# Variability

Variability describes how far apart data points lie from each other and from the center of a distribution

Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data.

These other measures we use with variability are

* Range
  + The difference between the highest and lowest values
* Interquartile range (IQR)
  + The range of the middle half of a distribution
* Standard deviation
  + Average distance from the mean
* Variance
  + The average of squared distances from the mean

# Why does variability matter?

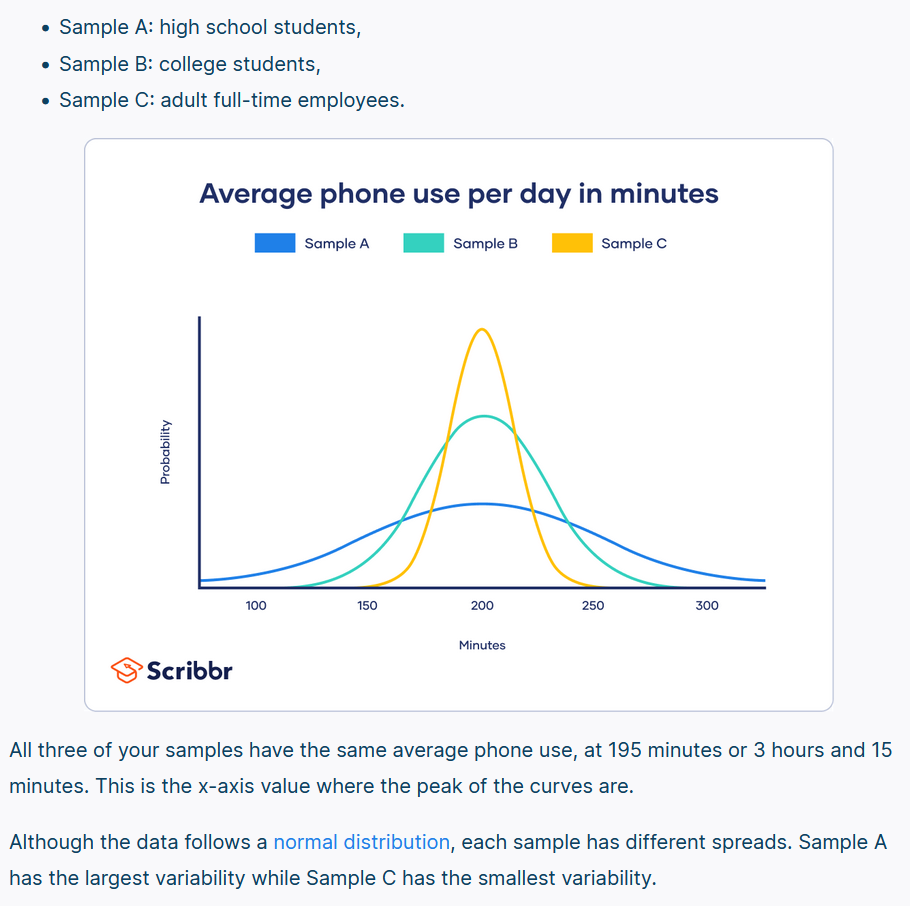
Central tendency tells you where MOST of you points lie

Variability summarizes how far apart they are.

This is important because it lets you know how well you can generalize results from the sample of your population

Low variability is ideal because it means that you can better predict information about the population based on sample data.

High variability means that the values are less consistent, meaning its hard to generalize/summarize the data in your population

****

# Range

The range tells you the spread of your data from the lowest to the highest values in the distribution

Sample set {72,110,134,190,238,287,305,324}

From the set the Range (R) is calculated in the notation

H = highest L = lowest

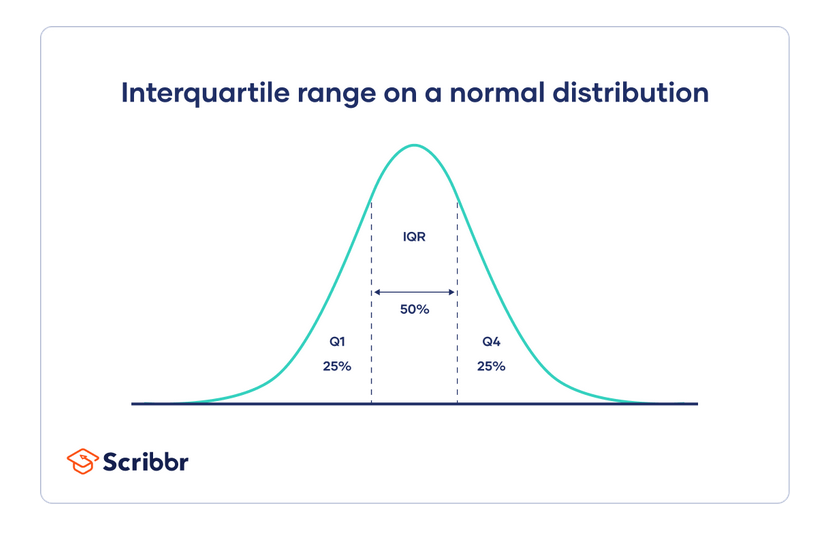
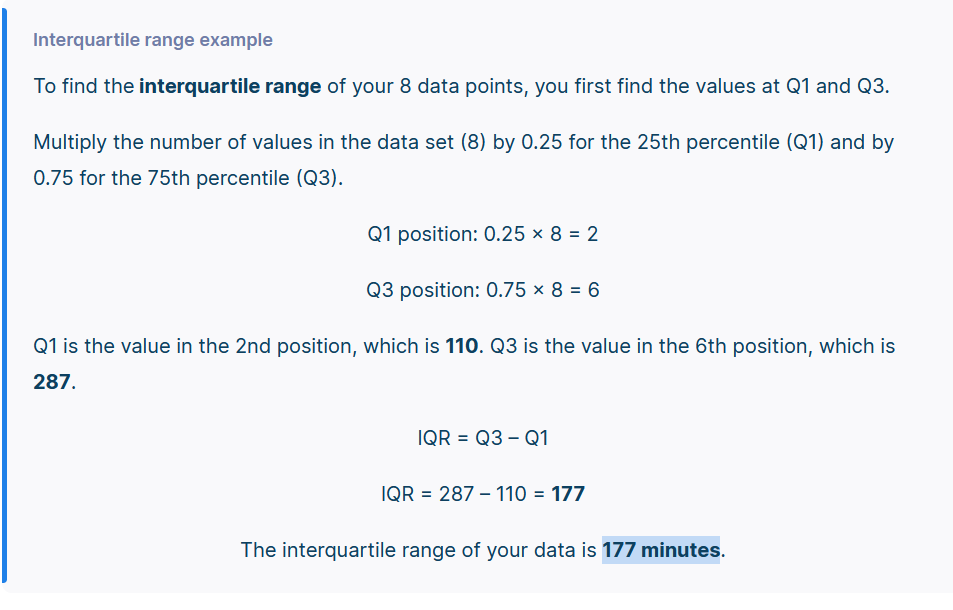
R = H – L

R= 324-72=252

Because only 2 numbers are use the range is influenced by outliers and does not give you any information about the distribution of values.

# Interquartile Range

IQR gives you the spread of the middle of your distribution

For any distribution that’s ordered from low to high, the IQR contains 50% of the values in the distribution. While the first quartile (Q1) contains the first 25% of values and fourth quartile gives you the last 25%the IQR is the third quartile(Q3) minus the first quartile(Q1). This gives us the range of the middle half of the datasetJust like the range, the IQR only use4s 2 values in its calculation.

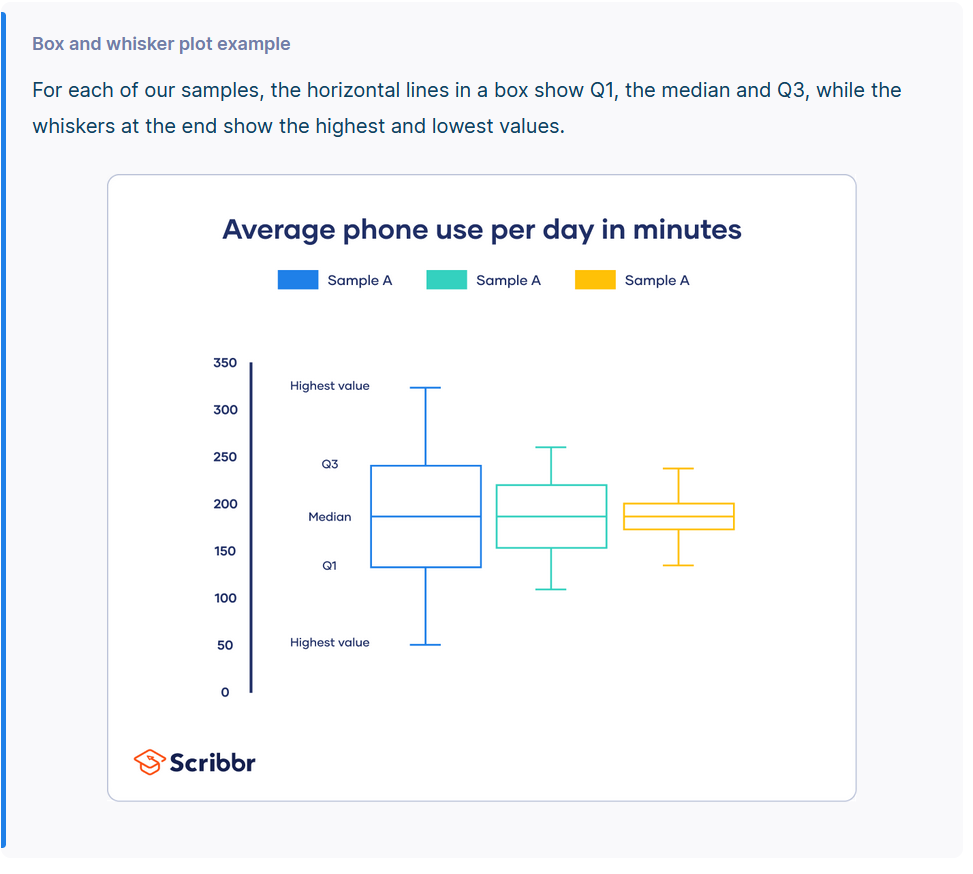
But the IQR is less affected by outliers: the 2 values come from the middle half of the data set so they are unlikely ot be extreme scrores

The IQR gives a consistent measure of variability for skewed as well as normal distributions.

## Five-number summary:

Every distribution can be organized using a fiove-number summary:

* Lowest value
* Q1:25th percentile
* Q2: the median
* Q3: 75th percentile
* Highest value (Q4)

These five number summaries can be easily visualized using box and whisker plots

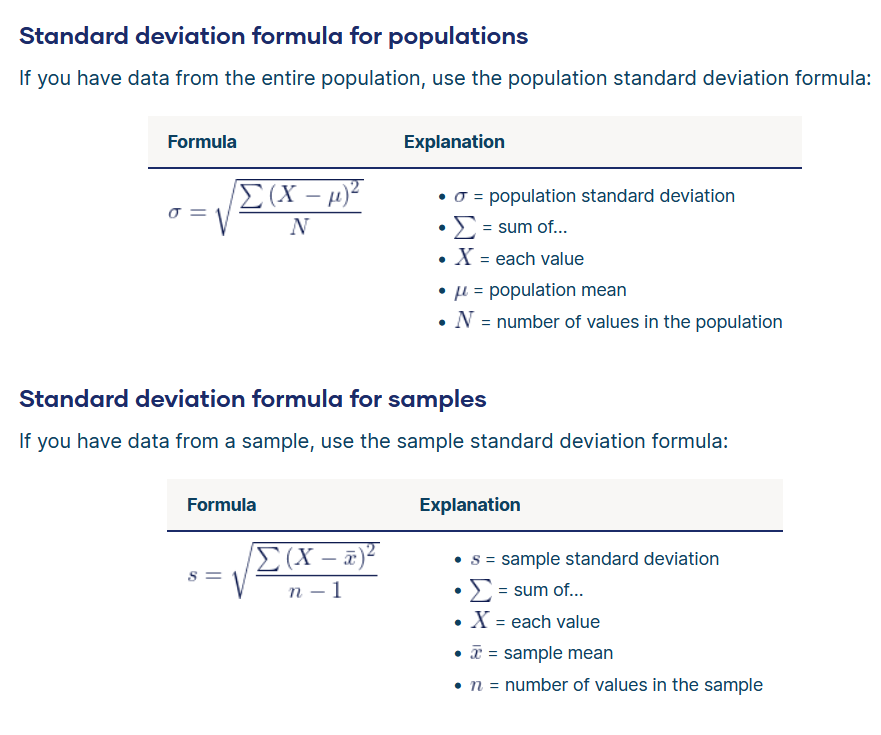
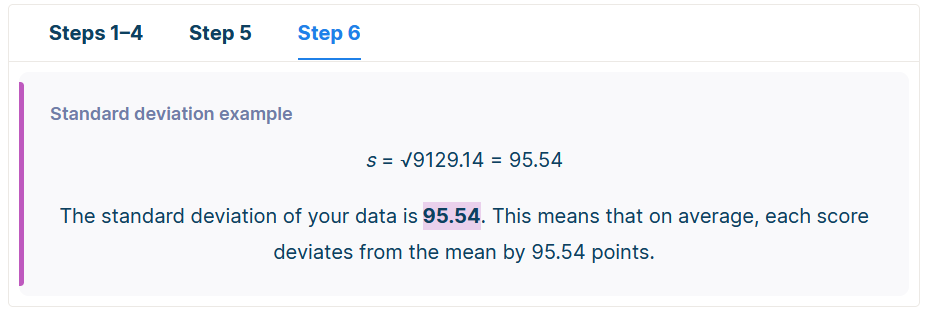
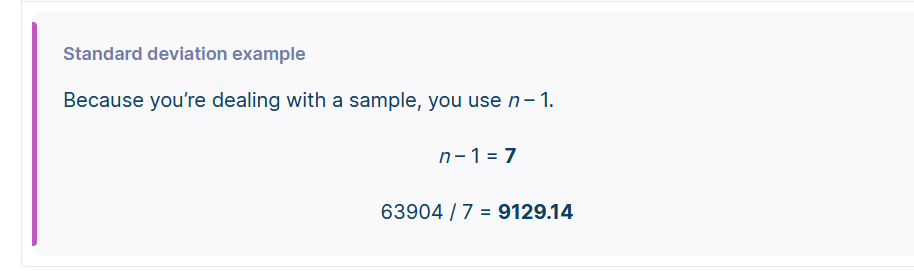
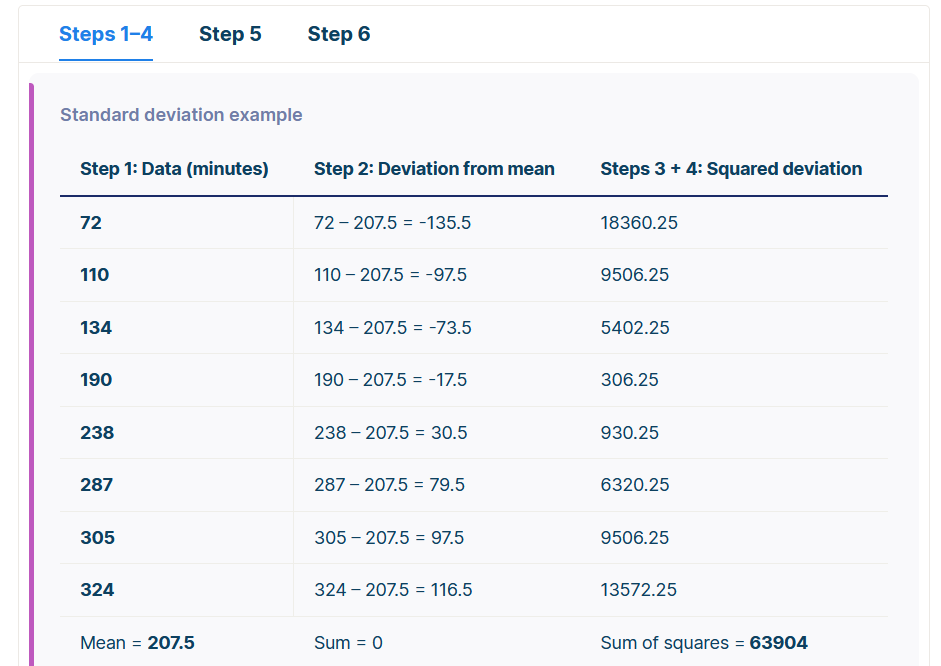
# Standard deviation

The standard deviation is the average amountof variability in your dataset.

It tells you on average how far each score lies from the mean. The larger the SD the more variable the data set is

There are six steps for finding the SD by hand:

1. List each score and find their mean
2. Subtract the mean from each score to get the deviation +/- away from the mean
3. Square each of these deviations
4. Add upp all the squared deviations
5. Divide the sum of the squared deviations by n-1(for a sample) or N(for a population
6. Find the square root of the number you found



## Why use n-1 for sample standard deviation?

reducing the sample n to n – 1 makes the standard deviation artificially large, giving you a conservative estimate of variability.

While this is not an unbiased estimate, it is a less biased estimate of standard deviation: it is better to overestimate rather than underestimate variability in samples.

The difference between biased and conservative estimates of standard deviation gets much smaller when you have a large sample size.

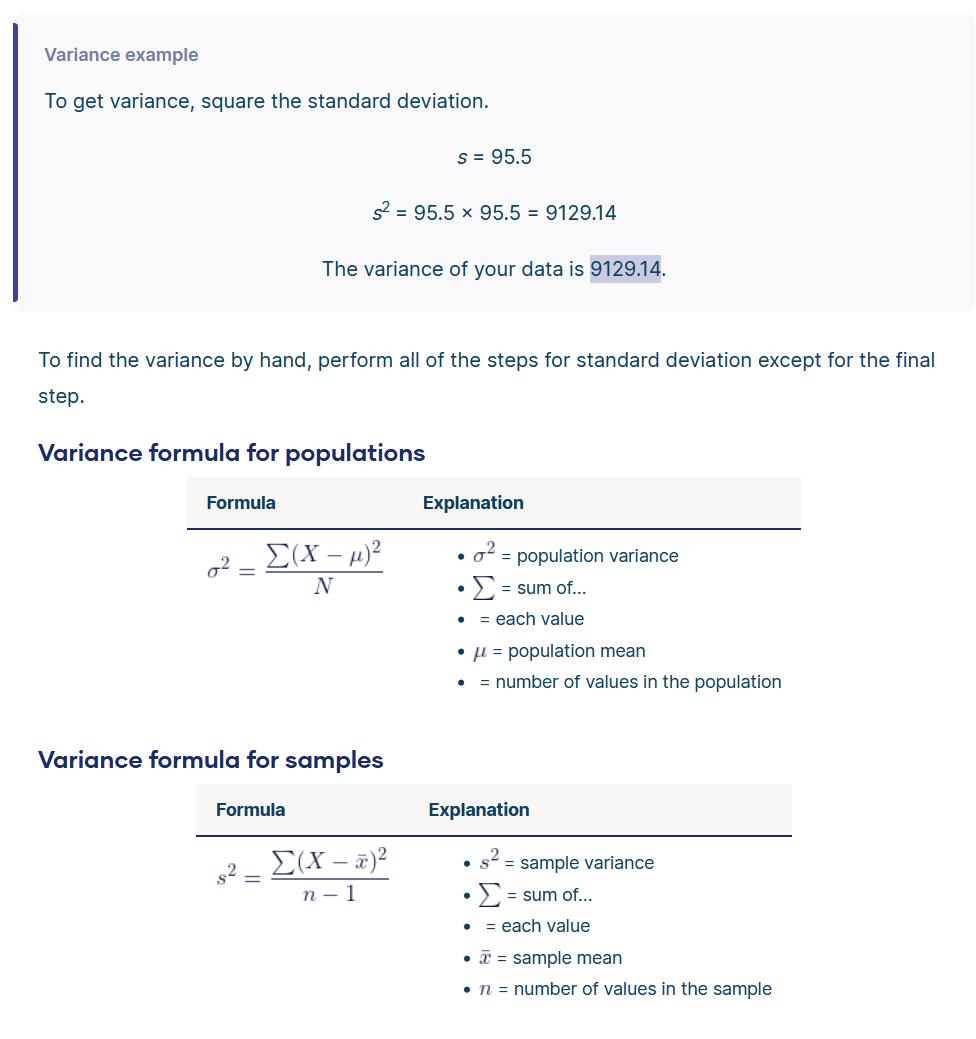
# Variance

Is he average of squared deviations from the mean. A deviation from the mean is how far a score lies from the mean

Variance is the square of the SD this means that the units of variance are much larger than those of a typical value of a data set

While it’s harder to interpret the variance number intuitively, it’s important to calculate variance for comparing different data set in statistical tests like ANOVAs

Variance reflects the degree of spread in the data set. The more spread the data the larger the variance is in relation to the mean



## Biased versus unbiased estimates of variance

An unbiased estimate in statistics is one that doesn’t consistently give you either high values or low values – it has no systematic bias.

Just like for standard deviation, there are different formulas for population and sample variance. But while there is no unbiased estimate for standard deviation, there is one for sample variance.

If the sample variance formula used the sample n, the sample variance would be biased towards lower numbers than expected. Reducing the sample n to n – 1 makes the variance artificially larger.

In this case, bias is not only lowered but totally removed. The sample variance formula gives completely unbiased estimates of variance.

So why isn’t the sample standard deviation also an unbiased estimate?

That’s because sample standard deviation comes from finding the square root of sample variance. Since a square root isn’t a linear operation, like addition or subtraction, the unbiasedness of the sample variance formula isn’t carried over the sample standard deviation formula.

# What’s the best measure of variability?

The best measure of variability depends on your level of measurement and distribution.

## Level of measurement

For data measured at an ordinal level, the range and interquartile range are the only appropriate measures of variability.

For more complex interval and ratio levels, the standard deviation and variance are also applicable.

## Distribution

For normal distributions, all measures can be used. The standard deviation and variance are preferred because they take your whole data set into account, but this also means that they are easily influenced by outliers.

For skewed distributions or data sets with outliers, the interquartile range is the best measure. It’s least affected by extreme values because it focuses on the spread in the middle of the data set.