جامعة نيويورك أبوظبي

NYU ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

Class 4: Describing our results - Central tendency and variability

Prof. Jon Sprouse
Psychology

Describing distributions

It is great to look at frequency tables, plots, and frequency polygons. But sometimes, we want single numbers that can describe some aspect of the distribution.

There are different types of information that one could be interested in. Here are two types that arise frequently:

Central Tendency: (or location)	A measure of location/central tendency gives a single value that is representative of the distribution as a whole (its expected value). The three most common measures of this are the mean, median, and mode.
Variability: (or spread/dispersion)	A measure of variability/dispersion/spread gives a single value that indicates how different the values in a distribution are from each other. The most common measures are variance and standard deviation, and sometimes the absolute deviation.

We will see these over and over again, but for now, I will simply define them so that we are all on the same page mathematically when they come up later.

Three measures of central tendency

Central Tendency: Mean

Let's start with the (arithmetic) mean. This is commonly called the average, but you should avoid that word. It is not precise enough. Say mean.

Mean: The sum of the values, divided by the number of values (the count) that were summed.

$$\textbf{Mean} = \frac{x_1 + x_2 + \dots x_n}{n}$$

The mean is by far the most common measure of central tendency, so you will encounter (and use it often). The primary benefit of the mean is that all data points contribute equally to it. But this is also a drawback, as it means that it is affected by outliers.

mean(1, 2, 3, 4, 5) = 3

mean(1, 2, 3, 4, 10) = 4

mean(1, 2, 3, 4, 100) = 22

outliers: Values that are much larger or much smaller than the rest of the values in a distribution.

Central Tendency: Median

The next most common measure of central tendency is the median.

Median: The median is the value in a set of values that divides the set into two halves (an upper half and a lower half). If there is an odd number of values in the set, the median will be one of the values in the set. If there is an even number, the median will be the mean of the two middle values.

The median is interesting for a number of reasons, but perhaps the most valuable aspect of the median is that it is **robust to outliers**. This is just a fancy way of saying that the median is not influenced by very large (or very small) numbers. This is because all that matters is the order of the values, not the size of the values. This is in stark contrast to the mean.

mean(1, 2, 3, 4, 5) = 3

mean(1, 2, 3, 4, 10) = 4

mean(1, 2, 3, 4, 100) = 22

median(1, 2, 3, 4, 5) = 3

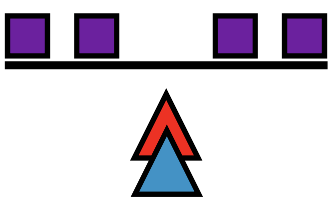
median(1, 2, 3, 4, 10) = 3

median(1, 2, 3, 4, 100) = 3

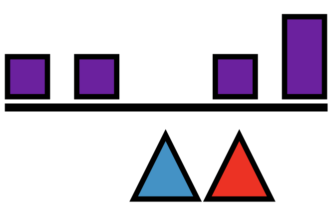
The Mean/Median see-saw analogy

I am not kidding when I say that there is a nifty visual analogy for means and medians involving a seesaw.

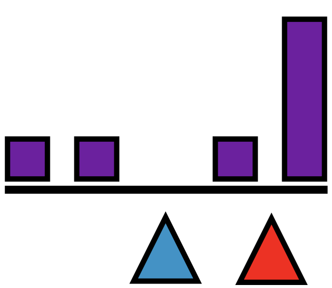
Imagine that you have 4 people on a see-saw, roughly split 2 on each side. Where do you place the fulcrum of the see-saw? The mean will be the balance point between their weights, and the median will be the point that keeps two people on either side of fulcrum.



If their weights are equal, the mean and median will be the same.



If one of them weighs more than the others, the median will still be in the same place - to split the teams. But the mean would move to adjust for the extra weight of that person.

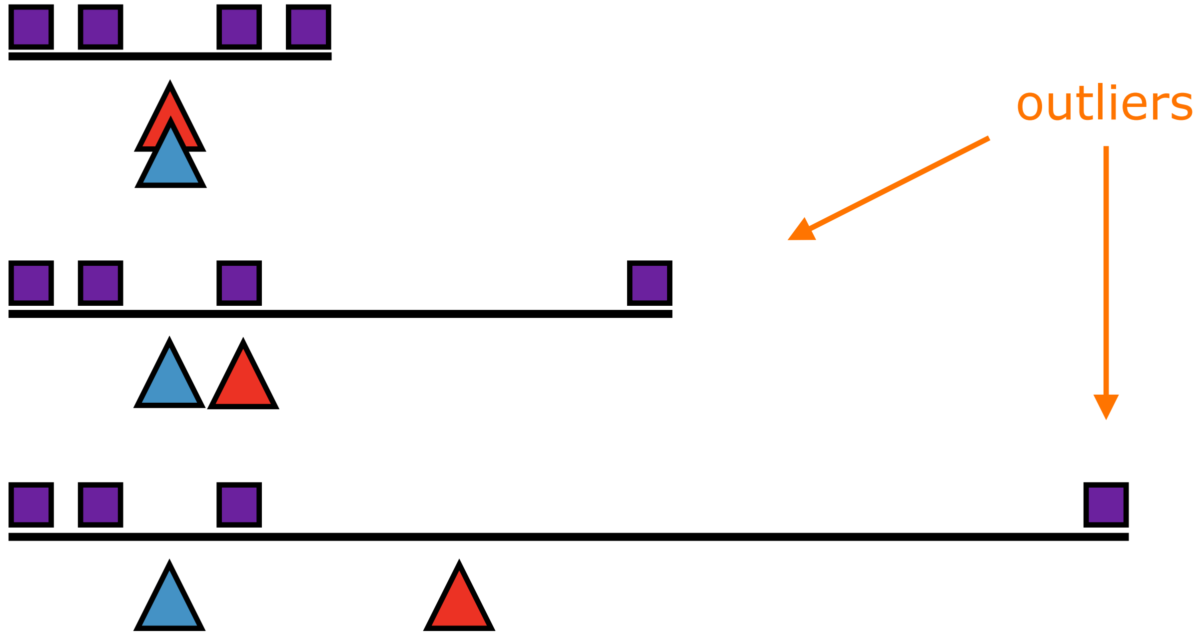


How far it moves depends on the how much more that person weighs.

The Mean/Median see-saw analogy

You can also think about the see-saw analogy with distances. This is closer to the way we plot distributions, so it is sometimes the way it is presented.

Here you need to imagine that the location on the see-saw indicates something about weight — kind of like the way something sitting on the edge of a long board will be heavier than something sitting on a short board (when you hold it by the other end). The same logic applies — the median will always split between 2 and 2. The mean will move based on the distance that the fourth person is from the group:



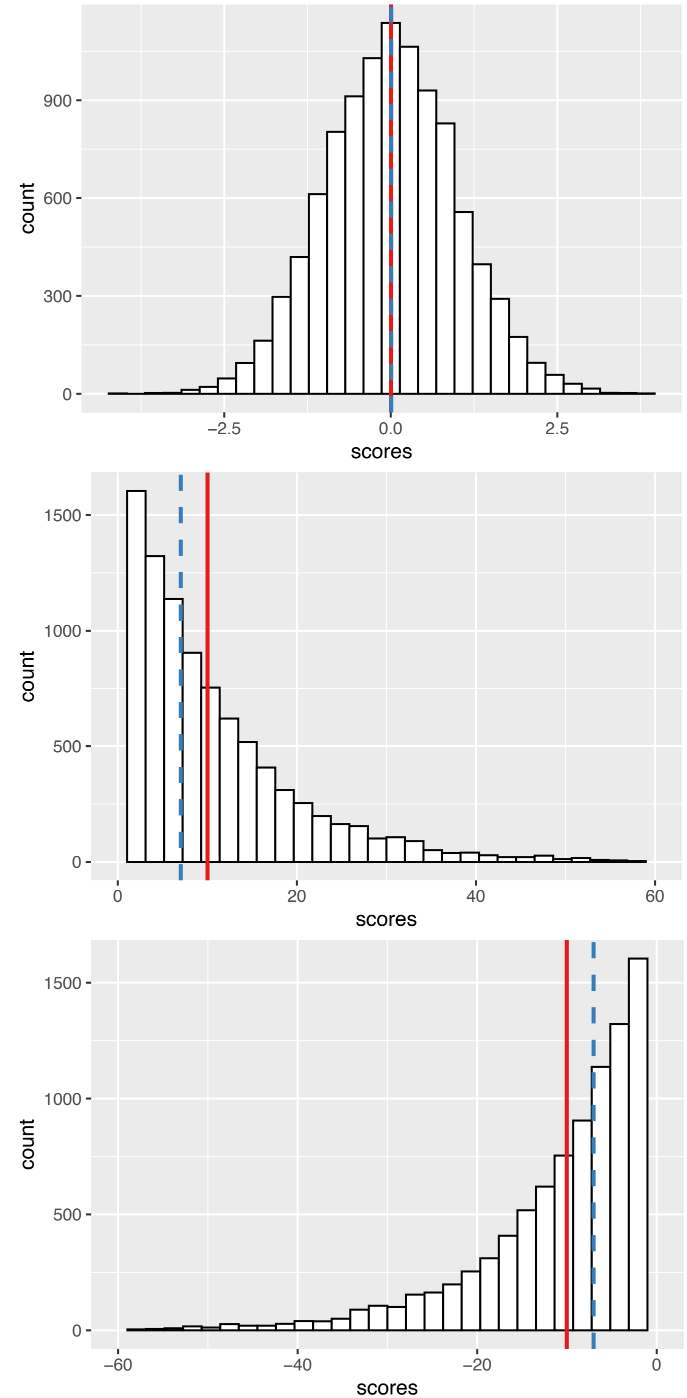
Mean/Median and the shape of distributions

If the distribution is **symmetric**, the mean and median will be identical.

Asymmetric distributions are said to be **skewed**. In a skewed distribution, the mean and median will not be identical.

A distribution is **positively skewed** if it has a long rightward tail (toward larger positive numbers).

A distribution is **negatively skewed** if it has a long leftward tail (toward smaller numbers).

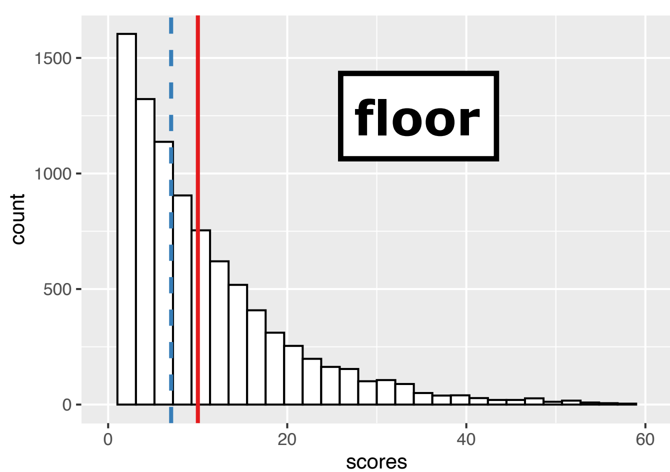


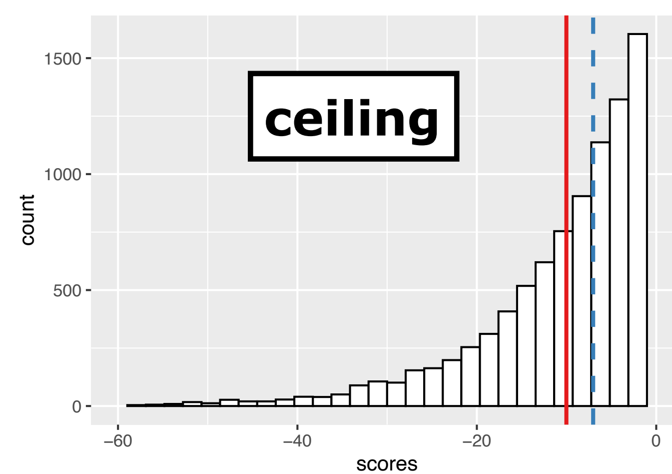
Skew and ceiling/floor effects

Skew can have many sources. For example, some phenomena simply have an exponential distribution that leads to a tail.

But the most common causes of skew are ceiling and floor effects.

Ceiling and floor effects arise when there is a hard boundary on one edge of the scale. A floor effect is when the hard boundary is on the lower end of the scale. A ceiling effect is when the hard boundary is on the higher end of the scale.





One very common floor effect in psychology is when we measure time - reaction times, study time, etc. Time has a natural floor of 0. This leads times to be positively skewed.

Choosing between mean and median

For the inferential statistics that we will do later in this course, we will use the mean. This is because much of frequentist statistics is built around the mean as the measure of central tendency.

But if you are just trying to describe your data set (descriptive statistics), you often have a choice:

The critical question is whether extreme values are relevant to your scientific theory/question or not. If they are relevant to your theory/question, then you want to use the mean so that they are included. If they are not relevant (they are outliers), then you may want to use the median.

Decisions about safety design

Injuries during car trips are rare. They are outliers. If you took the median car trip, it would tell you there are 0 injuries. And then you might not add safety features. So we use the mean.

Studies of income distribution

The very rich (like Jeff Bezos or Elon Musk) can distort our view of the average wealth of citizens in states like Washington or California. So we use the median instead.

Central Tendency: Mode

Mode: The most frequent score in a distribution.

The mode is not very common as a measure of central tendency. But it is the most flexible - it is defined for all four measurement types. It is worth noticing that the median is not defined for nominal measurements, and the mean is not defined for nominal or ordinal measurements:

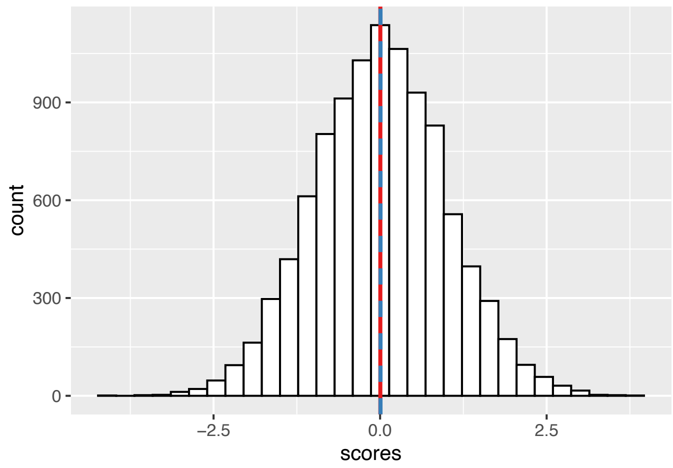
	nominal	**ordinal**	**interval**	**ratio**
mode	✓	✓	✓	✓
median	X	✓	✓	✓
mean	X	X	✓	✓

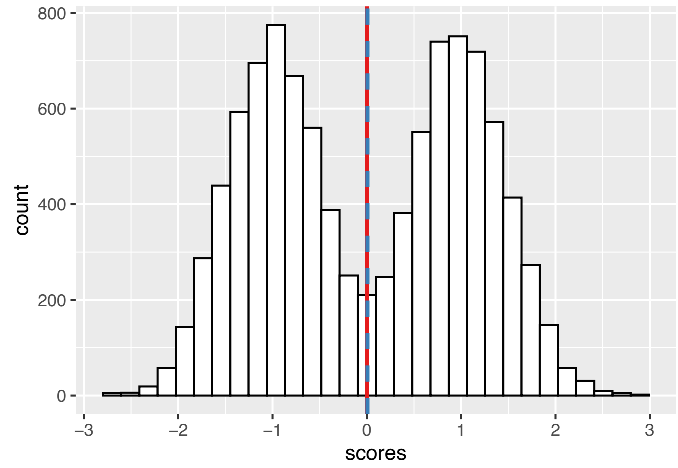
Mode and the shape of the distribution

Though the mode is not very common as a measure of central tendency, it is used as an additional dimension for reporting the shape of a distribution.

unimodal: One mode, visible as a single peak in the distribution.

bimodal: Two modes, visible as two peaks in the distribution.





You should always look at the distribution of your data before doing a statistical analysis. You should check to make sure it is unimodal. If it is bimodal, it may indicate that your participants were sampled from two different populations rather than being sampled from the same population.

Measures of variability

Range, Interquartile Range, and Semi-Interquartile Range

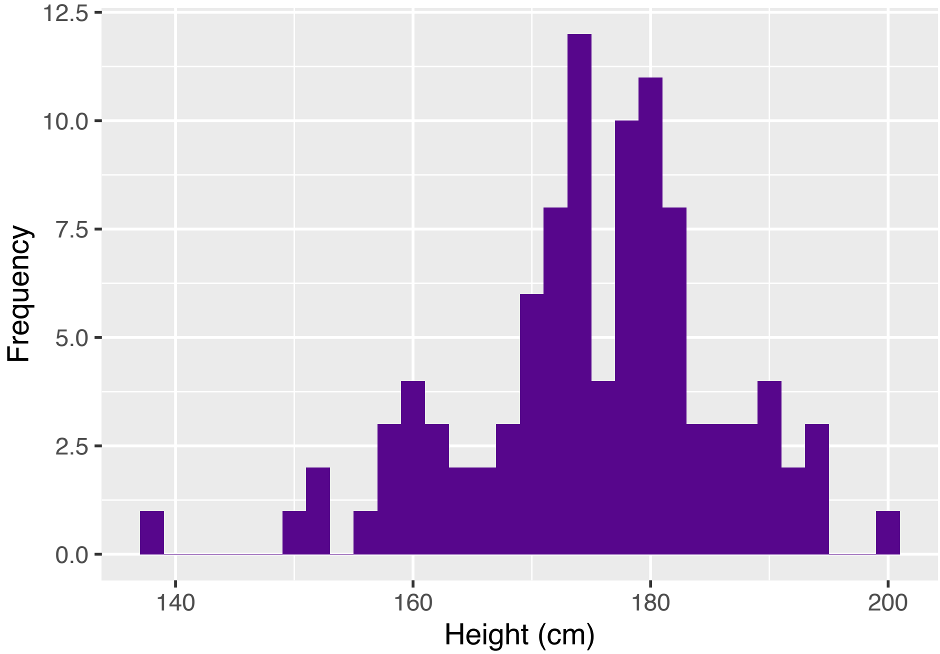
(You should know that these exist, but they are not as important as the standard deviation and absolute deviation measures we will see later)

The range

The **range** is sometimes defined as the difference between the smallest and largest value in a data set.

Let’s simulate height data again. The smallest value is 137.6 and the largest is 200. So the range is 62.4.

However, in practice, the range is the report of the smallest and largest values: 137.6 to 200. If you use the range() function in R, you will get a vector with these two numbers.

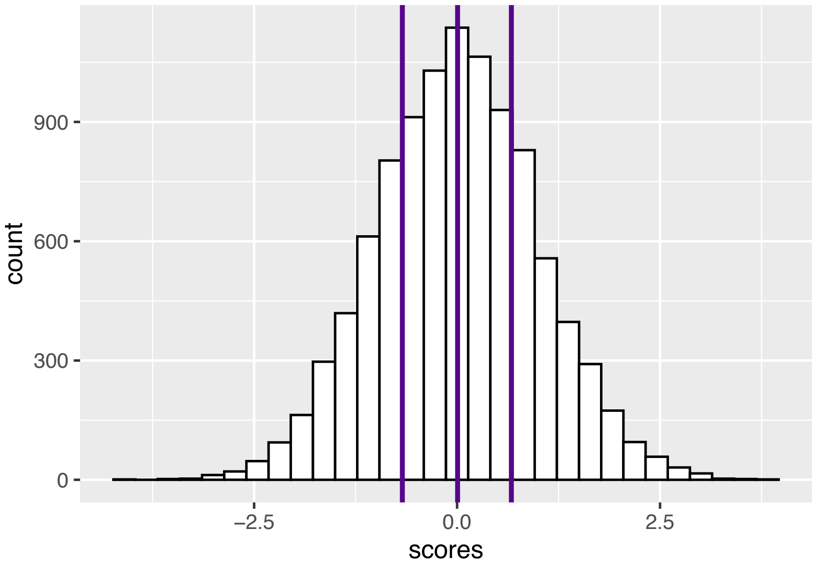


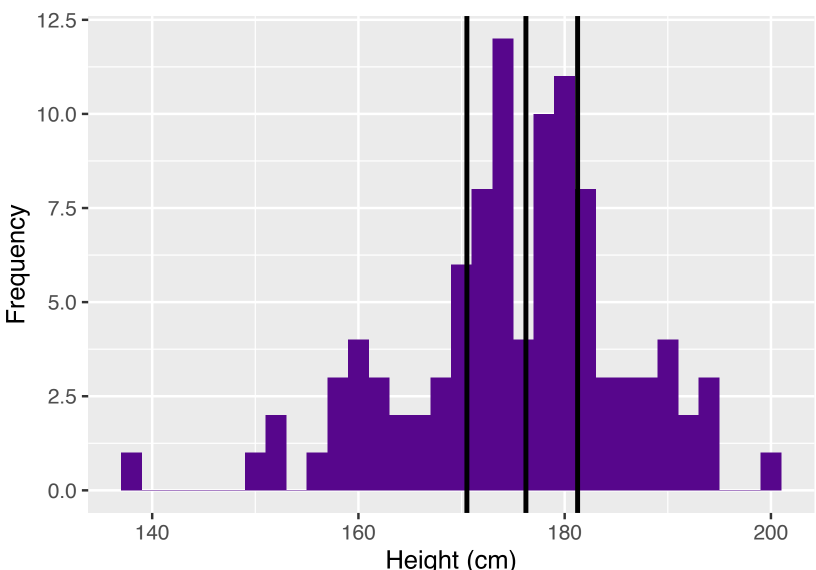
The main benefit of the **range** is that it includes every data point. It is exhaustive. The main drawback is that it is influenced by outliers in both directions. Therefore it is really only beneficial if you truly want to report the range of the measure. It is not so great as a measure of variability.

Quartiles

A **quantile** is a general term for the cut points that create adjacent chunks that each contain the same number of data points.

The most common quantiles are those that create chunks that contain 25% of the data set. They are called **quartiles** because there are four chunks: the one from 0 to 25%, the one from 25% to 50%, the one from 50% to 75%, and the one from 75% to 100%. The plot on the right is a symmetric distribution with 3 quartiles (cut points) in purple, creating 4 chunks.

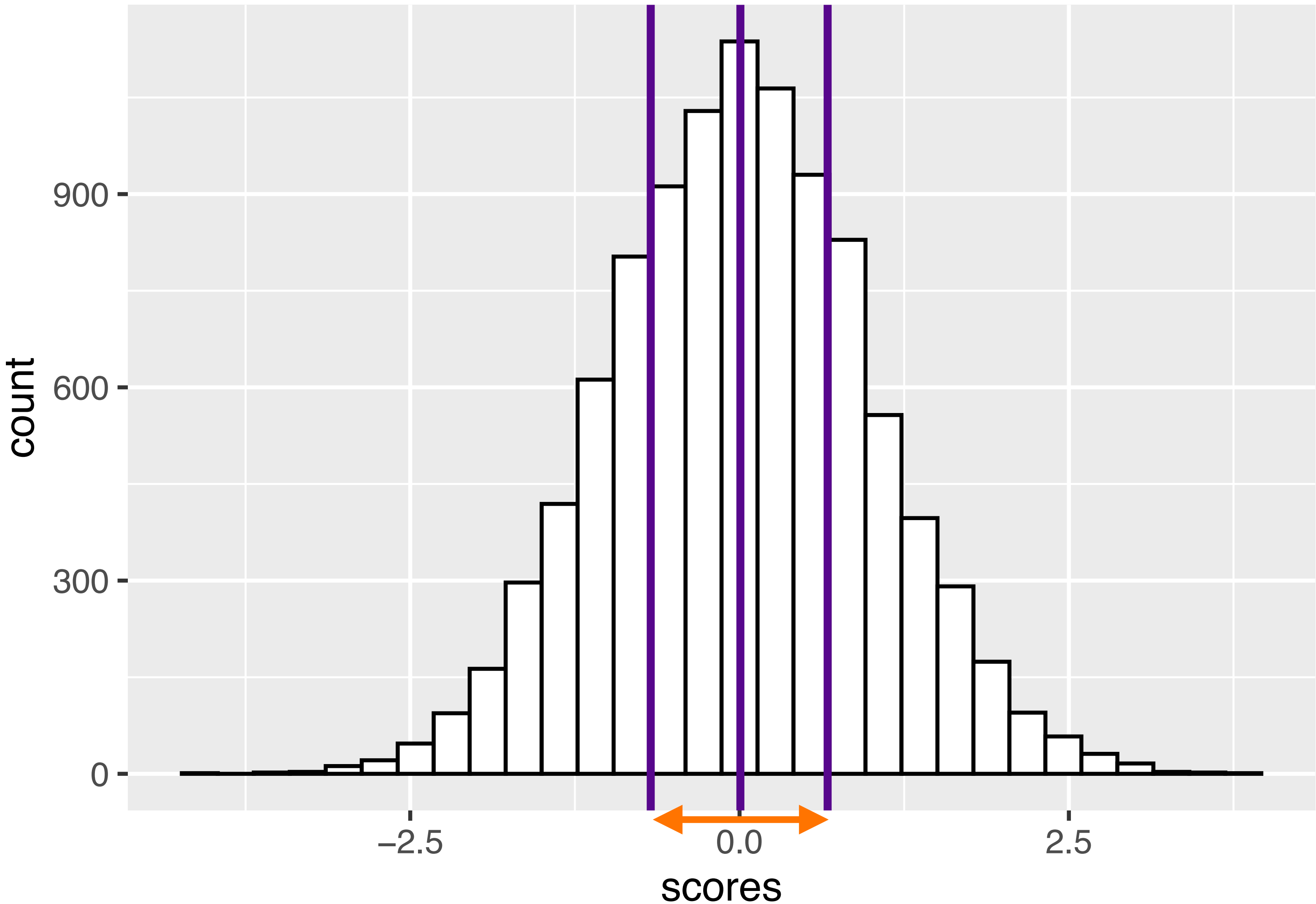




We can also do this with real data sets. The plot on the left is our simulated height data plotted with quartiles in black dividing the four chunks.

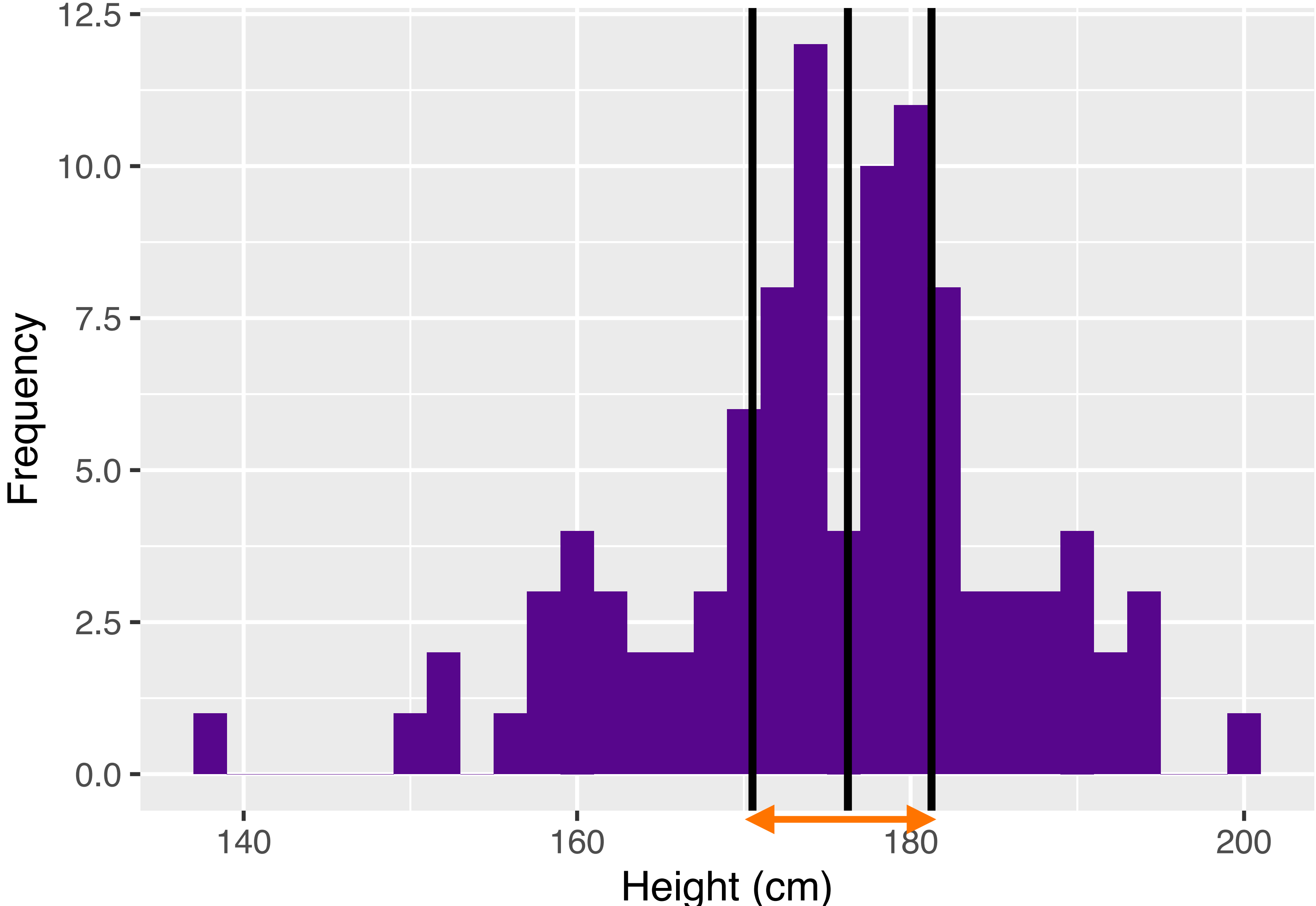
The interquartile range

The **interquartile range** is the range between the 1st quartile and the 3rd quartile. Visually, it is the range between the first line and the third line in these plots. It shows us the middle 50% of the data, so like the median (the middle line), it is not impacted by outliers!



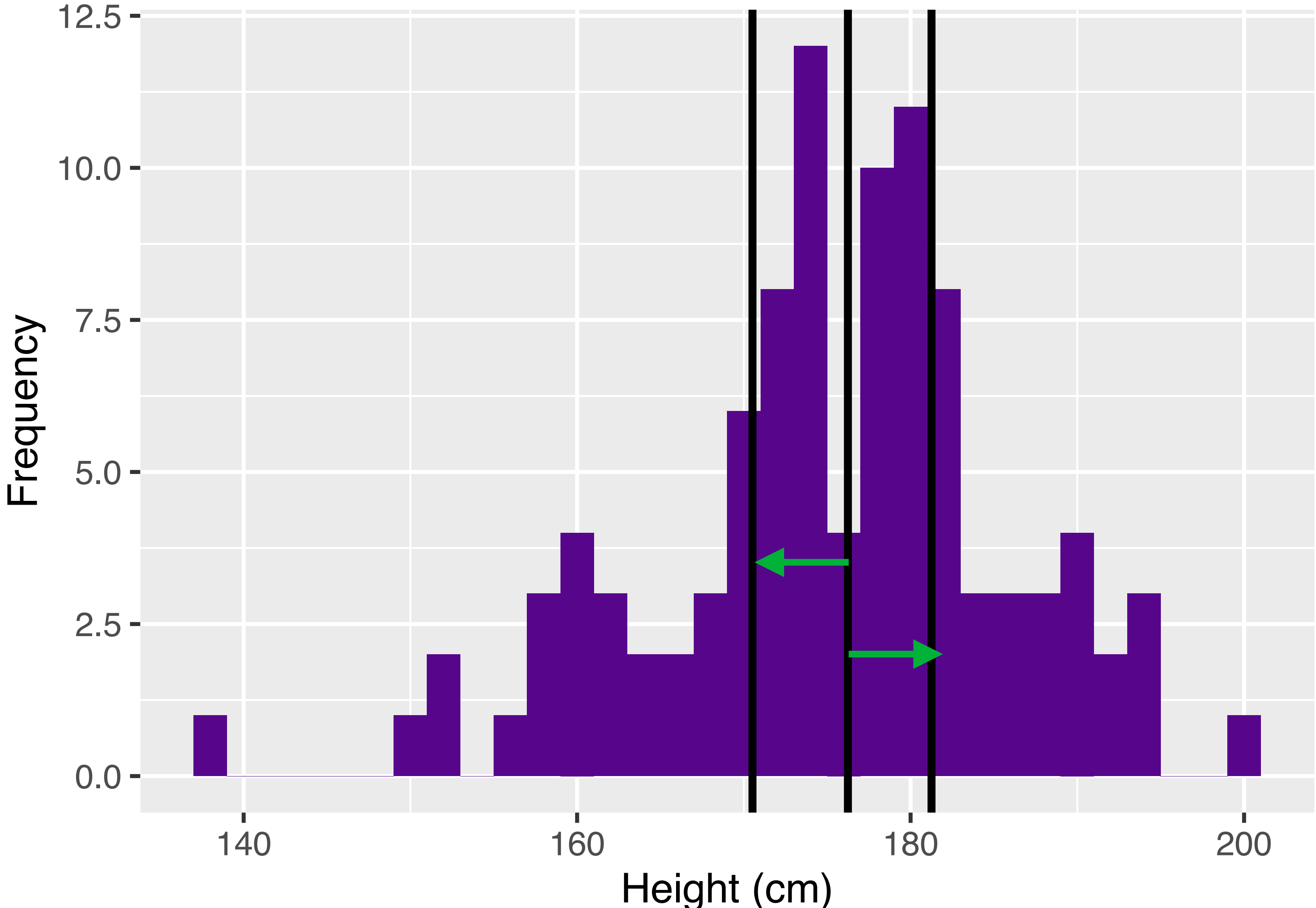
The interquartile range

The **interquartile range** for our simulated height data set is 10.7. You can find it using the R function IQR(). Unlike raw range, the interquartile range is reported as a single number (the technical definition of range) because it is seen as a measure of variability (it captures the middle 50% of the data).



The semi-interquartile range

A **semi-interquartile range** is the interquartile range divided by 2. So for our data set it is around 5cm. It is seen as a measure of variability moving in either direction away from the median. It is rarely seen because the median is rarely used in inferential statistics. But it is worth knowing what it is.



Standard deviation and Absolute deviation

(These are the important ones. And standard deviation is the most important of all.)

Deviation scores

The idea behind a deviation score is really straightforward. We choose a central tendency for the data set, and then ask how far each data point is from that central tendency. To calculate distance, we do subtraction:

deviation: The distance of each score from the central tendency.

Deviation = (x - CT)

Data set: 1, 2, 3, 4, 10

Central Tendency:
mean

Mean: 4

deviation of 1 = 1 - 4 = -3

deviation of 2 = 2 - 4 = -2

deviation of 3 = 3 - 4 = -1

deviation of 4 = 4 - 4 = 0

deviation of 10 = 10 - 4 = 6

A failed attempt at variability

What we want is a measure of variability for the entire data set. An obvious and logical idea would be to **sum** the deviations for each score, and use that sum as a measure of **variability** in the data set. Let's try this with our previous example:

Data set: 1, 2, 3, 4, 10

Central Tendency:
mean

Mean: 4

deviation of 1 = 1 - 4 = -3

deviation of 2 = 2 - 4 = -2

deviation of 3 = 3 - 4 = -1

deviation of 4 = 4 - 4 = 0

deviation of 10 = 10 - 4 = 6

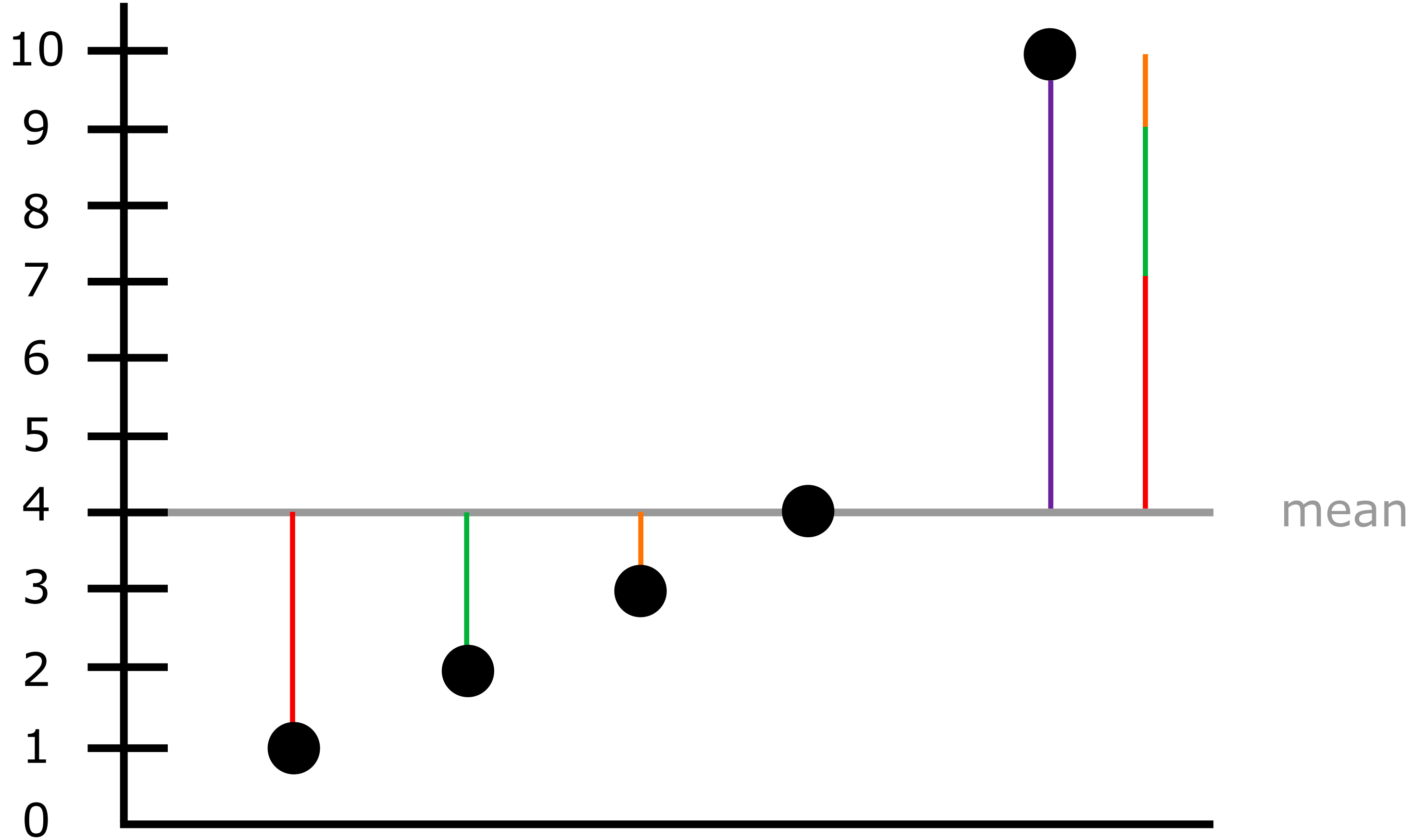
The problem with plain deviation scores is that they do not work as a measure of variability for the entire data set when the CT is the mean.

sum of the deviations:

-3 + -2 + -1 + 0 + 6 = 0

So what we learn here is that the sum of basic deviation scores is not going to be helpful for us when we use mean. It will always sum to 0.

The problem with the mean (visual)



Because of the definition of the mean, the deviation (from the mean) of the points below the mean will always equal the deviation of the points above the mean. So it is impossible to simply sum this deviation.

Squared deviations

The critical problem with deviations is that they are both positive and negative. But if we could eliminate the negative signs so they are all positive, we could sum them without running into a problem. One way to get rid of negative signs is to square them.

$$\text{squared deviation} = (x - CT)^2$$

We can now sum the squared deviations as a measure of total variability:

$$\text{sum of squares} = (x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2$$

Data set: 1, 2, 3, 4, 10
Central Tendency: mean
Mean: 4

sum of squares:

9 + 4 + 1 + 0 + 36 = 50

deviation² of 1 = $(1 - 4)^2 = 9$

deviation² of 2 = $(2 - 4)^2 = 4$

deviation² of 3 = $(3 - 4)^2 = 1$

deviation² of 4 = $(4 - 4)^2 = 0$

deviation² of 10 = $(10 - 4)^2 = 36$

Why we can't stop. The problem with SS.

The sum of squares is a measure of total variability in the data set:

$$\text{sum of squares} = (x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2$$

We could try to use the sum of squares as our measure of variability. But one problem with the sum of squares is that its size is dependent upon the number of values in the set. Larger sets could have larger sum of squares simply because they have more values, even though there might really be less variation. Here are two examples to illustrate this

[1, 2, 3, 4, 10]

sum of squares: 9 + 4 + 1 + 0 + 36 = **50**

[4,4,4,4,4,4,4,4,4,4,6,8,8,8,8,8,8,8,8,8,8]

sum of squares: (4x10) + (4x10) = **80**

Intuitively, you probably agree that the first set is more variable. The range is higher (9 vs 4). The number of different scores is higher (5 vs 3). But SS is higher for the second set simply because there are simply more values in the data set.

Variance

The solution to the problem (that SS tends to grow with the number of data points in the set) is to **divide by the number of data points**:

$$\text{variance} = \frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}$$

The intuition behind this is similar to the way we divide by the number of data points when we calculate the mean — dividing a sum by the number of data points gives us an "average" - each data point contributes a portion of itself to the sum. So, here we are calculating an average measure of variability. This gets a special name called **variance**.

The variance is a really important concept in statistics. So, you should memorize this. I find it easiest to memorize this by focusing on the logic we just went through — deviation scores, square them to eliminate negative signs, sum them for total variability, divide by n to get an average measure of variability.

Standard Deviation

Although variance is a useful measure, and we will see it often in statistics, it does have one problem. It is in really strange units - the units of measure are squared!

$$\text{variance} = \frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}$$

height squared?
stress ratings squared?
days ill squared?

These are not meaningful in the real world. So we probably don't want this as our only measure.

The fix for this should be obvious. We can simply take the square root of the variance to change it back into un-squared units. We call this the **standard deviation**:

$$\text{standard deviation} = \sqrt{\frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}}$$

same units as the original values - cm, days, etc.

Standard Deviation

The standard deviation is exactly what the name sounds like - a measure of the amount of deviation that a typical data point shows around the central tendency.

$$\text{standard deviation} = \sqrt{\frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}}$$

The **standard deviation** is a really important concept in statistics. So, you should memorize this. I find it easiest to memorize this by focusing on the logic we just went through — deviation scores, square them to eliminate negative signs, sum them for variability (sum of squares), divide by n to get an average measure of variability (variance), and take the square root to change it back to the original unit of measure (standard deviation). I know it is a lot. But just remember that each step in the chain has a logical motivation. If you can reconstruct the logic, you can always reconstruct the equation!

Which CT should we use? The mean.

sum of squares = $(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2$

$$\text{variance} = \frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}$$

$$\text{standard deviation} = \sqrt{\frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}}$$

I have written these with CT because they can, in principle, be used with either the mean or median. But variance and standard deviation are **always used with the mean**. To show you why, I need to take a step back and show you a different deviation score that we could have defined…

The absolute deviation - no squaring for us!

Let’s go back to our first logical move. We needed to eliminate the negative signs for our deviation scores, so we squared them. That probably sounded odd to you because you know we have another option — absolute values:

$$\text{absolute deviation} = |x - CT|$$

Then we can sum the absolute deviations:

$$\text{sum of absolute deviations} = |x_1 - CT| + |x_2 - CT| + \ldots + |x_n - CT|$$

And then divide by n to get an average of the sum of the absolute deviations:

$$\text{average absolute deviation} = \frac{|x_1 - CT| + |x_2 - CT| + \ldots + |x_n - CT|}{n}$$

And note that we don’t have to take a square root because the units are still the same as the original data set (we didn’t square anything).

(Average) Absolute Deviation

The (Average) Absolute Deviation is similar to the standard deviation. It tells us how a typical data point deviates from the central tendency.

$$\text{average absolute deviation} = \frac{|x_1 - CT| + |x_2 - CT| + \ldots + |x_n - CT|}{n}$$

And just like the standard deviation, it can be used with either the mean or the median:

$$\text{mean absolute deviation} = \frac{|x_1 - \text{mean}| + |x_2 - \text{mean}| + \ldots + |x_n - \text{mean}|}{n}$$

$$\text{median absolute deviation} = \frac{|x_1 - \text{median}| + |x_2 - \text{median}| + \ldots + |x_n - \text{median}|}{n}$$

But, in practice, you will typically see the absolute deviation with the median, just like you see the standard deviation with the mean.

Why mean w/SD and median w/AD?

Even though we could, in principle, use either central tendency for these measures of variability, we tend to use a very specific mapping:

standard deviation = $\sqrt{\dfrac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}}$ ← mean

absolute deviation = $\dfrac{|x_1 - CT| + |x_2 - CT| + \ldots + |x_n - CT|}{n}$ ← median

The reason is that we want **minimize the variability**. We want the smallest value that our equations will give so that we know it is a lower bound on the estimate of the variability. (As we saw with sum of squares, higher numbers can be due to reasons that don't really matter for our science.)

So we can ask a mathematical question: Which measure of CT minimizes the standard deviation? And which measure of CT minimizes the absolute deviation?

Why mean w/SD and median w/AD?

Even though we could, in principle, use either central tendency for these measures of variability, we tend to use a very specific mapping:

standard deviation = $\sqrt{\dfrac{(x_1 - CT)^2 + (x_2 - CT)^2 + \ldots + (x_n - CT)^2}{n}}$ ← mean

absolute deviation = $\dfrac{|x_1 - CT| + |x_2 - CT| + \ldots + |x_n - CT|}{n}$ ← median

The mean is the measure of central tendency that minimizes variance (and standard deviation). The variance we calculate by using the mean will always be smaller than (or equal to) the variance by using the median.

The median is the measure of central tendency that minimizes the absolute deviation. The absolute deviation we calculate by using the median will always be smaller than (or equal to) the absolute deviation by using the mean.

Let's simulate this!

Let's take 1000 samples of size 20. For each one, we will calculate all four possible measures:

mean standard deviation = $\sqrt{\dfrac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \ldots + (x_n - \text{mean})^2}{n}}$

median standard deviation = $\sqrt{\dfrac{(x_1 - \text{median})^2 + (x_2 - \text{median})^2 + \ldots + (x_n - \text{median})^2}{n}}$

---

mean absolute deviation = $\dfrac{|x_1 - \text{mean}| + |x_2 - \text{mean}| + \ldots + |x_n - \text{mean}|}{n}$

median absolute deviation = $\dfrac{|x_1 - \text{median}| + |x_2 - \text{median}| + \ldots + |x_n - \text{median}|}{n}$

Let's simulate this!

Let's take 1000 samples of size 20. For each one, we will calculate all four possible measures.

I am not showing you the R code - but these are the results of the code. I created a population with a mean of 0 and a standard deviation of 1, and had R create 1000 samples of size 20.

Then I asked R to calculate each of our 4 measures. What you see on the right are the four measures for the first 20 of the 1000 samples.

We can look at these to test our claims!

	MeanSD	MedianSD	MeanAD	MedianAD
1	1.0222072	1.0315435	0.9167606	0.9167606
2	0.9060478	0.9071818	0.8585974	0.8585974
3	0.8680902	0.8775020	0.8672157	0.8661544
4	0.7583353	0.7638103	0.7840383	0.7840383
5	1.1980147	1.2063635	1.0122249	1.0018999
6	1.0482317	1.0487225	0.9000681	0.9000681
7	0.8042305	0.8106485	0.8056556	0.7969203
8	0.9304095	0.9320397	0.8795878	0.8795878
9	0.9754150	0.9929334	0.9038638	0.8874771
10	0.8532289	0.8561514	0.8258634	0.8258634
11	0.9568814	0.9568934	0.8730349	0.8730349
12	0.8701299	0.8931737	0.8032382	0.7919271
13	0.9178570	0.9250142	0.8334220	0.8297906
14	1.0943475	1.0966130	0.8853802	0.8842262
15	1.0231321	1.0380066	0.8864093	0.8795511
16	0.7793890	0.7916873	0.7918395	0.7824341
17	0.9809680	1.0007803	0.9393706	0.9393706
18	0.9573879	0.9573894	0.8691631	0.8691631
19	1.2327010	1.2388635	0.9838237	0.9797697
20	1.0253671	1.0357164	0.9072601	0.8983947

Testing our claims

Claim 1:

For Standard Deviations, the mean will always be smaller than or equal to the median.

Look at the table on the right. What do you see? The MeanSD is always smaller than or equal to the MedianSD!

Claim 2:

For Absolute Deviations, the median will always be smaller than or equal to the mean.

And that is what we see!

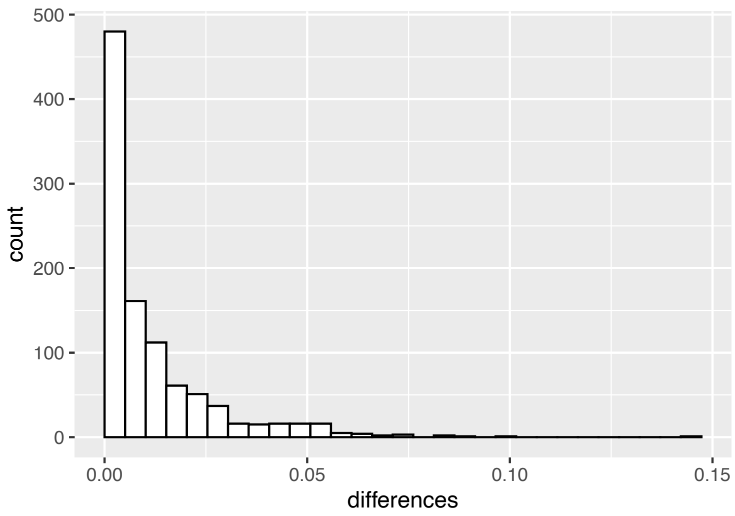
	MeanSD	MedianSD	MeanAD	MedianAD
1	1.0222072	1.0315435	0.9167606	0.9167606
2	0.9060478	0.9071818	0.8585974	0.8585974
3	0.8680902	0.8775020	0.8672157	0.8661544
4	0.7583353	0.7638103	0.7840383	0.7840383
5	1.1980147	1.2063635	1.0122249	1.0018999
6	1.0482317	1.0487225	0.9000681	0.9000681
7	0.8042305	0.8106485	0.8056556	0.7969203
8	0.9304095	0.9320397	0.8795878	0.8795878
9	0.9754150	0.9929334	0.9038638	0.8874771
10	0.8532289	0.8561514	0.8258634	0.8258634
11	0.9568814	0.9568934	0.8730349	0.8730349
12	0.8701299	0.8931737	0.8032382	0.7919271
13	0.9178570	0.9250142	0.8334220	0.8297906
14	1.0943475	1.0966130	0.8853802	0.8842262
15	1.0231321	1.0380066	0.8864093	0.8795511
16	0.7793890	0.7916873	0.7918395	0.7824341
17	0.9809680	1.0007803	0.9393706	0.9393706
18	0.9573879	0.9573894	0.8691631	0.8691631
19	1.2327010	1.2388635	0.9838237	0.9797697
20	1.0253671	1.0357164	0.9072601	0.8983947

Showing it for the full 1000 samples

Claim 1: For Standard Deviations, the mean will always be <u>smaller</u> than the median.

Prediction: If we calculate the difference median-mean, the result will always be <u>positive</u>.

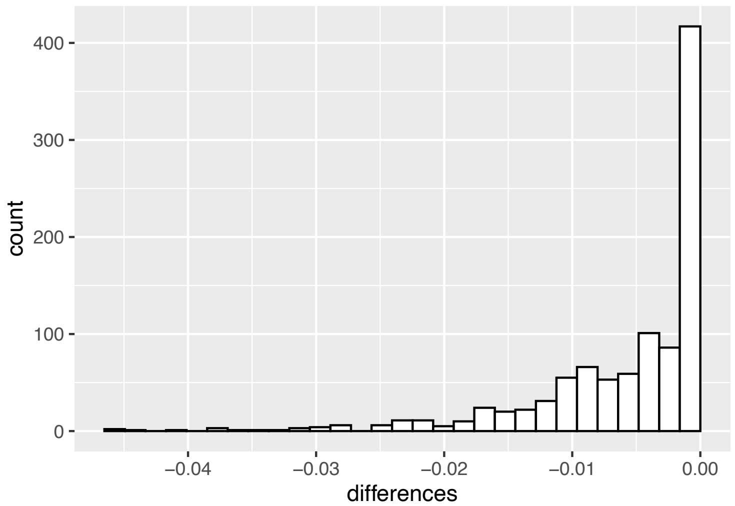
Prediction: All 1000 of the differences are <u>positive</u>. The mean is always smaller.



Claim 2: For Absolute Deviations, the median will always be <u>smaller</u> than the mean.

Prediction: If we calculate the difference median-mean, the result will always be <u>negative</u>.

Prediction: All 1000 of the differences are <u>negative</u>. The median is always smaller.



Why mean w/SD and median w/AD?

Even though we could, in principle, use either central tendency for these measures of variability, we tend to use a very specific mapping:

standard deviation = $$\sqrt{\frac{(x_1 - CT)^2 + (x_2 - CT)^2 + \dots + (x_n - CT)^2}{n}}$$ ← mean

absolute deviation = $$\frac{|x_1 - CT| + |x_2 - CT| + \dots + |x_n - CT|}{n}$$ ← median

The mean is the measure of central tendency that minimizes variance (and standard deviation). The variance we calculate by using the mean will always be smaller than (or equal to) the variance by using the median.

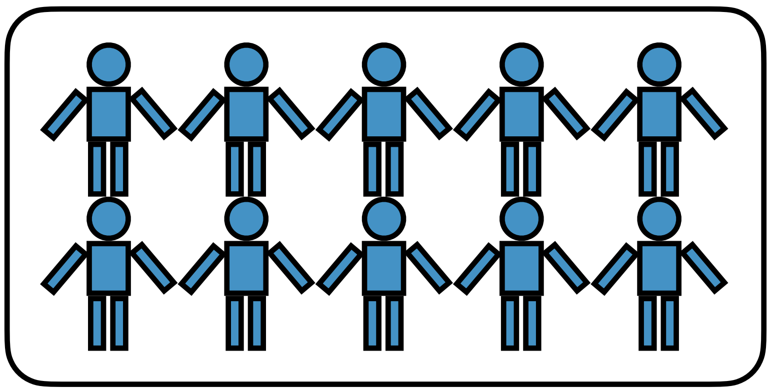
The median is the measure of central tendency that minimizes the absolute deviation. The absolute deviation we calculate by using the median will always be smaller than (or equal to) the absolute deviation by using the mean.

Parameters versus Statistics

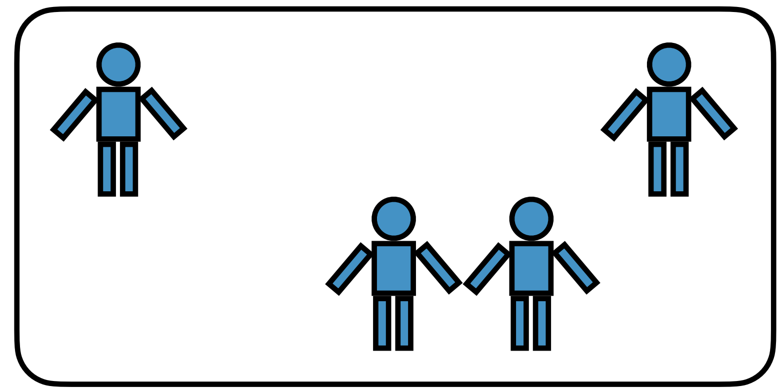
(Wherein we learn why Statistics is called “Statistics”)

Populations and Samples

As scientists, we study populations, but we work with samples.



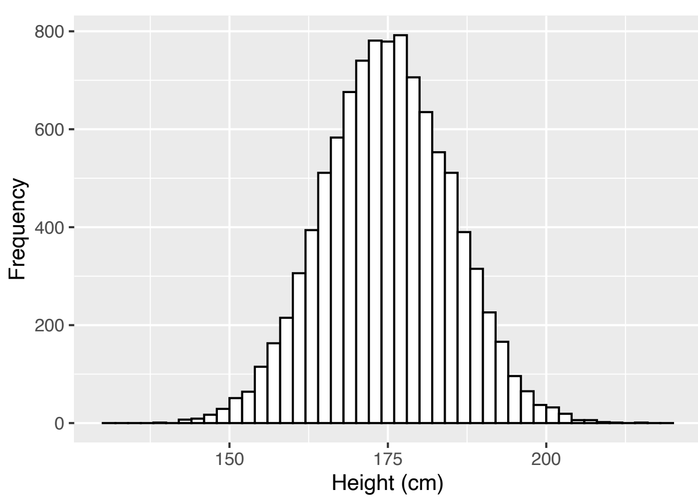
sampling



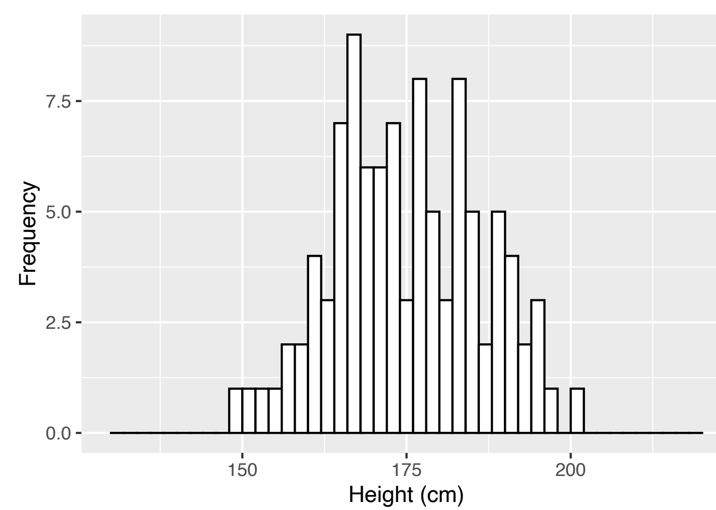
population

sample

Both are just sets (or distributions). So we can describe both with the same mathematical concepts: mean, standard deviation, etc.

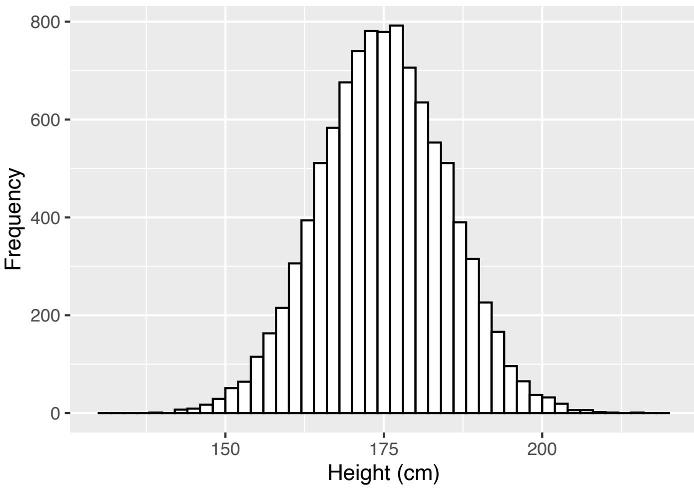


sampling



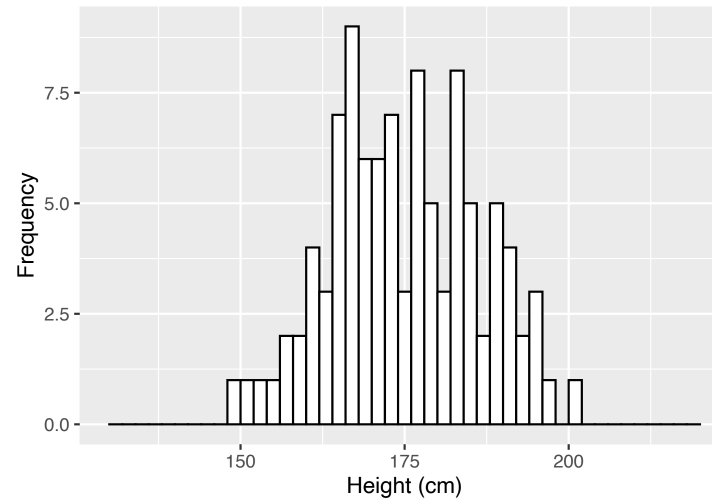
Parameter and Statistic

Mathematically, populations and samples are the same thing - a distribution



population

sampling



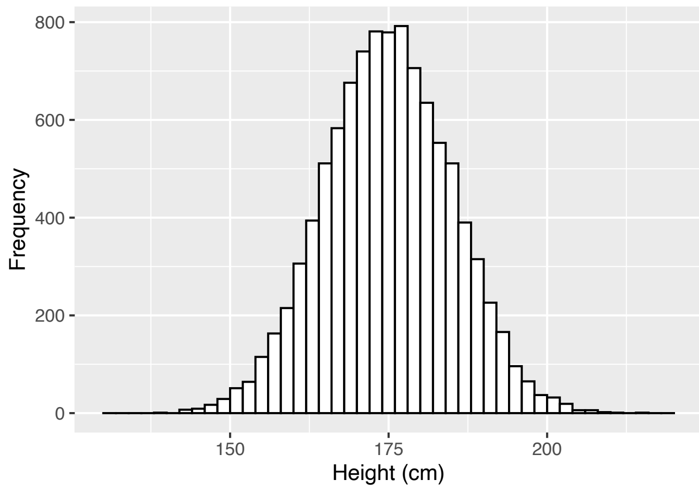
sample

So we can describe them using the same mathematical concepts - mean, standard deviation, etc. But when we do that, we use a different term for the numbers that describe populations and the numbers that describe samples:

Parameter: A number (like mean, standard deviation) that describes an aspect of a population. Usually written with a Greek letter.

Statistic: A number (like mean, standard deviation) that describes an aspect of a sample. Usually written with a Roman/Latin letter.

The symbols for parameters and statistics



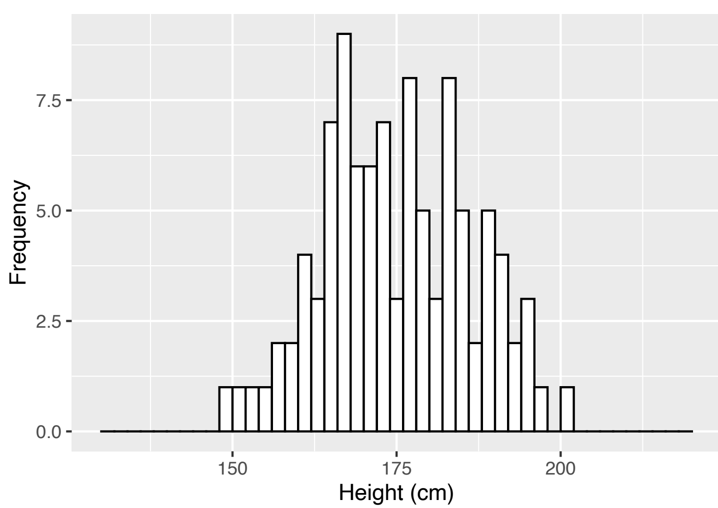
population

mean = μ

variance = σ^2

standard deviation = σ

Parameters use Greek letters



sample

mean = $\bar{x}$

variance = s^2

standard deviation = s

Statistics use Roman/Latin letters