huber-data-1

In this course, I'll be teaching you the basics of data analysis. It probably makes sense to start by defining the word DATA.

According to Wikipedia, "Data are values of qualitative or quantitative variables, belonging to a set of items."

Often the "set of items" that we are interested in studying is referred to as the POPULATION. Data analysis usually involves studying a subset, or SAMPLE, of an entire population.

Here is a diagram showing the relationship between a population and a sample.

Data analysis should always start with a specific question of interest. For example, we might ask "What percentage of people living in the United States are over six feet tall?"

Here, our population of interest is everyone living in the US. Since it's impractical to measure the heights of over 300 million people, we could instead choose 100 people at random and measure their heights. Our hope would be that this sample of 100 people is REPRESENTATIVE of the entire US population.

Lets quickly test your understanding of the term REPRESENTATIVE. If you were interested in studying the health of men living in the US, ages 18-25, which sample would be more representative of the target population– a sample of 50 men who live in a nearby retirement home, or a sample of 50 men who are students at a local university?

1. Men living at the retirement home
2. College students

College students

Would you like to watch a video on these topics now?

The purpose of analyzing a sample is to draw conclusions about the population from which the sample was selected. This is called INFERENCE and is the primary goal of INFERENTIAL STATISTICS.

In order to make any inferences about the population, we first need to describe the sample. This is the primary goal of DESCRIPTIVE STATISTICS.

If we want to describe our sample using just one number, how would we best do it? A good start is to find the center, the middle, or the most common element of our data. Statisticians call this the CENTRAL TENDENCY.

There are three different methods for finding such a number and the applicability of each method depends on the situation. Those three methods are called the MEAN, MEDIAN, and MODE.

Mean, median, and mode are all measures of _____.

1. variation
2. significance
3. deviation
4. central tendency

central tendency

Which of the following terms are of most importance when describing the central tendency of a data set?

1. median, mode, range
2. statistics, population, mode
3. population, sample, representative
4. mode, median, mean

mode, median, mean

To illustrate these concepts, we will now look at a real dataset from the 'openintro' R package, which has already been loaded for you. Type 'cars93' and press Enter to see the dataset we'll be working with.

```
cars93
```

```
## # A tibble: 54 × 6
##    type    price mpg_city drive_train passengers weight
##    <fct>   <dbl>    <int> <fct>            <int>  <int>
##  1 small    15.9       25 front                5   2705
##  2 midsize  33.9       18 front                5   3560
##  3 midsize  37.7       19 front                6   3405
##  4 midsize  30         22 rear                 4   3640
##  5 midsize  15.7       22 front                6   2880
##  6 large    20.8       19 front                6   3470
##  7 large    23.7       16 rear                 6   4105
##  8 midsize  26.3       19 front                5   3495
##  9 large    34.7       16 front                6   3620
## 10 midsize  40.1       16 front                5   3935
## # … with 44 more rows
```

You will notice the rows are numbered 1 through 54, each representing exactly one car in the dataset. For each car, the following VARIABLES, or characteristics, are reported: 'type' (small, midsize, large), 'price' (USD), 'mpg_city' (city miles per gallon), 'driveTrain' (4WD, front, rear), 'passengers' (total capacity), and 'weight' (lbs).

We will be focusing on the 'mpg_city' variable in this lesson. For simplicity, let's extract it from our dataset and store it in a new variable.

Access the 'mpg_city' variable from the 'cars93' dataset using the 'dataset$variable' notation.

```
cars93$mpg_city
```

```
##  [1] 25 18 19 22 22 19 16 19 16 16 21 17 20 20 29 23 21 29 20 31 23 21 18 46 42
29 22 20
## [29] 17 18 18 17 18 29 28 19 19 29 18 29 21 23 19 31 19 19 28 33 25 39 32 22 25
20
```

Now store the contents of the 'cars93$mpg_city' in a new variable called 'myMPG'.

```
myMPG <- cars93$mpg_city
```

The ARITHMETIC MEAN, or simply the MEAN or AVERAGE, is the most common measurement of central tendency. To calculate the mean of a dataset, you first sum all of the values and then divide that sum by the total number of values in the dataset.

However, when there are many values of interest, it becomes tedious to do this calculation by hand.

Luckily, R has a built-in function for computing the mean. The syntax for doing so is 'mean(variable)'.

Compute the mean value for the 'myMPG' variable now.

```
mean(myMPG)
```

```
## [1] 23.31481
```

Extreme values in our dataset can have a significant influence on the mean. For instance, if there was a car in our dataset that got 200 miles per gallon, this would inflate the mean upwards. This could be misleading since none of the other cars93 get anywhere near this gas mileage.

An alternative to the mean, which is not influenced at all by extreme values, is the MEDIAN. The median is computed by sorting all values from least to greatest and then selecting the middle value. If there is an even number of values, then there are actually 2 middle values. In this case, the MEDIAN is equal to the MEAN of the 2 middle values. Don't worry if this is a little confusing. It will become more clear with practice.

R also has a function for computing the median of a dataset and this is done by typing 'median(variable)'. Find the median value of your 'myMPG' variable now.

```
median(myMPG)
```

```
## [1] 21
```

Finally, we may be most interested in finding the value that shows up the most in our dataset. In other words, what is the most common value in our dataset? This is called the MODE and it is found by counting the number of times that each value appears in the dataset and selecting the most frequent value.

Use the 'table' function to see how many times each value appears for your 'myMPG' variable. The syntax for this function is the same as for the others you've seen.

```
table(myMPG)
```

```
## myMPG
## 16 17 18 19 20 21 22 23 25 28 29 31 32 33 39 42 46
##  3  3  6  8  5  4  4  3  3  2  6  2  1  1  1  1  1
```

Look at your table for the 'myMPG' variable that you created above. The first row gives you the value of your variable and the second row gives you the number of times it appears in your dataset. Since the mode is the value of our variable that appears most frequently, what is the mode of your 'myMPG' variable?

Congratulations! You've made it through your first lesson. We introduced basic concepts related to data and data analysis. Specifically, you learned three important measures of central tendency: mean, median, and mode. You also know how to compute these using R.