Permuted Data and Graphics

Unit 2 - Lab 6

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

Why permute data?

- **Permuting** data is a really simple technique to *shuffle* or *randomize* our data.
  - You actually permuted data when you completed the *Horror Movie Shuffle* lab.
- Why do we bother?
  - So we can compare our data's distributions with examples that we know for a fact (100% certain) are random.
- And why is this helpful?
  - It lets us see what *Chance World* looks like so that we can compare our *real world* to *Chance World*.

The Titanic

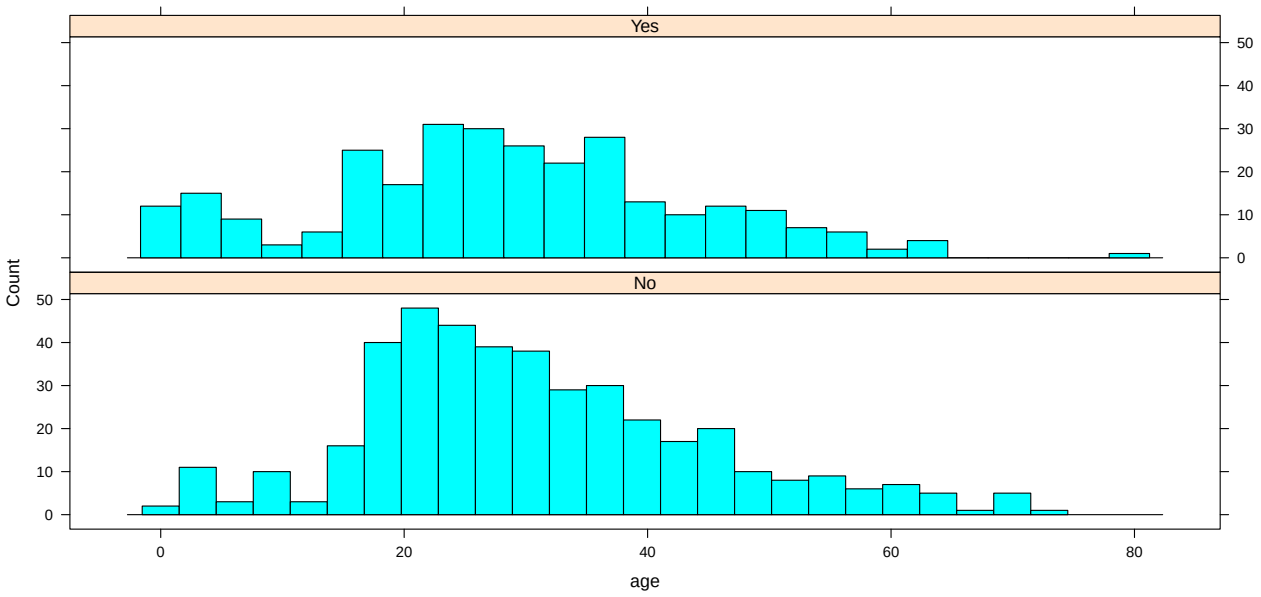
- The Titanic was a ship that sank en route to the U.S.A. from England after hitting an Iceberg in 1912.
- One common thought is that younger people were given priority when boarding the lifeboats.
- Use the following to load our `titanic` passenger and survival data.

```
data(titanic)
```

- In this lab, we'll take a brief look at missing data and practice comparing shuffled data to real data.

What do you think?

- Take a look at the following plot about the age of survivors and non-survivors.
- **Was the *typical* survivor younger than the *typical* non-survivor? Give an argument as to why.**



Let's find an answer!

- Let's start by calculating how much younger the *typical* survivor was than the *typical* non-survivor in our data.

```
mean(~age | survived, data = titanic)
```

- Notice that we get a warning message!
  - It appears we're missing 125 non-survivor's ages and 52 survivors ages.
  - Let's look at our `favstats` to see if this will be a problem.

```
favstats(~age | survived, data = titanic)
```

Missing values

```
favstats(~age | survived, data = titanic)
```

- Even though we're missing people in our data, we still have quite a few data points to use for calculating the *mean*.
- **Write down the number of people who are *NOT* missing for each group**
- **What do you think we mean when we say *missing data*? Why might it be missing?**

Back to our task

- Recalculate the mean `age` of survivors and non-survivors in our data.
- **How much younger is the average suvivor than the average non-survivor in our *actual* `titanic` data?**
  - **Write down this number in your data science journal.**

Randomizing our data

- Now that we've calculated the difference of our passenger's average ages, we'd like to compare it to data that we've randomized.
- Run the following to completely randomize the `survived` status for each person's `age`:

```
mean(~age | survived,
     data = resample(titanic,
                     shuffled="survived"),
     na.rm = TRUE)
```

Resampling (a.k.a. permuting)

```
mean(~age | survived,
     data = resample(titanic,
                     shuffled="survived"),
     na.rm = TRUE)
```

- When we write, `data = resample(titanic, shuffled="survived")`, we're telling R to:
  - Start with our original `titanic` data and ...
  - Take the the *Yes* and *No* values in our `survived` variable and mix them up...
- Then! With this randomized data:
  - Find the average `age` of our *randomized* survivors and non-survivors.

Why shuffle (or resample)?

- By shuffling the values of *Yes* and *No* in our `survived` variable:
  - We keep the same number of total survivors.
  - But! We *randomly* choose which passengers survived (or didn't ...)
- This means that, any relationship between `age` and those who `survived` is lost!

Answer the following:

- **How much younger is the average survivor than the average non-survivor in our *randomized* `titanic` data?**
- **Is this number very different than the difference we computed for our *actual* `titanic` data?**

But how "different" is 'different'?

- Comparing our actual data to the randomized data is not quite fair. -The randomized data is different every time.

- Instead of comparing to a single randomized data set, we want to compare to the *typical outcomes* we'd get from randomizing.
  - To do this, we need to randomize many, many times."
  - Then we can get a sense of how likely our *actual* values randomly appear.

do-ing things many times

- Just like how we computed our randomized data once, we can use the `do()` function to repeat our calculations many times.
- Run the following example:

```
do(3)*mean(~age|survived,
           data=resample(titanic,shuffled="survived"),
           na.rm = TRUE)
```

- **How many times did we randomize our data and then compute the average?**
- **Write down the code you would run to do our calculations 10 times.**
  - **Test your code to ensure it works.**

Calculate many means

- When we `do` a calculation many times, the results are stored as a `data.frame` type object.
  - This means we can save and manipulate our many calculations.
- **Re-run your code from the previous slide BUT ...**
  - **`do` the calculation 300 times AND ...***
  - **Assign the object the name `shuffled_means`**
- **Note**: Doing something 300 times can take a while. Be patient while your code runs.
  - You'll know it's done when the `>` symbol re-appears in your console.

Find many differences

we want to calculate how much younger our average chance-world survivor was than our average chance world non-survivor

- Now we want to calculate how much younger our average *Chance-World* survivor was than our average *Chance World* non-survivor.
- To do this, run:

```
shuffled_diffs <-
    transform(shuffled_means,
              Difference = Yes - No)
```

Looking at our differences

- After calculating our 300 randomized differences, make an appropriate plot to visualize them.
  - **Write down the code you used to make the plot of *differences*.**
- The values on the x-axis represent how much *younger* the average survivor was than the average non-survivor.
  - **What does it mean for our difference in average age to be *negative*?**

Making the call

- Compare the difference from our actual data, -2.2825, to the histogram of our randomized data.
  - **Is our *actual* difference close to the values in the center of our histogram? Or are they far away?**
  - **What does it mean for our *actual* value to be far away from the center of our randomized values?**

Making the call

- Read this part carefully:
  - If a *real-life event* is very common in *Chance-World*, then we might suspect that our *real-life* outcome was just due to chance, and not very *meaningful*.
  - But! if the *real-life* outcome is rare or unusual (i.e. Doesn't occur very often) in chance world, then we have evidence that the outcome is *meaningful*.
- **Based on your shufflings, do you think the difference in age of survivors is meaningful? Explain.**