Lab 2I - R's Normal Distribution Alphabet

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

## Where we're headed

* In the last lab, you were able to overlay a normal curve on histograms of data to help you decide if the data's distribution is close to a normal distribution.
  + We also saw that calculating the mean of random shuffles also produces differences that are normally distributed.
* In this lab, we'll learn how to use some other R functions to:
  + Simulate random draws from a normal distribution.
  + Calculate probabilities with normal distributions.

## Get set up

* Start by loading the titanic data and calculate the mean age of people in the data but shuffle their survival status 500 times.
  + Assign this data the name shfls.
* After creating shfls, use mutate to add a new variable to the data set. This new variable should have the name diff and should be the age of those who survived minus those who died.
* Finally, calculate the mean and sd of the diff variable.
  + Assign these values the name diff\_mean and diff\_sd.

## Is it normal?

* Before we proceed, we need to verify that our diff variable looks approximately normally distributed.
  + **Is the distribution close to normal? Explain how you determined this. Describe the center and spread of the distribution.**
  + **Compute the mean difference in the age of the *actual* survivors and the actual non-survivors.**

## Using the normal model

* Since the distribution of our diff variable appears normally distributed, we can use a normal model to estimate the probability of seeing differences that are more extreme than our actual data.
* Fill in the blanks to calculate the probability of an even smaller difference occurring than our actual difference using a normal model.

pnorm(\_\_\_\_, mean = diff\_mean, sd = \_\_\_\_)

## Extreme probabilities

* The probability you calculated in the previous slide is an estimate for how often we expect to see a difference smaller than the actual one we observed, by chance alone.
  + **Draw a sketch of a normal curve. Label the mean age difference, based on your shuffles, and the actual age difference of suvivors minus non-survivors from the actual data. Then, shade in the areas, under normal the curve, that are *smaller* than the actual difference.**
* If you wanted to instead calculate the probability that the difference would be larger than the one observed, we could run (fill in the blanks):

1 - pnorm(\_\_\_\_, mean = diff\_mean, sd = \_\_\_\_)

## Simulating normal draws

* We can simulate random draws from a normal distribution with the rnorm function.
  + Fill in the blanks in the following two lines of code to simulate 100 heights of randomly chosen men. Assume the mean height is 67 inches and the standard deviation is 3 inches.
  + Plot your simulated heights with a histogram.

draws <- rnorm(\_\_\_\_, mean = \_\_\_\_, sd = \_\_\_\_)

histogram(draws, fit = \_\_\_\_)

## P's and Q's

* We've seen that we can use pnorm to calculate *probabilities* based on a specified *quantity*.
  + Hence, why we call it "P" norm.
* Now we'll see how to do the opposite. That is, calculate a the *quantity* for a specific *probability*.
  + Hence why we'll call this a "Q" norm.
* How tall can you be and still be in the shortest 25% of heights if the mean height is 67 inches with a standard deviation of 3 inches?

qnorm(\_\_\_\_, mean = \_\_\_\_, sd = \_\_\_\_)

## On your own

* Using the titanic data, answer the following statistical question:
  + **Were women on the Titanic typically younger than men?**
  + **Use a histogram, 500 random shuffles and a normal model to answer the question in the bullet above.**