Homework 3

Amit Arora

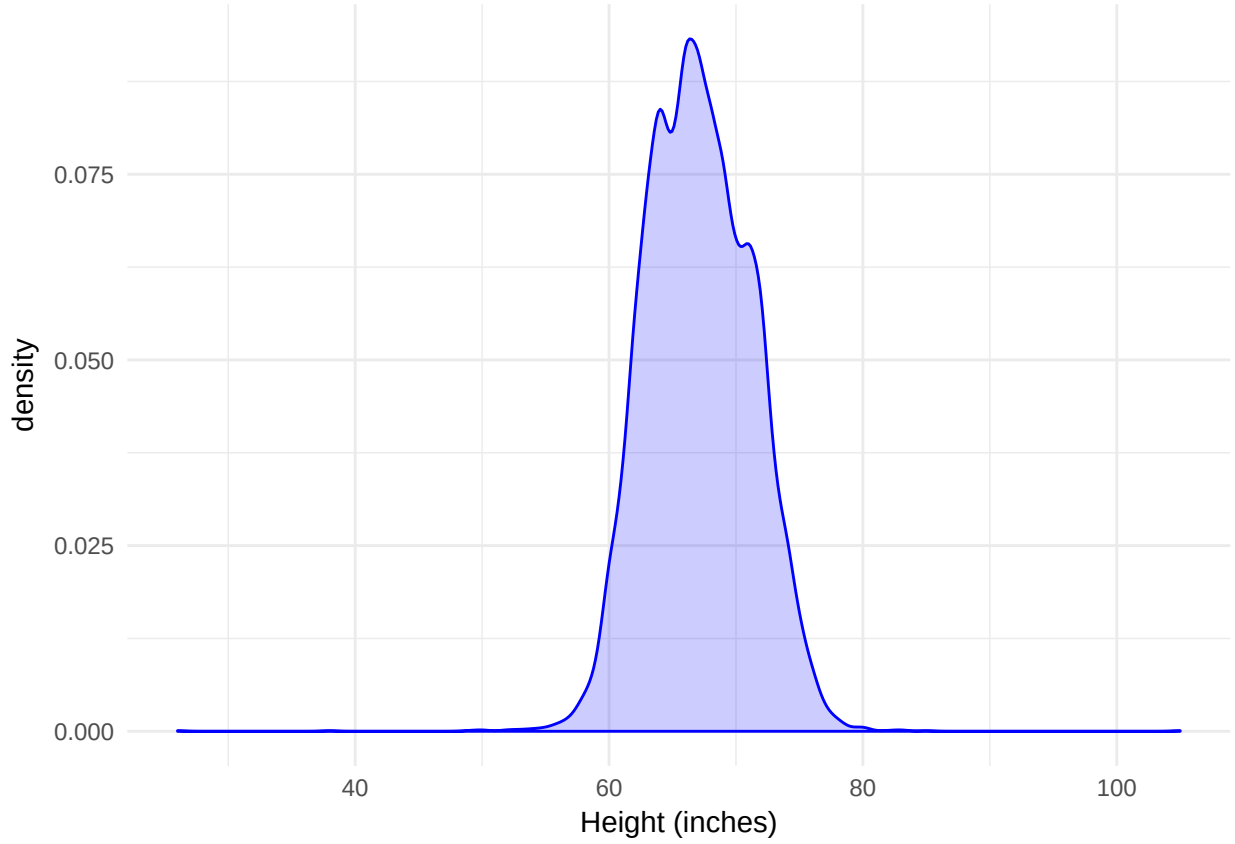
July 18, 2018

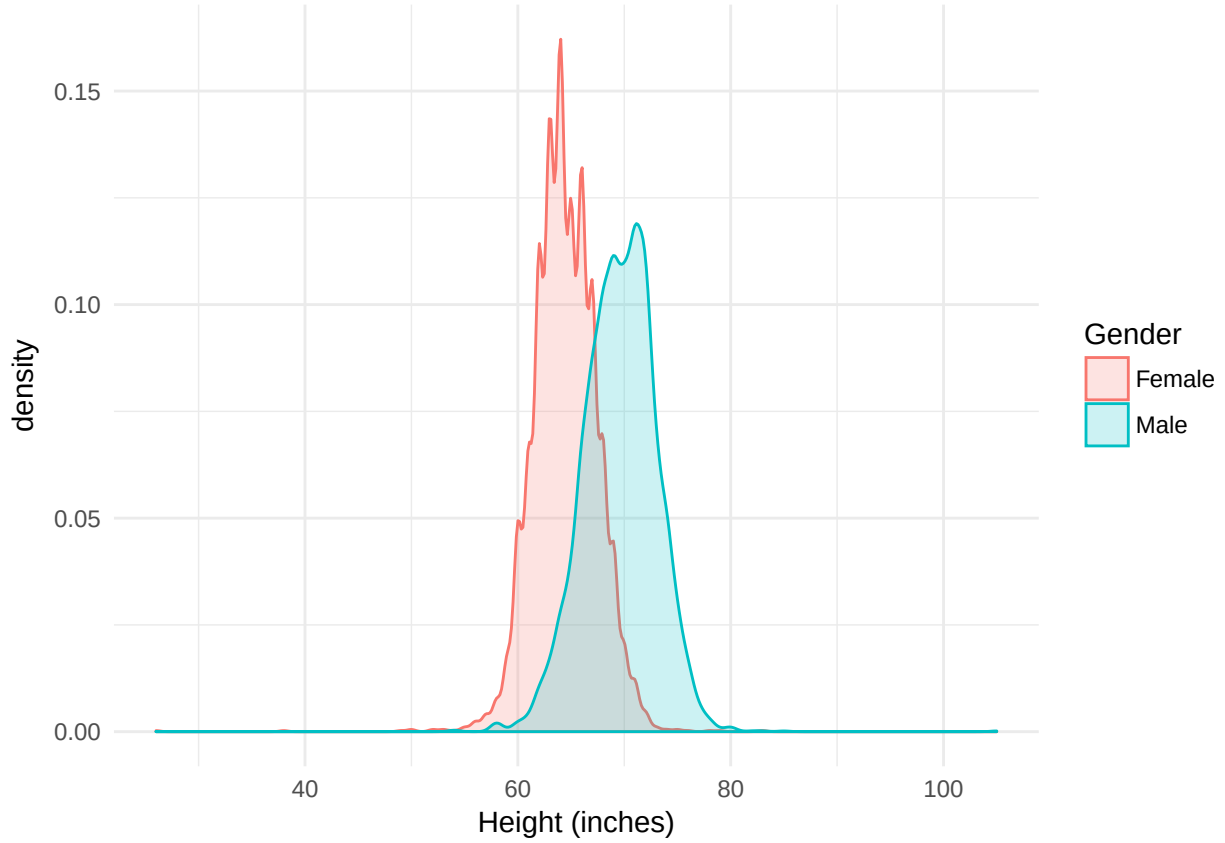
Distributions

Lets see distributions in action, for that we go back to the youth dataset.

1. Read the youth.csv file into a dataframe.
2. Plot densities for `Height (inches)` the field. Write a few lines about what you read from this density plot.
3. To make things interesting, lets plot the densities by Gender. Use the "Gender" field to color code the densities. What do you observe? You would need to use the "color" field and "fill" field in the geom_density in the aes function (you want to say *fill=Gender*).

The plots should like as shown below. What you observe here is also (in a very simple way) an introduction to machine learning, we will discuss this more in class.





Expected value

We discussed in class that the expected value can be thought of as the long term average of repeated experiments. For example if we take a random sample from a population and calculate a point estimate such as the mean of the sample, then we repeat it and repeat it again and again, now we calculate the mean of means we will find that if we repeat the experiments a large number of times it will converge to the actual population means.

Lets see this in action. Take the `Height (inches)` field from the youth dataset and do the following.

1. Read the dataset into a dataframe.
2. Store the `Height (inches)` field in a new variable called *height*, this is just for ease of coding since we are only going to work with this one field for now.
3. Take a random sample of 100 values from the height vector (remember we stored the `Height (inches)` in a new variable). This can be done using the *sample* function. So if you do sample (100, 10) it will get you a set of 10 random numbers from 1 to 100, then you can use these as indices into the height array to get a set of 10 random values from the height array, for example *height[sample(100, 10)]*. The key thing is that you dont want 10 out of 100 samples but you want 100 samples out of total number of elements in the height vector.

3.1 Use the map function to run a loop from 1 to 100 and calculate the mean of every sample (where each sample contains 100 values). You would need to wrap the map function in an *unlist* function to return the output as a vector of length 100.

4. Plot density curve for the returned mean values. What do you observe?

5. Calculate the mean of the 100 mean values. How does this mean compare with the mean of the height vector.
6. Repeat this experiment 10,000 times, now repeat it 100,000 times. What do you observe?