Measuring the Center of a Dataset

Averages measure the "center" of datasets, the properties that datasets, when graphed, have in their middle portions.

Mean: the "mathematical average."

$$\mu = \Sigma(v_a)/n$$

v_a: val of datapoint *a*

n : number of datapoints in set

or

$$\mu = \Sigma(v_a \times f_a)$$

v_a: val of datapoint(s) of kind *a*

f_a: % datapoint(s) of kind *a* are of the entire set of datapoints (as decimal)

Median: middle datapoint/value (when odd), or the value between the two middle datapoints/values (when even number of datapoints).

- ↳ When there are an even number of datapoints, there is no middle datapoint/value. So, to find the median, take the mean of the middle two values.

Mode: most commonly occurring datapoint/value.

Measuring Spread of a Dataset

There is sometimes an equivocation on the term "spread." Spread can simply be synonymous with **range** (the highest and lowest values in a dataset), or can mean **clustering** (how the datapoints and their values are grouped together).

Range = (highest datapoint of the set) - (lowest datapoint of the set)

$$\textbf{range} = v_n - v_1$$

Interquartile Range (IQR) = the range of the middle 50% of datapoints.

IQR = (highest value of the inner 50% of datapoints) - (lowest value of the inner 50% of datapoints)

i.e.

IQR = 3rd Quartile - 1st Quartile = Q3- Q1

- ↳ **Quartiles:** Quartiles are the dividing lines between evenly spaced, 25% intervals. Each interval is divided at the median/midpoint. Q1 is the dividing line between the lowest 25% of datapoints and the next 25%. Q3 is the dividing line between the highest 25% of datapoints and the previous 25%. That is, Q1 is the dividing line between the interval of the lowest 25% of datapoints and the middle 50%, and Q3 is the dividing line between the interval of the highest 25% of datapoints and the middle 50%.

↳ **Finding Quartiles:** Order data from least to greatest. Find the median (midpoint)--this is your middle quartile (cutoff value) (Q2). Split the data into two sets along the median. Find the median of each set--these medians are the cutoffs of each quarter interval (Q1 and Q3). Remember that the quartiles are the cutoff values/points, *not* the highest or lowest members of each interval.

Outlier Rule for Quartiles:
Any points that are more than 1.5 × IQR above Q3 or below Q1 are considered outliers.

Five Number Summary:

Min	
Q1	
Q2 (Median)	
Q3	
Max	

What Range and IQR can Show:
Range (highest-lowest) simply shows how far the spread of the data *can be*. That is, it shows how far apart the highest and lowest extremes of the entire dataset are. It *cannot* indicate how the datapoints cluster together. For two very different datasets--one with tight clusters and one with loose clusters of data--can have the same IQR.

Since IQR is just the range of the middle 50% of a dataset, IQR is limited in the same way: it shows how far apart the highest and lowest extremes *of the middle 50% of data* are. It shows how far the spread of the middle 50% of data points *can be*. But, again, it cannot indicate how the datapoints cluster together. For two very different datasets--one with tight clusters and one with loose clusters of data--can have the same IQR.

Proof: These two datasets have the same IQR. Yet, they have very different clustering, and therefore graphs. This proof will work both for range and IQR, since IQR is just a kind of range.

$D_1 = \{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20\}$
$D_2 = \{0,1,2,3,4,5,6,9,9,9,10,11,11,11,12,15,16,17,18,19,20\}$

Comparative Use 1: If D1 has higher IQR than D2, then the middle 50% of the data points *can* spread out a greater distance from the median of the dataset in D1 than in D2. (Again, keep in

mind that this does not show how the data clusters, only the limits of how it *can* cluster. It is a very vague measure.)

Comparative Use 2: When you know the range and IQR of a dataset, you can estimate how much the bottom and top 25% intervals spread apart from the middle 50%. The greater the difference between range and IQR, the more the low and high 25% intervals can deviate from the middle 50%. This also indicates (is an indirect measure of) how far the low and high 25% intervals deviate from the median. (But, again, keep in mind that this cannot show you how the data clusters--it can only show you the limits of how far the data can cluster.)

> **Principles of Use:** Thus, if D1 has a greater difference between range and IQR than D2, then the lows and highs of D1 *can* spread out much farther (proportionally, relative to the median and middle 50% of each, not absolutely) than the lows and highs of D2. And, if D1 and D2 have equal IQRs, then the absolute distance the lows and high 25% of datapoints *can* spread will be greater for D1 than D2.

Measuring Clustering

But, IQR did not do for us what we really wanted it to do--it is too limited of a measure. What we want is a way of quantitatively measuring how the data *clusters*, not just the limits of where it can cluster. That is, I want a quantity that reflects how the datapoints are spread out on average, high tightly they cluster together, not just how far apart they *can* be spread out, or how tightly or loosely they can cluster together. I want to be able to tell, at a glance, that D1 has tighter clusters than D2. But range and IQR are incapable of showing this alone. This is why we need to construct new measures. Ultimately, this will culminate in Standard Deviation. But, to understand why we need Standard Deviation, here are attempts at providing a useful measure, and why they fail where Standard Deviation succeeds.

> **Naive Average Spread:** The average (mean) of all the distances of each datapoint from the median/midpoint.

$$naive\ avg.\ spread\ =\ \sum_{i}^{n}(x_i - median)/n$$

> **Naive Average Deviation:** The average (mean) of all the distances of each datapoint from the mean. The output is a proportion of the spread of each datapoint to the sample size. Thus, hypothetically, if we knew that two datasets had the same sample size (*n*), but one had higher naive avg dev than the other, then the datapoints of the first would be less clustered together than the second.

$$naive\ avg.\ dev.\ = \frac{\sum_{i}^{n}(x_i - \mu)}{n}$$

But this has almost no usefulness because of a trivial cancellation issue. You can't simply take the "naive" average deviation from the mean and expect it to be useful. For no matter what the values of the dataset are, this function will always output 0. If we're trying to compare datasets with this mean, but every dataset scores the same value, there would be nothing to differentiate, and therefore nothing to compare.

Average Mean Deviation: We could then try to modify the naive average deviation to avoid the uselessness of all sets resulting in 0 by including absolute value in the numerator. It would still reflect the proportion of spread to sample size, without having the cancelation issue. So, it still would allow us to see that, if two datasets had the same sample size (*n*), but one had higher avg mean dev than the other, then the datapoints of the first would be less clustered together than the second. This is called the average mean deviation:

$$\frac{\sum_{i}^{n}|(x_i-\mu)|}{n}$$

This function solves the problem of all the values of the dataset canceling out to 0. However, it faces another, equally troubling issue. For infinitely many datasets, this function will produce the same value, even if the datasets have *drastically* different spreads and clustering. Try calculating the average mean deviation for the following two datasets:

D_1 = {-4,-4,4,4}
D_2 = {-6,-2,1,7}

You will find that, despite their drastic differences, both datasets produce an average mean deviation of 4. Comparing these datasets in terms of this value would not tell us anything about the spread of their data points relative to each other. Their differences would be apparent on a graph, but this value would not help us predict how their graphs might compare. Variance attempts to overcome this.

Variance: the average of deviation *squared* from the mean. Variance avoids the problem of all outputs being 0 or being the same for very different datasets by squaring each deviation.

$$variance = \frac{\sum_{i}^{n}(x_i-\mu)^2}{n}$$

x_i: the value of starting datapoint *i*
μ: the mean of the values of all data points in the dataset
n: sample size (the number of data points in the dataset)

Consult our dataset used previously again:

D_1 = {-4,-4,4,4}
D_2 = {-6,-2,1,7}

With average mean deviation, the final resulting score of these two datasets were the same, and so that value was useless for comparing these sorts of datasets. But with variance, D_1 ends up with a score of 16, and D_2 with a score of 22.5 (try it yourself). These are more useful for comparisons.

Variance thus *indirectly* reflects the deviation of each datapoint from the mean, and thus the clustering of the datapoints around the mean (numerator). And it puts this in proportion to the sample size. So, it still would allow us to see that, if two datasets had the same sample size (*n*), but one had higher variance than the other, then the datapoints of the first would be less clustered together than the second.

However, the variance outputs will be very cumbersome to work with--large, complex numbers. This is why we take standard deviation--to reduce the cumbersome quantity of the SD score.

Standard Deviation (SD): SD attempts to overcome these issues and leave us with a useful measure of clustering and spread without being cumbersome by taking the square root of the variance. SD is just variance squared. Squaring these variance values preserves their proportionality while giving us smaller, easier numbers to work with. But, it is still the fundamentally same measure as the variance: it is a function of the spread (deviation) of each datapoint from the mean, and the sample size, and thus reflects both. Here is the simplest version of the SD function:

$$S = \sqrt{\frac{\sum_{i}^{n}(x_i - \mu)^2}{n}}$$

x_i: the value of starting datapoint *i*

μ: the mean of the values of all data points in the dataset

n: sample size (the number of data points in the dataset)

The result of our previous dataset: SD outputs D_1 = 4, D_2 = 4.74. From this, we can see that D_2 has an 18.5% larger SD score than D_1. Since we know the both have the same sample size, this enables us to infer that the clustering *around the mean* of D_2 is 18.5% larger than that of D_1.

SD allows us to have an indirect measure of the clustering of datapoints *around the mean* by *indirectly* measuring the *average spread (deviation) of the datapoints from the mean.*

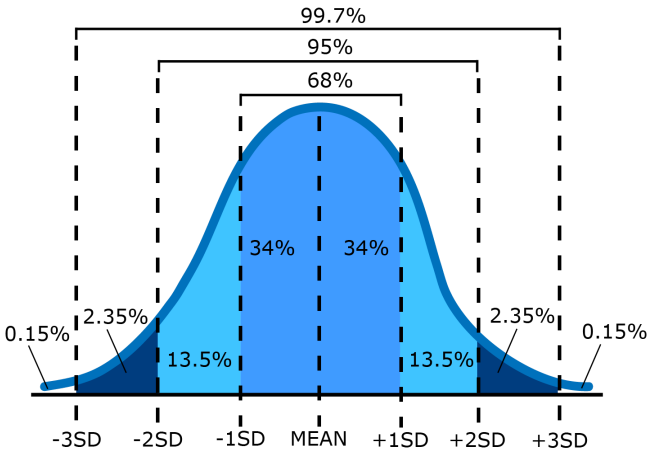
Using SD as an External, Comparative Measure:

- When *n* of each dataset is the same, the higher SD has looser clustering.
- The bigger sample size lowers the SD value or “score.” The smaller sample size increases the SD value. But, since variance

and SD are a *proportion* of clustering to sample size, you do not need to know sample size to use SD to compare datasets. The SD value is an indirect measure of degree of clustering proportional to the sample size. So even if the sample size differs, the output SD value will still reflect the amount of clustering on a graph. On two datasets with the same SD but very different sample sizes, the scatterplot distribution of the graphs will looks very similar, if not the same, but the graphs will just scale up or down.

Using SD as an Internal Measure

When we say “this datapoint is one standard deviation away,” what do we mean? We find the SD of a dataset. That value is now a unit of measurement. If SD = 13, then +1 SD is (*mean* + 13); -1 SD = (*mean* - 13). Internally, the SD of a dataset simply measures distance of some datapoint from the mean--there is nothing more significant to it than that. However, if we know the shape of a graph (the “distribution”), knowing how far some datapoint is away from the mean may help us predict that datapoint’s prevalence or probability using the *empirical rule.*



Empirical Rule (68-95-99.7 rule)**:** In a **normal or “gaussian” distribution** graph, that is, a **bell curve**, 34% of the data is one standard deviation above the mean (+1 SD), and 34% of the data is one standard deviation below the mean (−1SD). Therefore, according to the Standard Deviation Rule, 68% of the data falls between one standard deviation from the mean. This is merely a trivial result from the definition of normal distribution, and is not a prediction about real world datasets.

Normal Distribution: a dataset has a normal distribution if its median is also its mean and mode, and the data points fall symmetrically around either side of the mean. Normal distributions can be flatter or sharper than the one pictured above, what matters is the symmetrical distribution around the mean.

Right-Skewed Distribution: most of the datapoints fall to the right (are greater than) the median. A **Left-Skewed Distribution** would be the opposite: most of the datapoints fall to the left (are less than) the median.