Spread of a Data Set

Students learn how to evaluate the spread of a quantitative column using box plots, and explore how this offers a different perspective on shape from what can be achieved with a histogram. After applying these concepts to a contrived dataset, they apply them to their own datasets and interpret the results.

Prerequisites	Measures of Center
Relevant Standards K12CS CSTA CC-Math NGSS	*Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).*
Lesson Goals	Students will be able to... • apply one approach to measuring and displaying spread of a data set • compare and contrast information displayed in a box plot and a histogram
Student-facing Lesson Goals	• Let's compare different uses for box plots and histograms when talking about data.
Materials	• Lesson Slides (Google Slides) • Computer for each student (or pair), with access to the internet • Student workbook, and something to write with
Preparation	• Make sure all materials have been gathered • Decide how students will be grouped in pairs • All 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	

Types	Functions	Values
Number	`num-sqrt, num-sqr, mean, median, modes`	`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

box plot :: the box plot (a.k.a. box-and whisker-plot) is a way of displaying a distribution of data based on the five-number summary: minimum, first quartile, median, third quartile, and maximum

interquartile range :: (IQR) is one possible measure of spread, based on dividing a data set into four parts. The values that divide each part are called the first quartile (Q1), the median, and third quartile (Q3). IQR is calculated as Q3 minus Q1.

median :: the middle element of a quantitative data set

quartiles :: three values that divide a data set into four equal-sized groups

range of a data set :: the distance between minimum and maximum values

shape :: The aspect of a dataset that tells which values are more or less common
spread :: the extent to which values in a data set vary, either from one another or from the center

---

Measures of Spread

30 minutes

Overview

Students are introduced to the notion of *spread* in a dataset. They learn about quartiles, box plots, and how to use them to talk about spread.

Launch

A teacher may report that her students averaged a 75 on a test, but it's important to know how those scores were spread out: did all of them get exactly 75, or did half score 100 and the other half 50? When Data Scientists use the mean of a sample to estimate the mean of a whole population, it's important to know the spread in order to report how good or bad a job that estimate does.

Suppose we lined up all of the values in the pounds column of the animals data set from smallest to largest, and then split the line up into two equal groups by taking the ***median***. We can learn something about the ***spread*** of the data set by taking things further: The middle of the lighter half of animals is called the first ***quartile*** - or "Q1" - and the middle of the heavier half of animals is the third quartile (also called "Q3"). Once we find these numbers, we can say that the middle half of the animals' weights are spread between Q1 and Q3.

The first quartile (Q1) is the value for which 25% of the animals weighed that amount or less. What does the third quartile represent?

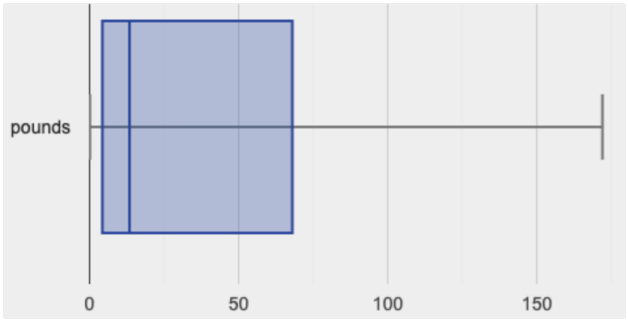
Besides looking at the median as center, and the spread between Q1 and Q3, we also gain valuable information from the spread of the entire data set—that is, the distance between minimum and maximum. This is called the ***range of a data set***. We can use ***box plots*** to visualize all of this information. These plots are constructed using **just five numbers**, which makes them convenient ways to display both center and spread of a data set in a clear and simple way. Below is the contract for `box-plot`, along with an example that will make a box plot for the `pounds` column in the `animals-table`.

```
# box-plot :: (t :: Table, column :: String) -> Image
box-plot(animals-table, "pounds")
```

Box plots divide our sample into equally-sized groups, and show where those groups are spread thin or clumped together.

Type in this expression in the Interactions Area, and see the resulting plot.

This plot shows us the center and spread in our dataset according to those five numbers.



- The **minimum** value in the dataset (at the left of "whisker"). In our dataset, that's just 0.1 pounds.
- The **First Quartile (Q1)** (the left edge of the box), is computed by taking *the median of the lower half of the values*. In the pounds column, that's 4.3 pounds.
- The **Median** value (the line in the middle), which is the middle Quartile of the whole dataset. We already computed this to be 13.4 pounds.
- The **Third Quartile (Q3)** (the right edge of the box), which is computed by taking *the median of the upper half of the values* That's 68 pounds in our dataset.
- The **maximum** value in the dataset (at the right of the "whisker"). In our dataset, that's 172 pounds.

One way to summarize the spread in the dataset is to measure the distance between the largest value and the smallest value. When we talk about functions having many possible outputs, we use the term "Range" to describe them. (**Note**: the

term "Range" means something different in statistics than it does in algebra and programming!) When we look at the distance between the smallest and largest values in our dataset, we use the same term.

Extension Activity

In statistics, it is not uncommon to use *modified box plots*, which remove extreme datapoints from the box-and-whisker and draw them as dots outside of the blot. The box plot then represents only the "non-extreme" points. Modified box plots are also available in Bootstrap:Data Science, using the following contract:

```
# modified-box-plot :: (t :: Table, column :: String) -> Image
```

Investigate

- Turn to Summarizing Columns in the Animals Dataset (Page 61)
- Fill in the five-number summary for the `pounds` column, and sketch the box plot.
- What conclusions can you draw about the distribution of values in this column?

Data Scientists subtract the 1st quartile from the 3rd quartile to compute the range of the “middle half” of the dataset, also called the ***interquartile range***.

- Find the ***interquartile range*** of this dataset.
- What percentage of animals fall within the interquartile range?
- What percentage of animals fall below the First Quartile? Above the Third Quartile? What percentage fall anywhere between the minimum and the maximum?

Now that you’re comfortable creating box plots and looking at measures of spread on the computer, it’s time to put your skills to the test!

Turn to Interpreting Spread (Page 62) and complete the questions you see there.

Just as pie and bar charts are ways of visualizing categorical data, box plots and histograms are both ways of visualizing the shape of quantitative data. Box plots make it easy to see the 5-number summary, and compare the Range and Interquartile Range. Histograms make it easier to see skewness and more details of the shape, and offer more granularity when using smaller bins.
Left-skewness is seen as a long tail in a histogram. In a box plot, it’s seen as a longer left "whisker" or more spread in the left part of the box. Likewise, right skewness is shown as a longer right "whisker" or more spread in the right part of the box. Box plots and Histograms can both tell us a lot about the shape of a dataset, but they do so by grouping data quite differently. A box plot is always divided into four parts, which may fall on differently-sized intervals but all contain the same number of points. A histogram, on the other hand, has identically-sized intervals which can contain very different numbers of points.

Turn to Identifying Shape (Page 63) and see if you can describe box plots using what you know about skewness.

Challenge Questions: - Compare the for the `pounds` column of both cats and dogs in the dataset. Are their shapes different? How much overlap is there? - Compare histograms for the `age` column of both cats and dogs in the dataset. Are their shapes different? How much overlap is there? - Can you explain why the amount of overlap between these two distributions is different?

Possible Misconceptions

It is extremely common for students to forget that every quartile *always* includes 25% of the dataset. This will need to be heavily reinforced.

Synthesize

Histograms, box plots, and measures of center and spread are all different ways to get at the ***shape*** of our data. It’s important to get comfortable using every tool in the toolbox when discussing shape!

Modified Box Plots

Modified Box Plots

More Statistics- or Math-oriented classes will also be familiar with *modified box plots*. These are similar to traditional box plots, but the box-and-whisker just extends to minimum and maximum non-outliers. To call our attention to outliers, they are drawn as small dots or asterisks at the extreme ends of the graph (watch a video on modified box plots). Pyret also has a `modified-box-plot` function, with the same Domain as `box-plot`.

Comparing Box Plots

15 minutes

Overview

Students assess the degree of visual overlap of two numerical distributions.

Launch

Mutiple box plots are extremely useful for showing us the answer to a particular kind of **Relate Question**, such as "Do dogs take longer to get adopted than cats?" This is really asking us about the interplay between a categorical variable (species) and a quantitative one (weeks to adoption). Instead of creating a whole new display tool, all we have to do is extend our usual box plot display so we can look at how the weeks distributions compare for cats and dogs. This works fine as long as we're sure to use a common scale: Note that both box plots in the display below share the same axis for adoption times, which ranges from about 1 to 10 weeks.

Box plots make it easy to decide if values of a quantitative variable seem to be fairly similar or quite different, depending on which group an individual is in. The trick is to train your eyes to look for whether there's a lot of overlap in the two box plots, or if one is noticeably higher than the other.

Investigate

Have students break into groups of 3-4, and compare the box plot of weeks-to-adoption for cats with the one for dogs.

Note: they can generate the pair of box plots themselves, but we recommend simply giving them this image: cats v. dogs

1. Do the two box plots mainly overlap, or is one noticeably higher than the other?
2. Roughly how do the medians compare?

Next, each group examines the pair of box plots that compare weeks to adoption for fixed versus unfixed animals: fixed v. unfixed. Once again, consider how similar or different the two plots seem.

1. Do the two box plots mainly overlap, or is one noticeably higher than the other?
2. Roughly how do the medians compare?

Students should confirm that the box plots for adoption times of unfixed versus fixed animals have more overlap than the box plots for adoption times of cats versus dogs.

Box plots and histograms give us two different views on the concept of shape. In a histogram, the intervals between the bars are fixed with different numbers of datapoints in each interval. A box plot is the exact opposite: the intervals are variable, with a fixed number of datapoints in each one. In a histogram, we can think of a datapoints that fall into bins, filling them up so we can see how many are in each. A box plot treats the data more like pizza dough, dividing it into four equal quarters and squeezing or stretching it to show where the data is tightly clumped or spread out over a long interval. To compare the two, complete Matching Box-Plots to Histograms (Page 65).

Synthesize

Referring to our first side-by-side box plots, the one for dogs' adoption times was much higher than the one for cats' adoption times; the top half of the dogs' box plot doesn't overlap at all with the cats' box plot. Does this suggest that species *does* or *does not* play a role in how long it takes for an animal to be adopted?

Referring to our second pair of box plots, we saw that adoption times for unfixed and fixed animals overlapped a lot, and the

medians (shown by the lines through the middle of each box) were pretty close: both a bit less than 4. Does this suggest that being fixed or not does or does not play a role in how long it takes for an animal to be adopted?
Which variable seems to have more of an effect on adoption time: species (cat or dog) or whether an animal is fixed or not?
Have students share back their findings.

Your Analysis

flexible

Overview

Students repeat the previous activity, this time applying it to their own dataset and interpreting their own results. **Note: this activity can be done briefly as a homework assignment, but we recommend giving students an *additional class period* to work on this.**

Investigate

- Take 15 minutes to fill out Shape of My Dataset (Page 64) in your Student Workbook. Choose a column to investigate, and write up your findings.
- Students should fill in Measures of Center and Spread portion of their Research Paper, using the means, medians, modes, box plots and five-number summaries they've constructed for their dataset and explaining what they show.

Synthesize

Have students share their findings with one another.

Additional Exercises:

- Project: Project: Stress or Chill? (You will also need the Personality True Colors assessment)