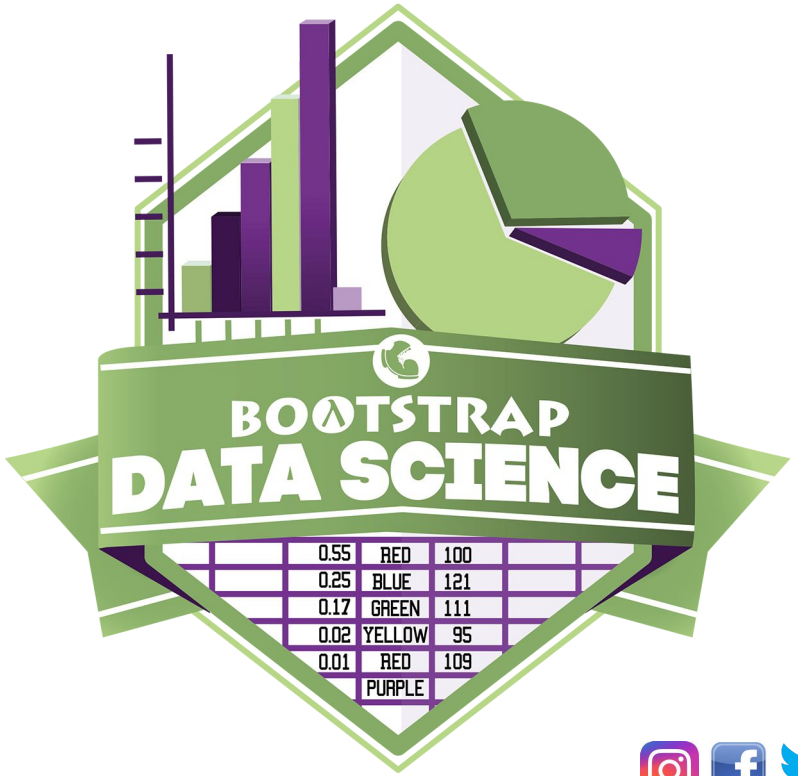
Measures of Center



@BootstrapWorld

Mean

According to the Animal Shelter Bureau, the average pet weighs almost 40 pounds.

Some medicines are dosed by weight: heavier animals need a larger dose that could be dangerous for smaller animals.

If someone from the shelter needs to give a dose of medicine to the animals, **is the "average" the best estimate we can use**?

Mean

“The average pet weighs almost 40 pounds” is a statement about the entire dataset, which summarizes a whole column of values with a single number.

Summarizing a big dataset means that some information gets lost, so it’s important to pick an appropriate summary. Picking the wrong summary can have serious implications!

Mean

- Students are sometimes summarized by two numbers -- their GPA and SAT scores -- which can impact where they go to college or how much financial aid they get.
- Schools are sometimes summarized by a few numbers -- student pass rates and attendance, for example -- which can determine whether or not a school gets funding.
- Adults are often summarized by a single number -- like their credit score -- which determines their ability to get a job or a home loan.
- When buying uniforms for a sports team, a coach might look for the most common size that the players wear.

Can you think of other examples where someone uses a number or two to summarize something complex?

Mean

The arithmetic mean is the number that "balances" all the other numbers in the sample. So let's do some real balancing!

Each group of three will need a ruler, 4-8 pennies, and at least one pen or pencil.

1. The ruler represents a number line with values (weight) distributed equally across the line. If there's values at every inch from 0 to 12, where should the pencil be placed in order to balance the ruler on top of it?
2. Place a penny at 1 and 11. Where must the pencil be placed to balance those two values? What is the mean of the values [1, 11]?
3. Place pennies at 1, 9 and 11. Where must the pencil be placed to balance those two values? What is the mean of of the values [1, 9, 11]?
4. Suppose you were to place two pennies at 2, and a third penny at 8. Can you *predict* where the pencil should be placed?

Mean

- If we plotted all the pounds values as points on a number line, what could we say about the **average** of those values?
- Is there a **midpoint**?
- Is there a point that shows up **most often**? Each of these are different ways of "measuring center".

Mean

The mean of a dataset is the sum of values divided by the number of values. To take the average of a column, we add all the numbers in that column and divide by the number of rows.

Pyret has a way for us to compute the mean of any quantitative column in a Table:

```
# mean :: Table, String -> Number
```

What is the function's name? Domain? Range?

Mean

- Open your saved Animals Starter File, or make a new copy.
- Type `mean(animals-table, "pounds")`.
- What does this give us?
- Does this support the Bureau's claims?
- Turn to Summarizing Columns in the Animals Dataset. In the "measures of center" section, *fill in the computed mean.*

Median

You computed the mean of that column to be almost exactly 40 pounds, but if we look at the dataset we'll quickly see that most of the animals weigh less than 40 pounds!

In fact, more than half of the animals weigh less than *15 pounds*. What is throwing off the average so much?

Median

Kujo and Mr. Peanutbutter!

In this case, the mean is being thrown off by a few extreme data points. These extreme points are called ***outliers***, because they fall far outside of the rest of the dataset.

Calculating the mean is great when all the points are fairly balanced on either side of the middle, but it distorts things for datasets with extreme outliers.

The mean may also be thrown off by the presence of ***skewness***: a lopsided shape due to values trailing off to the left or right.

Median

Make a `histogram` of the `pounds` column, and try different bin sizes. Can you see the huge number of animals clumped to the left, with Kujo and Mr. Peanutbutter as outliers skewed to the right?

A different way to measure center is to line up all of the data points -- in order -- and find a point in the center where half of the values are smaller and the other half are larger. This is the ***median***, or "middle" value of a list.

Median

Consider this list of ACT scores:

`25, 26, 28, 28, 28, 29, 29, 30, 30, 31, 32`

Here 29 is the median, because it separates the "bottom half" (5 values below it) from the top half" (5 values above it).

The algorithm for finding the median of a quantitative column is:

1. Sort the numbers
2. Cross out the highest and lowest numbers
3. Repeat until there is only one number left. If there are two numbers left at the end, take the *mean* of those numbers

Median

- Pyret has a function to compute the median of a list as well:

```
# median :: Table, String -> Number
```

- Compute the median for the `pounds` column in the Animals Dataset, and add this to Summarizing Columns in the Animals Dataset.
- Is it different than the mean?
- What can we conclude when the mean is so much greater than the median?
- For practice, compute the mean and median for the weeks and age columns.

Median

By looking at the histogram, we can see whether it's probably better to use the mean or median.

- Strong left skewness and/or low outliers can pull the mean down below the median.
- Right skewness and/or high outliers can pull the mean above the median.

Mean is generally the best measure of center, because it includes information from every single point. But it's misleading for highly-skewed datasets, so statisticians fall back to the median.

Modes

The third measure of center is called the ***modes*** of a dataset. The ***modes*** of a dataset are the values that appear *most often*.

Median and Mean always produce one number, but if two or more values are equally common, there can be more than one mode. If all values are equally common, then there is no mode at all!

Modes

Consider the following three datasets:

```
1, 2, 3, 4
1, 2, 2, 3, 4
1, 1, 2, 3, 4, 4
```

- The first dataset has no mode at all!
- The mode of the second dataset is 2, since 2 appears more than any other number.
- The modes (plural!) of the last dataset are 1 and 4, because 1 and 4 both appear more often than any other element, and because they appear equally often.

Modes

In Pyret, the mode(s) are calculated by the modes function, which consumes a Table and the name of the column you want to measure, and produces a *List* of Numbers.

```
# modes :: Table, String -> List
```

Modes

Compute the `modes` of the `pounds` column, and add it to Summarizing Columns in the Animals Dataset. What did you get?

Modes

The most common animal weights are 0.1 and 6.5! These are well below our mean and even our median, which is further evidence of outliers or skewness.

At this point, we have a lot of evidence that suggests the Bureau's use of "mean" to summarize animal weights isn't ideal.

Modes

We have three reasons to suspect that ***mean*** isn't the best value to use:

- The median is only 11.3 pounds.
- The modes of our dataset are only 0.1 and 6.5 pounds, which suggests clusters of animals that weigh mere fractions of the mean.
- When viewed as a histogram, we can see the right skewness and high outliers in the dataset. Mean is sensitive to datasets with skewness and/or outliers.

Modes

“In 2003, the average American family earned $43,000 a year -- well above the poverty line! Therefore very few Americans were living in poverty."

Do you trust this statement? Why or why not?

Consider how many policies or laws are informed by statistics like this!

Knowing about measures of center helps us see through misleading statements.

Modes

When should each measure of center be used?

- If the data doesn't show much skewness or have outliers, ***mean*** is the best summary because it incorporates information from every value.
- If the data has noticeable outliers or skewness, ***median*** gives a better summary of center than the mean.
- If there are very few possible values, such as AP Scores (1–5), the ***mode*** could be a useful way to summarize the dataset.

Data Exploration Project (Measures of Center)

Let's review what we have learned about computing and interpreting three measures of center - mean, median, and modes.

- Describe how to compute mean, median, and modes.
- When ***mean*** provide the best summary?
- When does ***median*** provide the best summary?
- When are ***mode***(s) a useful way to summarize a dataset?

Data Exploration Project (Measures of Center)

Let's connect what we know about measures of center to your chosen dataset.

- Open your chosen dataset starter file in Pyret.
- Choose two quantitative columns that you'd like to analyze.
- Use Pyret to compute the mean, median and modes of one of them.
- *It's time to add to your Data Exploration Project.*
- Locate the "Measures of Center and Spread" section of your Exploration Project and, in the slide following the example, replace `Column A` with the title of the column you just investigated.

Data Exploration Project (Measures of Center)

- Then type in the mean, median and modes that you just identified. Leave the other rows blank. We will come back to them another day.
- On the next slide, repeat with `Column B` using the second column you're interested in.

Data Exploration Project (Measures of Center)

Add your interpretations to the two “Measures of Center and Spread” slides and record any questions that emerged in the "My Questions" section at the end of the slide deck.

Data Exploration Project (Measures of Center)

Share your findings!

Did you discover anything surprising or interesting about your dataset?

Which measures of center do you think were the most useful for their two chosen quantitative columns?

What questions did the measures of center inspire you to ask about their dataset?

When you compared their findings with other students, did they make any interesting discoveries?

Additional Activities

- Mode(s) (Desmos)
- Critiquing Written Findings
- Data Cycle: Measures of Center