Standard Deviation

(Also available in Pyret)

Students learn how standard deviation serves as Data Scientists' most common measure of "spread": how far all the values in a dataset tend to be from their mean. When we looked at box plots, we visualized spread based on range and interquartile range. Now we'll return to histograms and picture the spread in terms of standard deviation.

Lesson Goals	Students will be able to… • apply one approach to measuring and displaying spread of a dataset • 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.
Prerequisites	• Introduction to Data Science • Exploring CODAP • Dot Plots and Bar Charts • Histograms • Visualizing the "Shape" of Data • Measures of Center • Box Plots
Materials	• PDF of all Handouts and Page • Data Exploration Project Slide Template • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• Additional Printable Pages for Scaffolding and Practice

Glossary

histogram :: a display of quantitative data that uses vertical bars positioned over bins (or 'intervals'); each bar's height reflects the count data values in that bin.

mean :: a representation of the center, or 'typical' value in a set of numbers, calculated as the sum of those numbers divided by the number of values.

outlier :: observations whose values are very different from the other observations in the same dataset, perhaps due to experimental error. Outliers can also be indicative of data belonging to a different population from the rest of the established samples.

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).

spread :: the extent to which values in a dataset vary, either from one another or from the center

standard deviation :: a number that measures spread of a dataset using the typical distance of values from their mean

Measuring "Deviance" *30 minutes*

Overview

Students review the notion of ***spread*** itself, and build up to the formula by annotating ***histograms***.

Launch

The Animal Shelter Bureau reports that the *mean* age of shelter cats is 3 years.

Take a look at the the Animals Dataset on the spreadsheet or on this page (for those using a printed workbook, you'll find it at the front).

- Does a mean age of 3 years translate to all of the cats being close to 3 years old? Why or why not?
  - *No, we cannot assume all cats are close to 3 years old. There are some outliers in the dataset.*

In the activity that follows, students will look at ten cats from the shelter to consider the distribution of their ages.

Turn to Computing Standard Deviation, and complete numbers 1-3.

- What did you get for the mean? Does it match what the Animal Shelter Bureau says?
  - *The mean is 3; yes, it matches what the Animal Shelter Bureau says.*
- Can you think of four ages, such that the mean age for all of them is 3?
  - *Some possibilities include: {3,3,3,3}, {1,2,4,5}, {1,1,4,6}… any four ages that add up to 12 will work!*
- Can you think of a *different* spread of four ages that would have the same mean?
  - *See above.*
- How many different sets of four ages can you think of, which all have a mean of 3?
  - *See above.*

Without a measure of ***spread***, just knowing the mean doesn't tell us enough about the shape of the data.

When summarizing a column, we'd like to use a measure that gathers data from every value. We already have one method of measuring spread: calculating the Five Number Summary and using it to generate a box-plot.

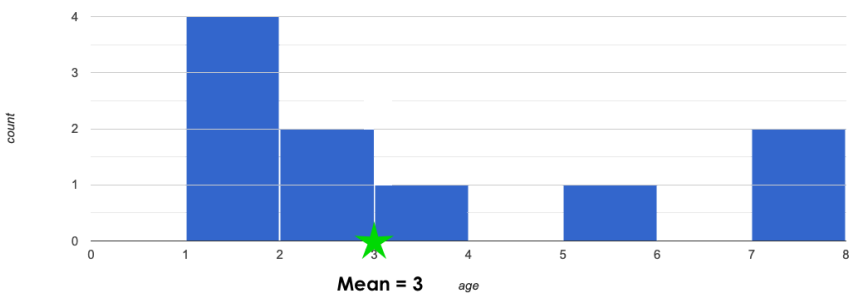
Unfortunately, that measure comes from only a small number of data points! If possible, we'd like to have a measure that summarizes the spread across *all* the points.

Standard deviation is the most useful way to summarize ***spread*** of a quantitative column.

Instead of focusing on the handful of data points used in our Five Number Summary, another way to measure spread is to focus on *the "typical" distance from the mean*. In other words, we want to know what kind of deviation is "standard" for all the points.

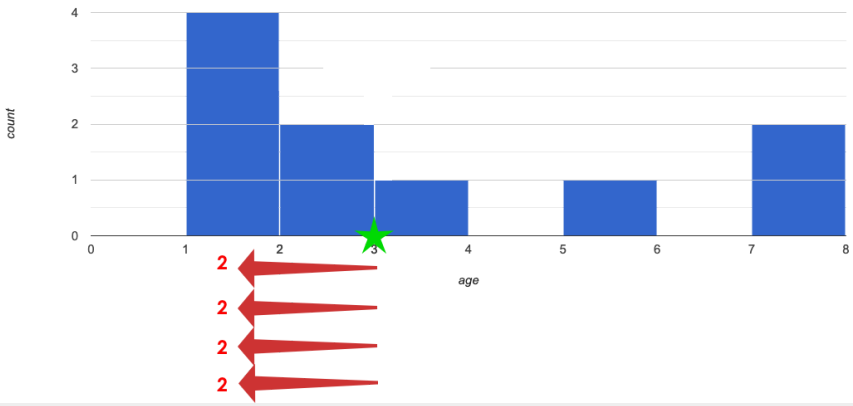
Investigate

We could imagine a shelter where every cat is between 2 and 4, so **each cat only deviates from the mean by 1 year**! But we could also imagine a shelter with only kittens and very old cats, where **cats deviate by as much as 10 years from the mean!**



How far away is each data point from 3?

In this image, we've draw an arrow for each of the 1-year-old cats. That means there are four arrows running from the mean at 3 to the interval at 1, and each arrow has the label 2.



Next, complete numbers 4 to 6 of Computing Standard Deviation.

Mean Average Deviation?

In this section of the worksheet, students will need to stretch their visual imaginations a bit! In problem number 6, they are asked to summarize all 10 distances from the mean into a single number. The goal here is for students to make an educated guess about standard deviation (SD) *before* learning the algorithm for computing it. Invite and encourage discussion about students' different approaches for guessing at the best summary number *before* sharing the key idea about standard deviation!

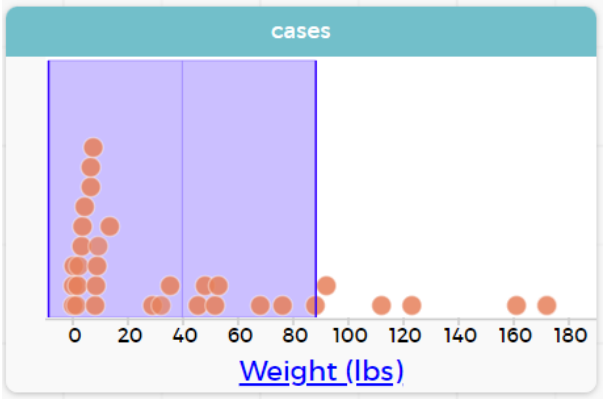
Students are likely to hone in on the *Mean Average Deviation*, or MAD. Both SD and MAD measure variability or "spread" by computing individual deviations from the mean, but MAD averages these deviations and SD transforms them via square/square-root.

To compute the standard deviation we add the squares of all *N* distances, divide by *N-1*, then take the square root of the result.

The process of finding standard deviation manually is a bit laborious. Keeping organized is crucial; a partially-completed table is provided on the bottom half of worksheet to support students in doing so.

Complete numbers 7-10 of Computing Standard Deviation, where you will utilize the algorithm for computing standard deviation.

To compute standard deviation in CODAP, create a graph with only one quantitative attribute. Open the `Measure` menu, then select the button that says "Measures of Spread." (Note that this button only appears when one quantitative attribute displayed.) Selected Standard Deviation. Move your cursor back to the display, and hover over the edge of the purple shading that appears.



- What is the standard deviation for the weights of *all* the animals at our dataset?
  - *Approximately 48.5*

Optional: For additional practice, have students complete Computing Standard Deviation (2).

Synthesize

- Can you explain why two datasets can have the same mean, but different standard deviations?
  - *Mean is a measure of* ***central tendency****, whereas standard deviation measures the* ***variation*** *of some sample.*
- What kind of dataset would have a standard deviation of zero?
  - *A standard deviation of zero means that every number in the sample is exactly the same.*

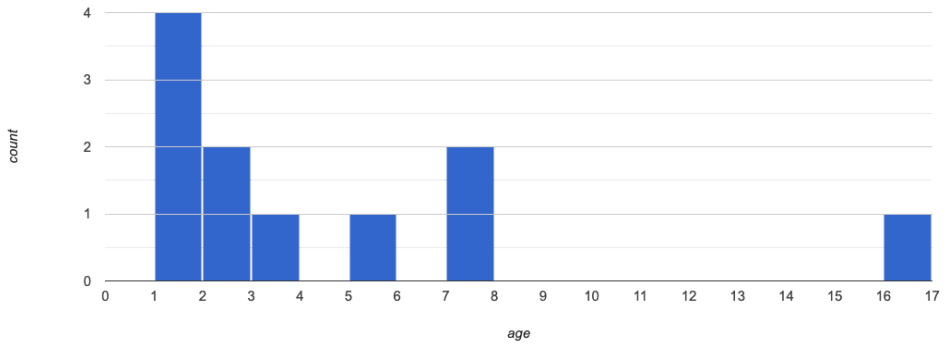
Comparing Standard Deviations *20 minutes*

Overview

Students compare centers and (more importantly) spreads - of two quantitative datasets by comparing their histograms. Both ***mean*** and ***standard deviation*** can be affected by ***outliers*** and/or ***skewness***.

Launch

Take a look at the histogram below. It is the same histogram we saw in the previous section, but now with an 11th cat that is 16 years old. That's quite an outlier!



- What is the shape of this histogram?
  - *The histogram has high outliers, therefore it is skewed right.*
- How does it differ from the one we just looked at?
- The previous histogram - with the 16-year-old cat omitted - was roughly symmetric.

Turn to The Effect of an Outlier to explore the extent to which the inclusion of an outlier will affect the center and spread of a quantitative dataset.

- What did this outlier do to the mean? Refer back to Computing Standard Deviation to help you.
  - *Previously, the standard deviation was ~2.45; now it is ~5.83.*
- What did this outlier do to the standard deviation?
  - *The outlier caused the standard deviation to increase by ~3.38.*

Optional: To see how changes in data values affect the mean and standard deviation, complete Matching Mean & Standard Deviation to Data.

Investigate

The mean and standard deviation tell us where the data is centered and how far the data strays from that center. For example, when writing about the ages of cats in our shelter, we might say "the mean age is 3 and the standard devation is 2.45, so most cats are between the ages of 1 and 5 years old."

- The mean time-to-adoption is 5.75 weeks. Does that mean most animals generally get adopted in 4-6 weeks? *Solicit students' ideas, but do not reveal the answer.*
- Turn to Data Cycle: Standard Deviation in the Animals Dataset to get some practice using the Data Cycle to answer this question, then write your findings in the space at the bottom.

Mean Average v. Standard Deviation

MAD and SD are both measures of a certain kind of *distance*, literally asking "how are far from the mean are all the points in the dataset?". With each point being independent from the other, we can imagine a dataset with two points as a right triangle with two legs: how far apart are these points?

Before learning the distance formula, students might guess at a number of ways to compute the hypotenuse. They can quickly rule out the sum of the legs, and the difference between them. At some point they might suggest *averaging* the lengths of the legs. Mean Average Deviation (MAD) does exactly that, by flattening each points' deviation into a single "dimension".

Of course, these legs exist on separate axes - so we need a formula for distances in more than one dimension. Computing the SD involves the *square root of a sum of squares*. That should sound suspiciously like the distance formula! Indeed, computing the SD for a dataset with two points is basically finding the (normalized) length of the hypoteneuse!

The pythagorean distance works in 3-dimensions as well (right pyramids!) - or for any number of dimensions - as does the formula for standard deviation. By treating each point as a separate dimension, DS allows each deviation to be considered independantly.

Why use one measure of spread instead the other? The answer is closely related to the difference between two measures of *center*! `Mean` incorporates data from every point, while `median` does not. However, `mean` is sensitive to the effect of extreme outliers or ***skew***. In those cases, `median` is considered to be the better measure of center.

Treating each point independantly allows each deviation to contribute to the measure of spread, just as `mean` computes the measure of center. This is why SD is used most often, but like `mean` it is sensitive to extreme outliers or skew. In those cases, the MAD is considered a better measure of spread.

Synthesize

- How much did adding an outlier change the mean? The standard deviation?
- Extreme values affect both the mean and standard deviation of a dataset.
- Unusually low values *decrease* the mean, while unusually high values *increase* it. Unusually low or high values increase the standard deviation, because it summarizes distance from the mean in either direction.

Data Exploration Project (Standard Deviation) *flexible*

Overview

Students apply what they have learned about standard deviation to their chosen dataset. In their Data Exploration Project Slide Template, they will complete the final row of the "Measures of Center and Spread" table, adding the standard deviation for two quantitative columns. They will also interpret the standard deviations they found, and record any interesting questions that emerge. To learn more about the sequence and scope of the Exploration Project, visit Project: Dataset Exploration. For teachers with time and interest, Project: Create a Research Project is an extension of the Dataset Exploration, where students select a single question to investigate via data analysis.

Launch

Let's review what we have learned about standard deviation.

- Do we compute standard deviation with categorical data or quantitative data? How many columns of data does standard deviation tell us about?
  - *Standard deviation is a measure that tells us about the spread of a single quantitative column of data.*
- Standard deviation is a measure of ***spread***. In your own words, what does ***spread*** mean?
  - *Spread is the extent to which values in a dataset vary, either from one another or from the center.*
- How can two datasets have the same mean, but different standard deviations?
  - *Mean is a measure of central tendency, whereas standard deviation measures the variation of some sample.*
- Both unusually low and unusually high values (outliers) **increase** the standard deviation. Explain why.
  - *Standard deviation summarizes distance from the mean in **either** direction.*

Investigate

Let's connect what we know about standard deviation to your chosen dataset.

Reminder: Students have the opportunity to choose a dataset that interests them from our List of Datasets in the Choosing Your Dataset lesson.

- Open your chosen dataset starter file in CODAP.

- Remind yourself which two columns you investigated in the Measures of Center lesson and use CODAP to compute the standard deviation for one of them.
    - *Note: Consider recommending that students choose the same column they used when they found their Measures of Center. If students use a different column, they will need to copy/paste additional slides into their slide deck.*
- What question does your computation answer?
    - *Possible responses: How is the data for a certain column distributed? How does the standard deviation compare to the mean?*
- Write down that question in the top section of the Data Cycle: Standard Deviation in My Dataset.
- Complete the rest of the data cycle, recording how you considered, analyzed and interpreted the question.
- Repeat this process for the other column you explored before (and any others you are curious about).
    - *Note: If students want to investigate new columns from their dataset, they will need to copy/paste additional Measures of Center and Spread slides into their Explorartion Project and calculate the mean, median, modes and 5-number summaries for the new columns.*

Invite students to discuss their results and consider how to interpret them.

- **It's time to add to your Data Exploration Project Slide Template.**
- Locate the "Measures of Center and Spread" section of your Exploration Project. Type the standard deviations you just identified into the tables for the corresponding columns.
- Now, add your interpretations of the standard deviations and record any questions that emerged in the "My Questions" section at the end of the slide deck.

Synthesize

Share your findings with the class!

Did you discover anything surprising or interesting about your dataset?

What questions did the standard deviations inspire you to ask?

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