any association between whether students eat breakfast daily and their GPA. They surveyed 106 students at their college and asked if they ate breakfast and what their current GPA was. The results can be found in the CSV file titled 'BreakfastGPA' on teams or use the line of code below.

```
breakfast <- read_delim("http://www.isi-stats.com/isi/data/chap6/AP/BreakfastGPA.txt")
```

a) How many variables are we looking at? Are they categorical or quantitative?

We are looking at two variables. One is categorical and one is quantitative. This will be a difference in two means (two-sample t-test).

b) Complete the table below, listing the mean, standard deviation, and size of each group.

Note: If you want to get more decimal places from your **summarize code, save your filtering as its own named object. Then you can open that item from your environment to see your mean and SD to more decimal places. E.G., BreakfastSummary <- Breakfast %>% group_by(Breakfast) %>% %summarize(...*

	Average GPA	Standard Deviation	Count
Breakfast	3.561905	0.3103061	63
No Breakfast	3.412791	0.4346883	43

c) Plot a split histogram comparing both groups. Do we meet the validity conditions?

The histograms are sparse and it is hard to tell the true shape. Ideally more observations would be taken. This could be argued for or against theoretical methods here, but there is not evident skew which would definitively prevent theoretical methods.

d) Report the standardized statistic, p-value, and 99% confidence interval for the difference in GPAs between the two groups. Use theoretical computations regardless of your thoughts on ***b*** above.

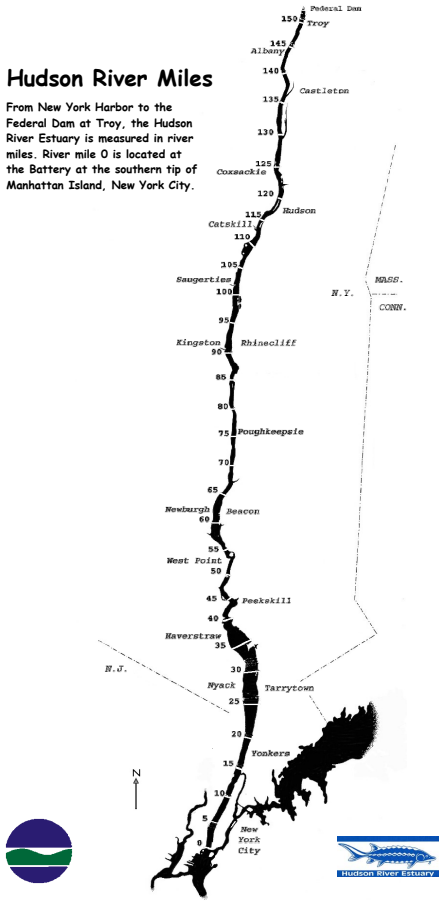
t = 1.937578
p-value = 0.05538688
CI = (-0.05282109, 0.35104922)

e) Interpret the results of your findings in terms of the original problem.

With a p-value of 0.055, we have only moderate evidence that there is an association between eating breakfast and average GPA. With a significance level of 0.01, we could not reject the null and must accept it as plausible that there is no association between the two.

4) The FDA limits PCBs within fish to 2 parts per million (ppm) or less to be considered fit for human consumption. The dataset HudsonFish.csv located on our Teams page lists samples of fish caught and reported from the Hudson River between 2001 and 2011. The dataset tracks which mile marker the fish was caught at (0 is the base of the Atlantic, while West Point is about mile marker 50 and Newburgh is 60) using variable *River MILE*. It also tracks which season the fish was caught (*Season*) and the amount of PCB (*Total PCB(ppm)*).

Your buddy wants to plan a fishing trip and tries to convince you that fish caught in the spring aren't as bad as the ones caught later in the year. Is there something to your friend's claim?



a) What is your research question?

Do fish caught in the Hudson River, on average, have less PCBs in the spring than those caught in different months?

b) How many variables are we interested in here? Are they categorical or quantitative?

We are going to look at two variables. One is categorical (Season) and one is quantitative (PCB(ppm)). This means we will be looking at a difference of two means.

c) Write out your null and alternate hypothesis.

$H_0 : \mu_1 - \mu_2 = 0$. The average PCB in in fish is the same regardless of the season caught.

$H_a : \mu_1 - \mu_2 < 0$. The average PCB in fish caught in spring is less than the PCBs of fish caught in other

seasons.

d) Generate a split histogram comparing the *Total PCB(ppm)* variable by Season. Comment on the shape of each.

(*Note: There are some very high (concerning levels of PCB) outliers which may make it difficult to view your results. To "zoom in", you may add "+ xlim(c(0,5))" to your chunk of ggplot code.*)

We have a large amount of fish caught in Spring and lower counts in Summer and Fall. There doesn't appear to be a large amount of skew.

```
fish %>%
ggplot(aes(x = Total.PCB.ppm.)) +
geom_histogram() +
facet_grid(Season ~.) +
xlim(c(0,5))
```

e) You may notice there are three seasons here, but we only interested in Spring or Not Spring Run the code below to create a new dataset which converts our categorical variable to a Success/Fail binary categorical variable to continue our analysis.

```
fish2 <- fish %>%
  mutate(Spring2 = Season=="Spring")
```

f) Generate a split histogram comparing the *Total PCB(ppm)* variable with your new variable, *Spring2*. Do we meet validity conditions?

Yes, we meet validity conditions - we have at least 20 observations in each group and there is not significant skew.

```
fish2 %>%
 group_by(Spring2) %>%
 summarize(mean = mean(Total.PCB.ppm.),
           s = sd(Total.PCB.ppm.),
           count=n())

fish2 %>%
  ggplot(aes(x = Total.PCB.ppm.)) +
  geom_histogram() +
  facet_grid(Spring2 ~.) +
  xlim(c(0,5))
```

g) Conduct a two-sample t-test. Report your standardized statistic, p-value, and 95% confidence interval.

t = -8.187246

p-value = $2.116905e^{-16}$

CI = $(-1.155223, -0.708777)$

h) In your own words, interpret your findings.

There is strong evidence that there is an association between time of the year and average PCB levels in the Hudson River. In fact, we are 95% confident that, on average, there are between 0.709 and 1.155 parts per million less PCBs in fish caught in Spring than those caught in other months. However, these findings are from the entire Hudson River. We could use the Mile Marker data to specify a range of the Hudson River that we intend to fish in to get more applicable levels of PCB risk. Furthermore, these are only reported fish. It is possible that most sickly looking fish are returned and not reporting, so these findings may be biased towards healthier fish.