Stat 140 - Quiz 1 Sample

What's Your Name? _____

Which section are you in? _____

This is a sample quiz. For the real quiz, I will use a different data set, but will pick roughly 2-3 of the questions that are below and adapt them to the new data set with minimal modification.

Below are the first few rows of a data frame named `NHANES`. NHANES stands for "National Health and Nutrition Examination Surveys", and the data frame contains information about the health of randomly sampled Americans.

```
##      ID Gender Age Weight Height   BMI BPSysAve BPDiaAve Diabetes
## 1 51624   male  34   87.4  164.7 32.22      113       85       No
## 2 51625   male   4   17.0  105.4 15.30       NA       NA       No
## 3 51630 female  49   86.7  168.4 30.57      112       75       No
## 4 51638   male   9   29.8  133.1 16.82       86       47       No
## 5 51646   male   8   35.2  130.6 20.64      107       37       No
## 6 51647 female  45   75.7  166.7 27.24      118       64       No
```

1. What is each observational unit in this data set?

Each observational unit is an American.

2. For each of the following variables, is that variable categorical or quantitative? If it is categorical, is it ordinal or nominal?

- `Gender`: categorical, nominal
- `Height`: quantitative
- `Diabetes`: categorical, ordinal (No is better than Yes)

3. The following command counts how many observational units are in each combination of levels of the gender and diabetes variables.

```
tally(Diabetes ~ Gender, data = NHANES)
```

```
##         Gender
## Diabetes female male
##      No    3088 3013
##      Yes    269  283
```

a. Calculate the joint distribution of Diabetes and Gender

The joint distribution of Diabetes and Gender says what proportion of all observational units in the data set are in each combination of levels for those variables. Divide the counts for the combinations of those two variables by the total number of observations in the data set. Leave the answer as a proportion (between 0 and 1).

The total is 3088 + 3013 + 269 + 283 = 6653.

female, no diabetes: 3088/6653 = 0.464

female, diabetes: 269/6653 = 0.040

male, no diabetes: 3013/6653 = 0.453

male, diabetes: 283/6653 = 0.043

b. Calculate the marginal distribution of Diabetes

The marginal distribution of diabetes is the distribution of diabetes across all people in the study. To find the marginal distribution, we need to first find the total number of people in each level of the diabetes variable:

total for no diabetes: 3088 + 3013 = 6101

total for yes diabetes: 269 + 283 = 552

The marginal distribution of diabetes is then the proportion of all observational units in each level of the diabetes variable:

no diabetes: 6101 / 6653 = 0.917

diabetes: 552 / 6653 = 0.083

c. Calculate the conditional distribution of Diabetes given that the subject's Gender is male

For this conditional distribution, we are restricting our attention to just the observational units whose gender was recorded as male. Among those people, what proportion had or did not have diabetes?

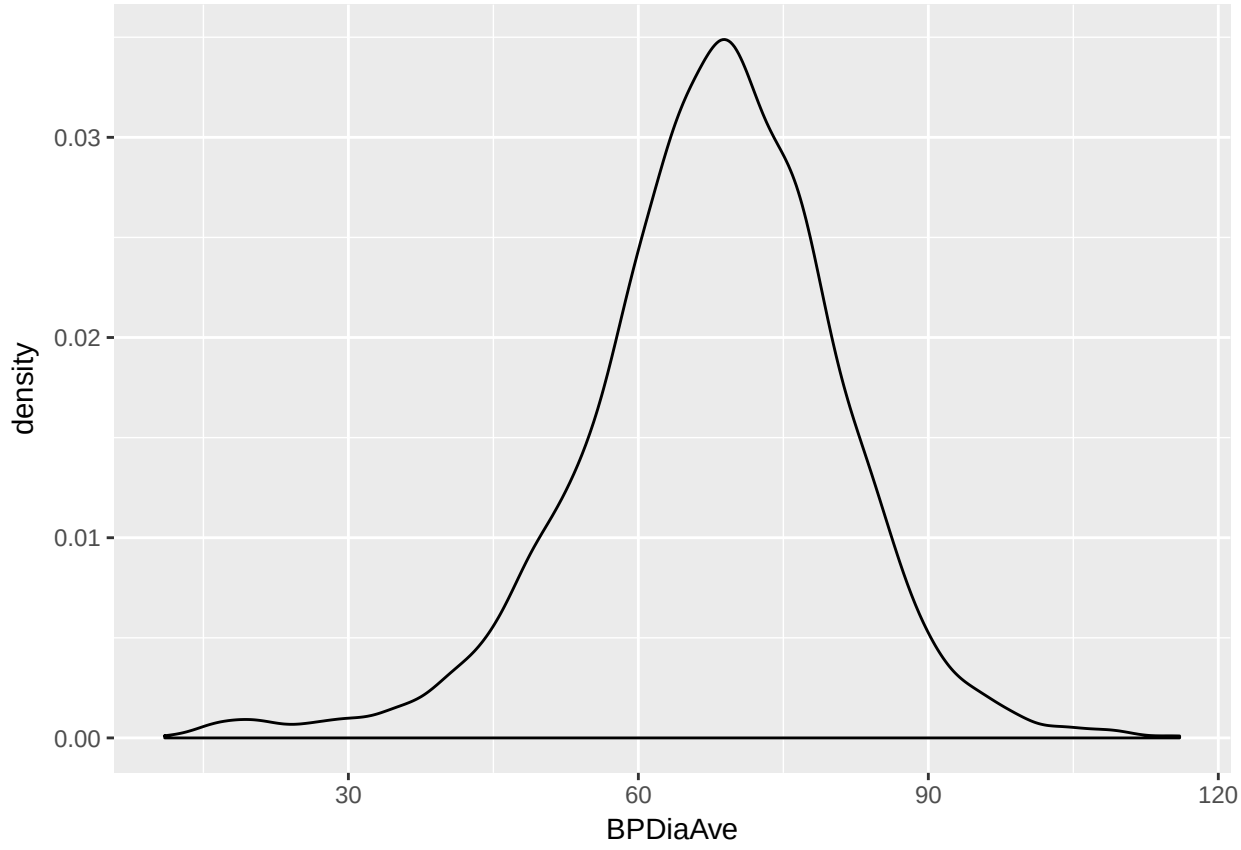
There were a total of 3013 + 283 = 3296 people in the data set whose gender was recorded as male.

no diabetes: 3013/3296 = 0.914

diabetes: 283/3296 = 0.086

4. Here is a plot of the study participants' blood pressure measurements:

```
ggplot(data = NHANES, mapping = aes(x = BPDiaAve)) + geom_density()
```



a. What statistics could you use to summarize the center of this distribution?

This distribution is unimodal and symmetric, so we could summarize its center with either the mean or the median. Most commonly, the mean would be used in this setting.

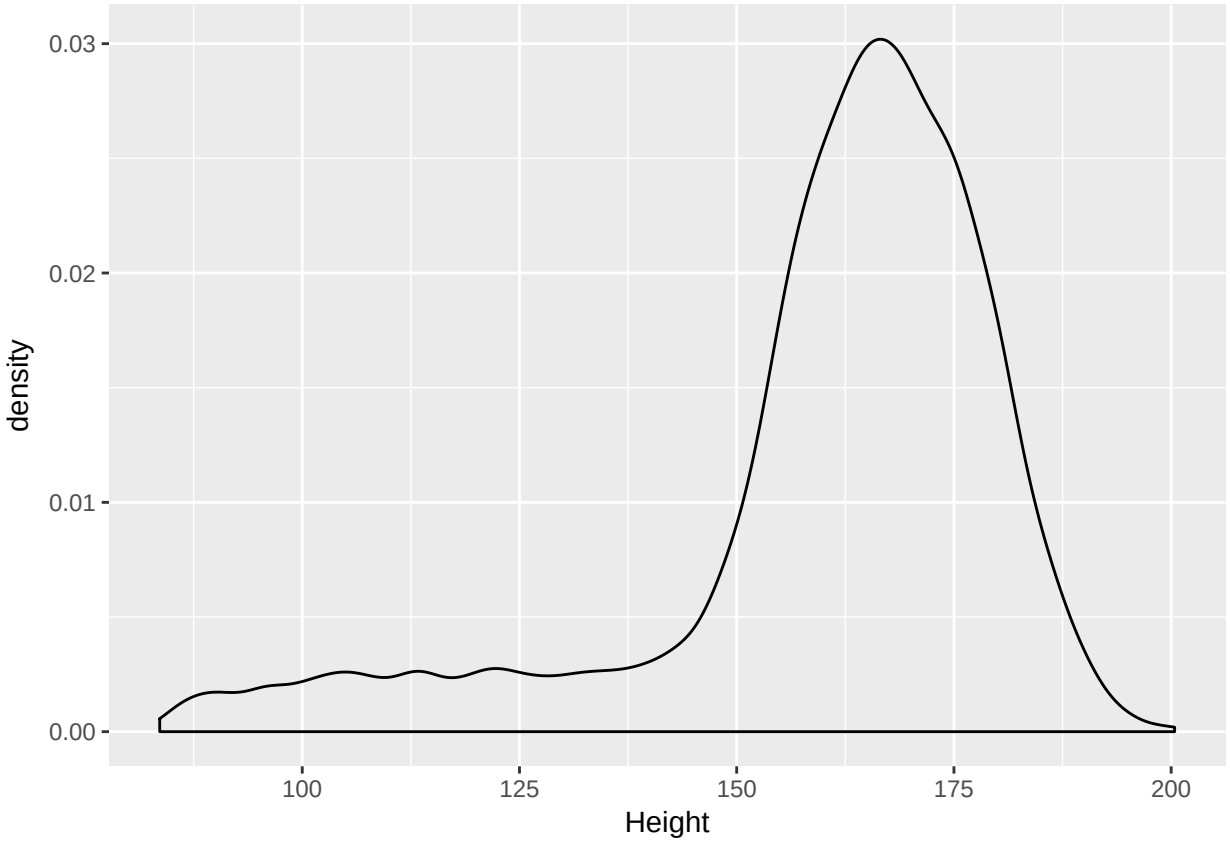
b. What statistics could you use to summarize the spread of this distribution?

This distribution is unimodal and symmetric, so we could summarize its spread with any of the standard deviation, variance, or interquartile range. Most commonly, the standard deviation would be used in this setting.

5. Here is a plot of the study participants' heights:

```
ggplot(data = NHANES, mapping = aes(x = Height)) + geom_density()
```

```
## Warning: Removed 298 rows containing non-finite values (stat_density).
```



a. What statistics could you use to summarize the center of this distribution?

This distribution is skewed to the left, so we would summarize its center with the median.

b. What statistics could you use to summarize the spread of this distribution?

This distribution is skewed to the left, so we would summarize its spread with the interquartile range.

6. For each of the following pairs of variables, circle the type of plot you would make, and write down the type of geometry you would use to make that plot. (more than one answer may be correct – if so, choose one)

6a. `Diabetes` and `Gender`

Bar Plot **This is the answer**

Box Plot

Density Plot with groups shown in different colors

Scatter Plot

Geometry type: geom_bar

6b. `Age` and `BPSysAve` (`BPSysAve` is a measure of blood pressure)

Bar Plot

Box Plot

Density Plot with groups shown in different colors

Scatter Plot **This is the answer**

Geometry type: geom_point

6c. `Diabetes` and `BPSysAve`

Bar Plot

Box Plot **This is one possible answer**

Density Plot with groups shown in different colors **This is one possible answer**

Scatter Plot

Geometry type: geom_boxplot or geom_density, depending on which option you chose above.

6d. `Weight` and `Height`

Bar Plot

Box Plot

Density Plot with groups shown in different colors

Scatter Plot **This is the answer**

Geometry type: geom_point