










Measures of Center

Students learn different ways to report the center of a quantitative data set: mean, median and mode(s). After applying these concepts to a contrived dataset, they apply them to their own datasets and interpret the results.

Prerequisites	Visualizing the “Shape” of Data																		
Relevant Standards	Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).																		
<div>OK</div> <div>K12CS</div> <div>NGSS</div>																			
Lesson Goals	Students will be able to... <ul style="list-style-type: none">Students explore the concept of center of a distribution, learning how to compute the mean, median and mode(s) of a datasetStudents find the mean, median and mode(s) of various columns in the Animals table																		
Student-facing Lesson Goals	<ul style="list-style-type: none">Let’s use mean, median, and mode to describe our data.																		
Materials	<ul style="list-style-type: none">Lesson Slides (Google Slides)Computer for each student (or pair), with access to the internetStudent workbook, and something to write with																		
Preparation	<ul style="list-style-type: none">Make sure all materials have been gatheredDecide how students will be grouped in pairsAll students should log into CPO and open the "Animals Starter File" they saved from the prior lesson. If they don’t have the file, they can open a new one																		
Supplemental Resources																			
Language Table	<table><tr><th>Types</th><th>Functions</th><th>Values</th></tr><tr><td>Number</td><td>num-sqrt, num-sqr</td><td>4, -1.2, 2/3</td></tr><tr><td>String</td><td>string-repeat, string-contains</td><td>"hello", "91"</td></tr><tr><td>Boolean</td><td>==, <, <=, >=, string-equal</td><td>true, false</td></tr><tr><td>Image</td><td>triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram</td><td>  </td></tr><tr><td>Table</td><td>count, .row-n, .order-by, .filter, .build-column</td><td></td></tr></table>	Types	Functions	Values	Number	num-sqrt, num-sqr	4, -1.2, 2/3	String	string-repeat, string-contains	"hello", "91"	Boolean	==, <, <=, >=, string-equal	true, false	Image	triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram	  	Table	count, .row-n, .order-by, .filter, .build-column	
Types	Functions	Values																	
Number	num-sqrt, num-sqr	4, -1.2, 2/3																	
String	string-repeat, string-contains	"hello", "91"																	
Boolean	==, <, <=, >=, string-equal	true, false																	
Image	triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram	  																	
Table	count, .row-n, .order-by, .filter, .build-column																		

Glossary

mean :: average, calculated as the sum of values divided by the number of values

median :: the middle element of a quantitative data set

mode :: the most commonly appearing categorical or quantitative value or values in a data set

outlier :: a data point that is unusually far above or below most of the others

skew :: lack of balance in a dataset’s shape, arising from more values that are unusually low or high. Such values tend to trail off, rather than be separated by a gap (as with outliers).

Do Now

Open your workbooks to [Summarizing Columns in the Animals Dataset \(Page 61\)](#). We've already filled in the answer to Question 1 for you (`pounds`). In your animals starter file, make a `box-plot` for the `pounds` column, and fill in the 5-number summary for that column in your workbook.

Mean

15 minutes

Overview

Students learn about mean (or "average"), and how it is one way (among others!) to summarize a quantitative column.

Launch

According to the Animal Shelter Bureau, the average pet weighs almost 41 pounds.

Some medicines are dosed by weight: heavier animals need a larger dose. If someone from the shelter needs to give a dose of medicine to the animals, is the "average" the best estimate we can use?

"The average pet weighs 41 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! Here are just a few examples of summary data being used for important things. Do you think these summaries are appropriate or not?

- 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 shut down.
- 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?

Every kind of summary has situations in which it does a good job of reporting what's typical, and others where it doesn't really do justice to the data. In fact, the shape of the data can play a huge role in whether or not one kind of summary is appropriate!

One of the ways that Data Scientists summarize quantitative data is by talking about its *center* - literally asking "what is a typical value in this sample?", in the hopes of inferring something about a larger population. But there are many different ways to define "center", and each method has strengths and weaknesses. Let's check the "41 pounds" claim and see if it's an appropriate measure of center. Later on, you'll have a chance to apply what you've learned to your own dataset, to find the best way to provide an overall summary of the data.

Investigate

Open your "Animals Starter File". (If you do not have this file, or if something has happened to it, you can always make a [new copy](#).)

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".

The Animal Shelter Bureau used one method of summary, called the *mean*, or "average". In general, the mean of a data set 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. It consumes a Table and the name of the column you want to measure, and produces the mean — or average — of the numbers in that column.

```
# mean :: (t :: Table, col :: String) -> Number
```

What is its name? Domain? Range?

Notice that calculating the mean requires being able to add and divide, so the mean only makes sense for quantitative data. For example, the mean of a list of Presidents doesn't make sense. Same thing for a list of zip codes: even though we can divide a sum of zip codes, the output doesn't correspond to some "center" zip code.

Type `mean(animals-table, "pounds")`. What does this give us? Does this support the Bureau's claims?

Open your workbooks to [Summarizing Columns in the Animals Dataset \(Page 61\)](#). Under the "measures of center" section, fill in the computed mean.

Median

15 minutes

Overview

Students learn a second measure of center: the *median*. They learn the algorithm and the code to find the median, as well as situations where taking the median is more appropriate than the mean.

Launch

You computed the mean of that column to be almost exactly 41 pounds. That IS the average, but if we scan the dataset we'll quickly see that most of the animals weigh less than 41 pounds! In fact, more than half of the animals weigh less than just 15 pounds. What is throwing off the average so much?

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 left or right of center.

Make a `histogram` of the `pounds` column, and try different bin sizes. Can you see the skew towards the right, with a huge number of animals clumped to the left?

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.

As an example, 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 (we did this for you in the above example).
2. Cross out the highest number.
3. Cross out the lowest number.
4. Repeat until there is only one number left. If there are two numbers left at the end, take the *mean* of those numbers.

Investigate

- Pyret has a function to compute the median of a list as well. Find the contract in your contracts page.
- Compute the median for the `pounds` column in the Animals Dataset, and add this to [Summarizing Columns in the Animals Dataset \(Page 61\)](#).
- 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.

Synthesize

By looking at the histogram, we can develop an intuition for whether it's probably better to use the mean or median. Pronounced left skewness and/or low outliers can pull the mean down below the median, while right skewness and/or high outliers can pull it up. Either way, such shapes distort the mean as a measure of what's typical for the data set. Data scientists generally prefer to use the mean as their measure of center, because it contains information from every single data value. However, if a data set has substantial skewness or outliers, they use median to report the center .

Modes

25 minutes

Overview

Students learn about the mode(s) of a dataset, how to compute the mode, and when it is appropriate to use this as a measure of center.

Launch

The third measure of center is called the *mode* of a dataset. The *mode* of a data set is the value that appears *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! Often there will be just one mode in the list of most common values: many data sets are what we call “unimodal”. But sometimes there are exceptions! 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 data set is 2, since 2 appears more than any other number.
- The modes (plural!) of the last data set are 1 and 4, because 1 and 4 both appear more often than any other element, and because they appear equally often.

Mode is rarely used to summarize quantitative data. It is very common as a summary of *categorical* data, telling us which category occurs most often.

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 :: (t :: Table, col :: String) -> List<Number>
```

Investigate

Compute the `modes` of the `pounds` column, and add it to [Summarizing Columns in the Animals Dataset \(Page 61\)](#). What did you get?

Synthesize

The most common number of pounds an animal weighs is 6.5! That's 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. Our mean weight agrees with their findings, but we have three reasons to suspect that *mean* isn't the best value to use:

- The median is only 13.4 pounds.
- The mode of our dataset is only 6.5 pounds, which suggests a cluster of animals that weigh less than one-sixth 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.

Closing

The Animal Shelter Bureau started with a fact: the mean weight is about 41 pounds. But then they reported a conclusion without checking to see if that was the best summary statistic to look at. As Data Scientists, we had to look deeper into the data to find out whether or not to settle for the Bureau's summary. This is why using tools like histograms that show shape can be so important when deciding on a summary tool.

"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.

You now have three different ways to measure center in a dataset. But how do you know which one to use? Depending on the shape of the dataset, a measure could be really useful or totally misleading! Here are some guidelines for when to use one measurement over the other:

- 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 data set.

Exercises

Critiquing Written Findings