Eyeballing Normal

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

We love the Normal curve, we do!

- Data scientists love the Normal curve.
- Lots of real distributions of data appear to follow the shape of the Normal curve.
- This opens the door to a lot of sweet applications for using the *Normal model*.
- When data appears to be *Normally distributed*, we can use the *Normal model* to:
- Simulate *Normally distributed* data.
- Easily compute probabilities.

Oh Normal curve we love you!

- So in this lab, we will look at some previous data sets to see if we can find data that is roughly Normally distributed.
- We will find out what is needed to be able to simulate Normally distributed data.
- We will use the Normal model to compute probabilites.

The ol' CDC data

- Type `load(cdc)` to load up our CDC data.
- Create a histogram for the `height` of the people in our `cdc` data.
- **Write down, based on what you know about the Normal distribution, whether or not you think the data is Normally distributed. Explain.**
- **Calculate the mean and standard deviation of `height`.**
- **Find the value of the mean on the histogram. Does the distribution seem roughly symmetric?**

Over-laying Normal curves

- When trying to determine if data may be Normally distributed, it's often helpful to overlay an actual Normal curve onto the data.
- To do this for our CDC `height` variable, type:

```
histogram(~height, data=cdc,
          fit='normal')
```

- When we include the argument `fit='normal'` to the `histogram` function, we are telling R to fit a Normal curve to our data's distribution.
- **Based on the histrogram with the Normal curve over-layed, do you think that people's heights are Normally distributed? Why?**

Eyeballing Normal distributions

- Often times, the best and simplest way to decide if something is Normally distributed is to just eyeball it.
- The distribution of people's heights really doesn't look very Normally distributed.
- The *peak* of the distribution seems too flat.
- The distribution also appears to spread out too much.
- Create another histogram of people's heights; but this time, *split* the plots by 'gender' and then overlay the normal curves.
- **Do the histograms for each gender look more Normal than the plot where they were together? Explain your reasoning.**

Before we go on:

- Data scientists often try to fit models to data so that they can carry out simulations.
- Based on your plots, answer the following:
- **Would you recommend a data scientist use the *Normal* model to simulate people's heights? If *yes*, explain why. If *no*, explain which values would occur too often.**
- **Would you recommend a data scientist use the *Normal* model to simulate each gender's heights? Why or why not?**

Practice, practice, practice

- Run the following to load the `movie` data:

```
data(movie)
```

- This data set contains a variety of information about movies from Rotten Tomatoes.
- One of the variables is `reviews_num`, which is the number of reviews Rotten Tomatoes used to create its movie rating.
- Let's practice and decide if the number of reviews is Normally distributed.

The Eye giveth ... and also tricketh

- Start by creating the following histogram

```
histogram(~reviews_num, data=movie,
          nint = 10, fit = 'normal')
```

- **Does the distribution look *Normal*? Why or why not?**
- Now type this:

```
histogram(~reviews_num, data=movie,
          nint = 11, fit = 'normal')
```

- **Does this plot look more or less Normally distributed than the previous one? Explain.**
- **Would you recommend using a *Normal* model? Why?**

Using Normal Models

- Data scientists like using Normal models because it often resembles real data.
- But not EVERYTHING is Normally distributed.
- As a data scientist in training, you must decide when a Normal model seems appropriate.
- No model is ever perfect 100% of the time.
- If you choose a model, you should be able to justify why you chose it.

On your own

- Load the `titanic` data set.
- **Which variables do you think might fit a Normal curve?**
- **Which variables would you say *definitely* would NOT fit a Normal curve?**
- Be sure to try altering the number of bins to see if it helps your data look more or less normal.
- Also try splitting the data based on `gender`, or whether the person `survived`.
- **Write down the variables you think look roughly Normal. Include the code and any relevant plots you used to make your decision.**