The Titanic Shuffle

Directions: Follow along with the slides and answer the questions in **BOLDED** font in your journal.

Previously . . .

- In the last lab, we learned that by using a `do`-loop and the `resample` function, we could simulate shuffling our data many times.
- This helps us determine how likely it is that a difference between groups is due to chance.
- For this lab, we will determine if there is any evidence to the belief that wealthier passengers on the Titanic were more likely to survive than poorer passengers.
- We will consider wealthier passengers to be those that paid a higher fare for their ticket.

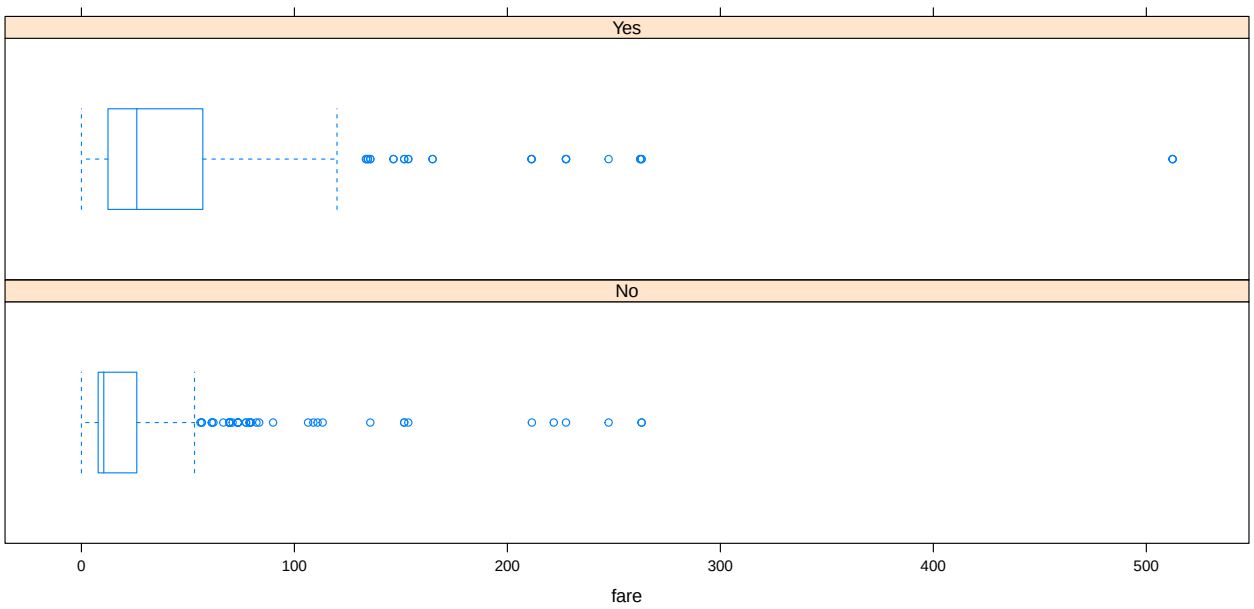
The Titanic

- The Titanic was a ship that sank en route to the U.S.A. from England after hitting an Iceberg in 1912.
- Use the following to load the `titanic` passenger and survival data.

```
data(titanic)
```

What do you think?

- Look at the plot below of fares paid separated by whether the passenger survived or not.
- **Do you believe richer passengers were more likely to survive? Why?**



Let's find an answer!

- Let's start by calculating how much more the *typical* survivor paid versus the *typical* non-survivor in our data.
- **Based on the boxplots from the previous slide, would you use the `mean` or `median` to calculate the *typical* fare paid by the passengers? Why?**
- **What was the *typical* fare paid by survivors? Non-survivers? How much more did the typical survivor pay? Write the code you used to find these answers.**

Do the shuffle!

- **Use a `do`-loop and the `resample` function to shuffle the passenger's survival status 300 times and compute each group's `median` fare paid.**
- **Write the code you used to perform your simulations**.
- **Use the `transform` function to add a variable called `Diff` to the shuffled medians you just calculated.**
- See the previous lab if you need help.
- The `Diff` variable should account for how much *more* the survivors paid than the non-survivors.
- **Write the code you used to calculate and add your `Diff` variable to your shuffled medians.**

Put your simulations data to use

- Create a plot of the differences in the fare paid for your randomized *survivors* and *non-survivors*.
- **What was the *actual* difference in the median fare paid by survivors and non-survivors in the data? Based on your plot, do you think this difference is *big*? Why?**

Convert & compare z-scores

- **What is the `mean` and standard deviaton of your 300 randomized differences?**
- **Convert the *actual* difference you computed to a z-score. Write down the z-score you calculated.**
- **How many standard deviations away from the mean is the actual median difference in fares paid by survivors and non-survivors?**

Extreme z-scores!

- Data scientists usually consider any z-score larger than 3 or smaller than -3 to be *extreme*.
- And by *extreme* we mean that they occur so rarely by chance alone that we start to believe that something besides chance alone is causing the z-scores to be so large.
- To show how rare these z-scores occur by random chance, let's use our 300 simulated median differences to estimate the probability of obtaining an *extreme* z-score.

Probably z!

- Follow along to convert your `Diff` variable to z-scores:

```
m_diff <- mean(~Diff, data=shfl_med)
```

```
s_diff <- stdev(~Diff, data=shfl_med)
```

```
shfl_med <-
  transform(shfl_diff,
            zscores=(Diff-m_diff)/s_diff)
```

- **Use the `subset` and `nrow` functions to compute the estimated probabilities of a z-score being larger than 3, smaller than -3, and larger than 3 OR smaller than -3.**

On your own

- **Redo your simulation and the analysis, BUT this time use the `mean` fare paid instead of the `median` fare paid.**
- **Does your conclusion change depending on the method you use to describe the *typical* fare paid by survivors and non-survivors? Explain.**
- After redoing your analysis using the `mean` instead of the `median`, answer the following:
- **If a journalist walked up to you right now and asked if the amount of fare paid for a Titanic ticket had an effect on a person's probability of surviving, what would you say? How would you justify your answer?**