**MA206**, Lesson 17 - Two Groups - Two Means

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

What are the validity conditions to use theoretical methods for 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?

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

What are the validity conditions to use theoretical methods for two proportions?

How do we compute the standardized statistic for two proportions?
z =

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

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

**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?

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

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

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

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

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

**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?

**b)** What is the null and alternate hypothesis?

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

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

**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?

**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** | | | |
| **No Breakfast** | | | |

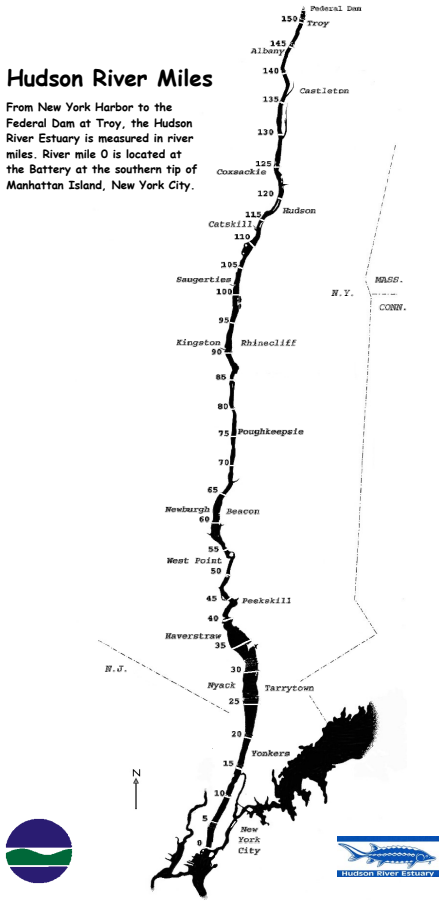
**c)** Plot a split histogram comparing both groups. Do we meet the validity conditions?

**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.

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

**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?

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

**c)** Write out your null and alternate hypothesis.

**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.*)

```
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?

```
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.

**h)** In your own words, interpret your findings.