Probability and Statistics

Evaluation

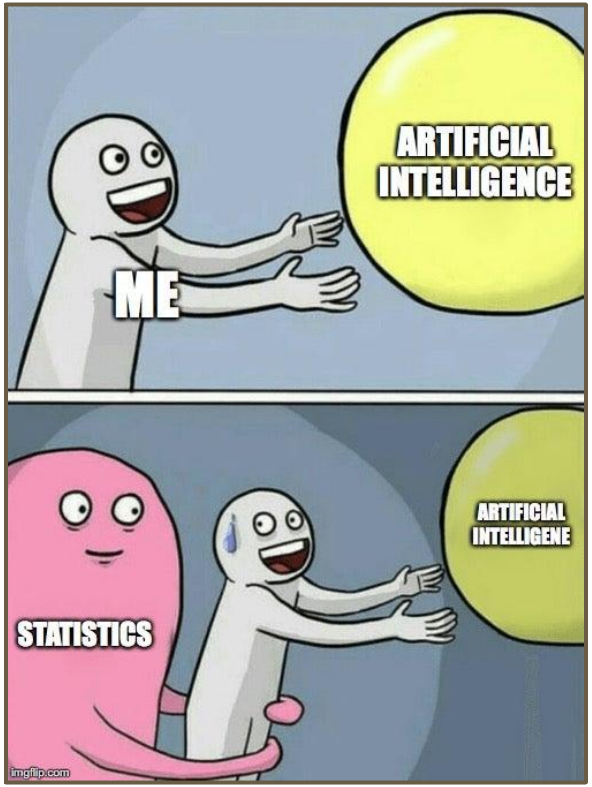
- 40% : Workshop Collection - “ *Week 1 to Week 12* “ **[Individual]**
- 60% : Unseen Written Examination **[Individual]**

Note: If you fail the exam you will fail the module.

Course Overview

- Descriptive Statistics
- Basics of Probability
- Discrete Probability Distributions
- Continuous Probability Distributions
- Inferential Statistics
- Regression and Correlation Analysis
- Bayesian Statistics

Why Probability & Statistics?



Week 1: Introduction to Basic Statistics

After the end of this week, You will know about:

- Basic Statistics
- Mean, Median, Mode, Central Tendency
- Five number Summary
- Variance & Standard Deviation

Basic Statistics

Statistics: a collection of mathematics used to summarize, analyze, and interpret a group of numbers or observations.

The basic statistics discussed in this unit are those most commonly used to describe collections of numerical values.

Understanding these concepts is a necessary foundation for both probabilistic and statistical calculations. These bases are prerequisite to all analysis of data.

Probability and Statistics are helpful tools. However, by themselves they cannot replace thoughtful experimental design, well-defined research questions, or the implementation of specific theories or models.

Many aspiring data scientists get caught up in the mathematics, or technological tools used by data scientists; however, an effective data scientist will understand that these are simply tools to answer complex questions.

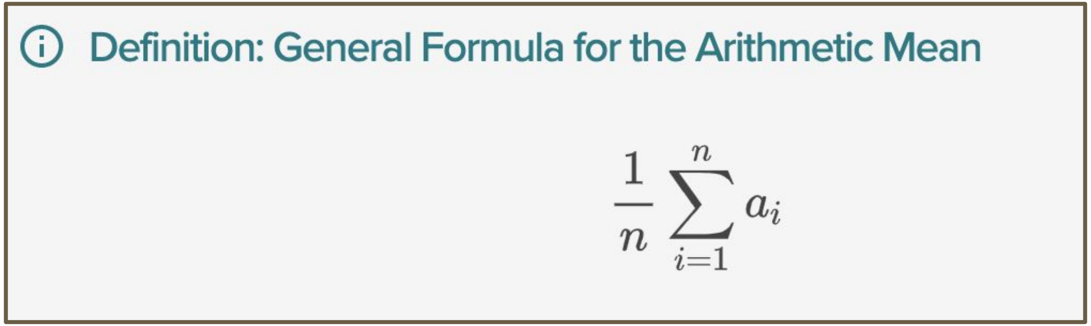
An effective data scientist understands that their most important role is **to ask the right questions**, the questions which will provide answers that can guide business, scientific, or research decisions and conclusions.

While the ability to ask the right questions is important, it is necessary to have the tools to answer those questions.

This unit will cover the key descriptive statistics: mean and median, mode, variance, standard deviation, and the five number summary.

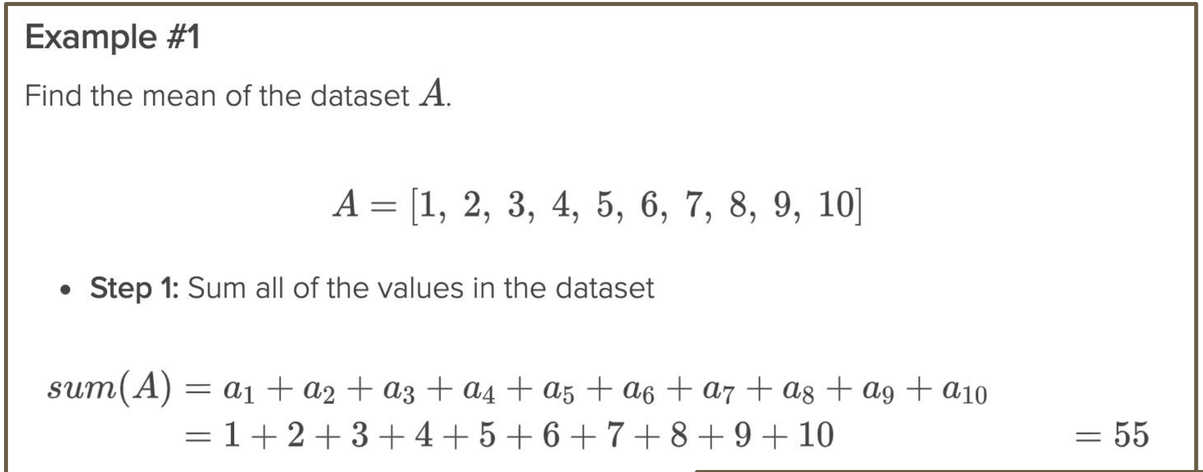
Mean

- The mean in statistics and probability is likely a familiar concept; the mean is commonly referred to as an average.
- A mean is derived by calculating a sum of all the values in a collection, then dividing that sum by the total number of items.



ⓘ Definition: General Formula for the Arithmetic Mean

$$\frac{1}{n}\sum_{i=1}^{n} a_i$$



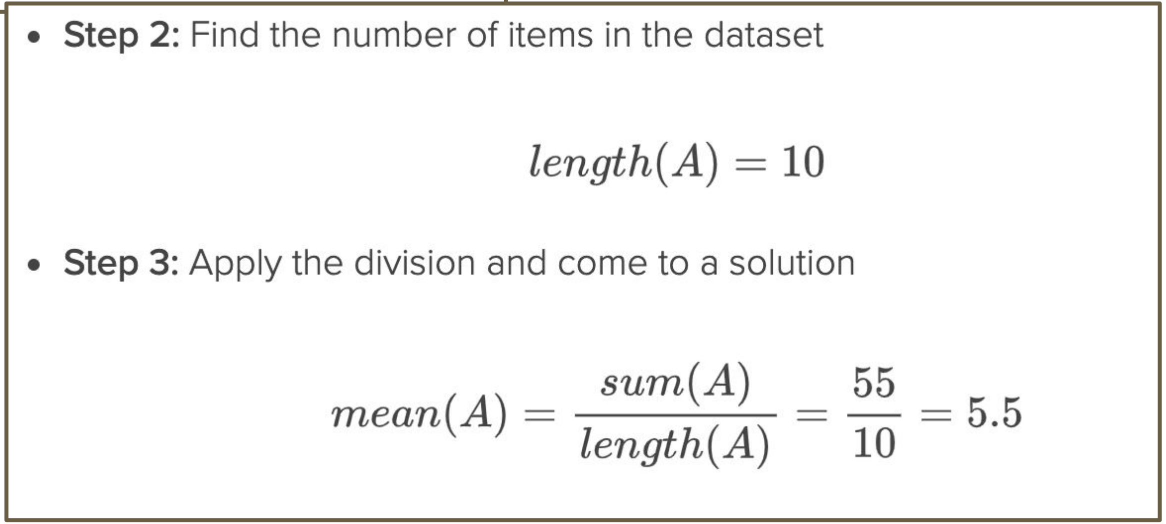
Example #1

Find the mean of the dataset A.

$$A = [1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10]$$

- **Step 1:** Sum all of the values in the dataset

$$\begin{aligned} sum(A) &= a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + a_9 + a_{10} \\ &= 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 \qquad = 55 \end{aligned}$$

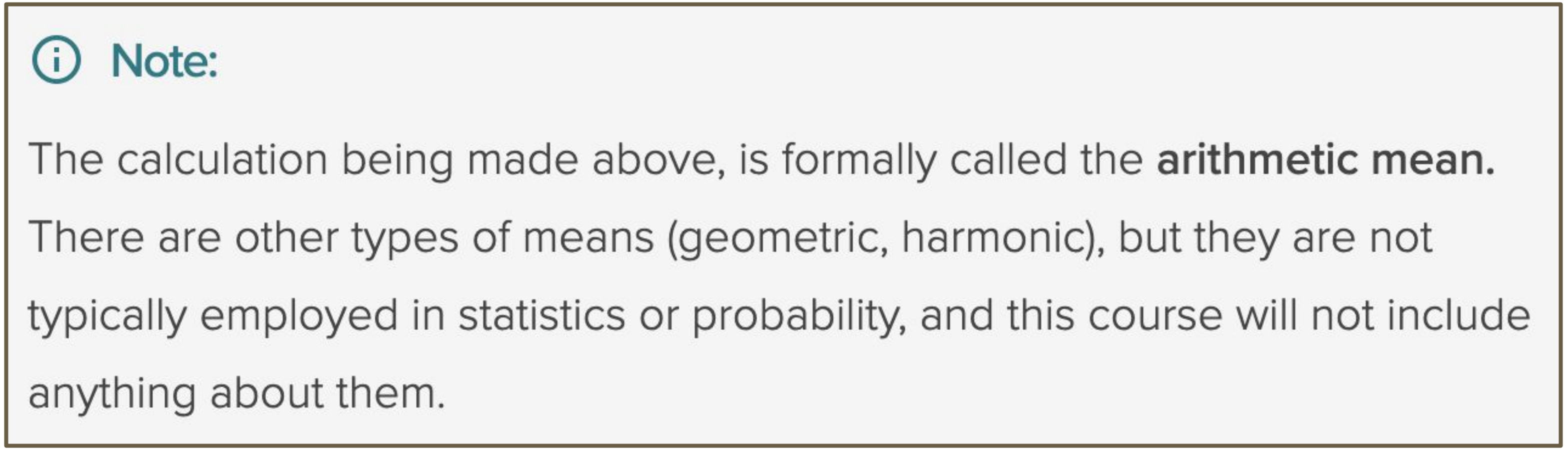


- **Step 2:** Find the number of items in the dataset

$$length(A) = 10$$

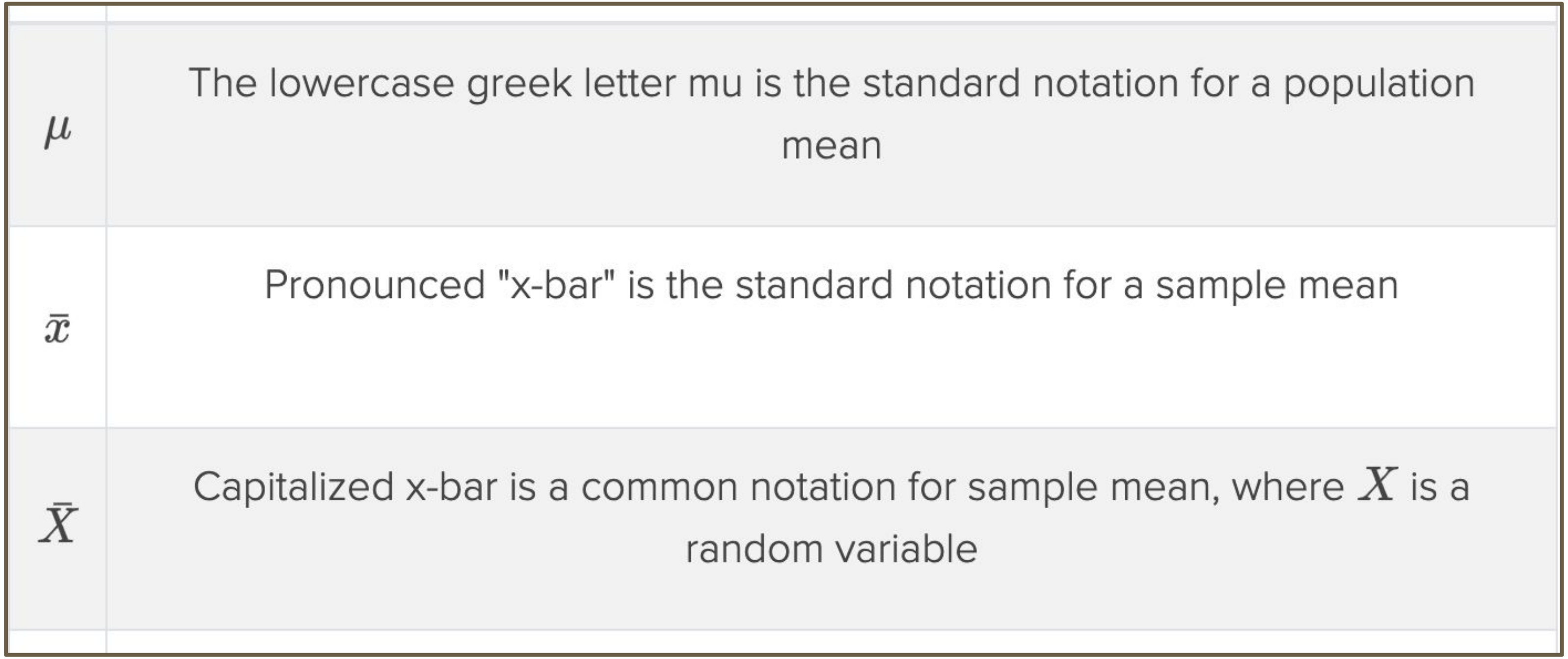
- **Step 3:** Apply the division and come to a solution

$$mean(A) = \frac{sum(A)}{length(A)} = \frac{55}{10} = 5.5$$



Note:

The calculation being made above, is formally called the **arithmetic mean.** There are other types of means (geometric, harmonic), but they are not typically employed in statistics or probability, and this course will not include anything about them.



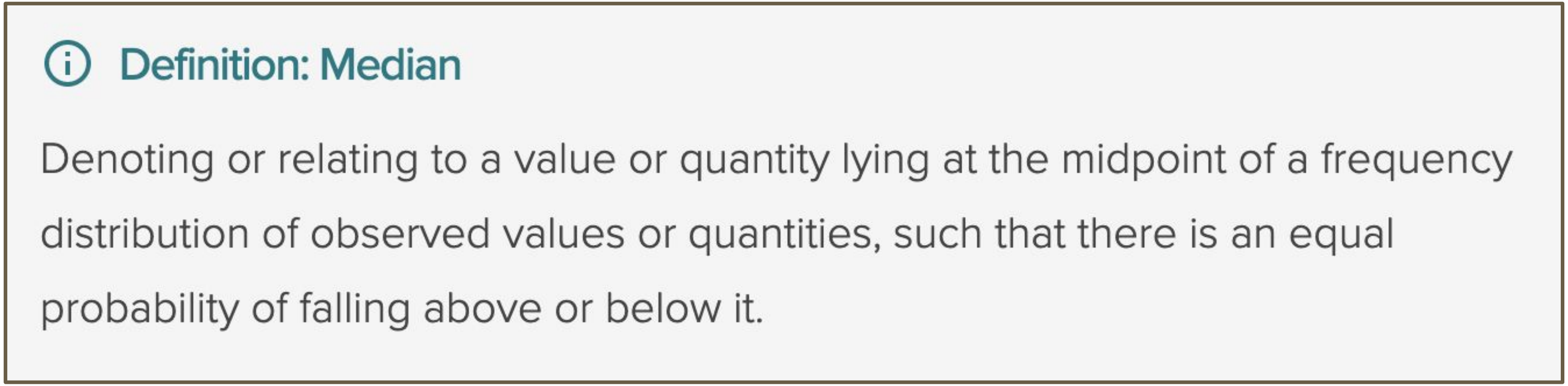
μ	The lowercase greek letter mu is the standard notation for a population mean
$\bar{x}$	Pronounced "x-bar" is the standard notation for a sample mean
$\bar{X}$	Capitalized x-bar is a common notation for sample mean, where X is a random variable

Population vs Sample…?

- *Mean* related mathematical challenges : **Tutorial**
- Finding the *Mean* using *Python*: **Workshop**

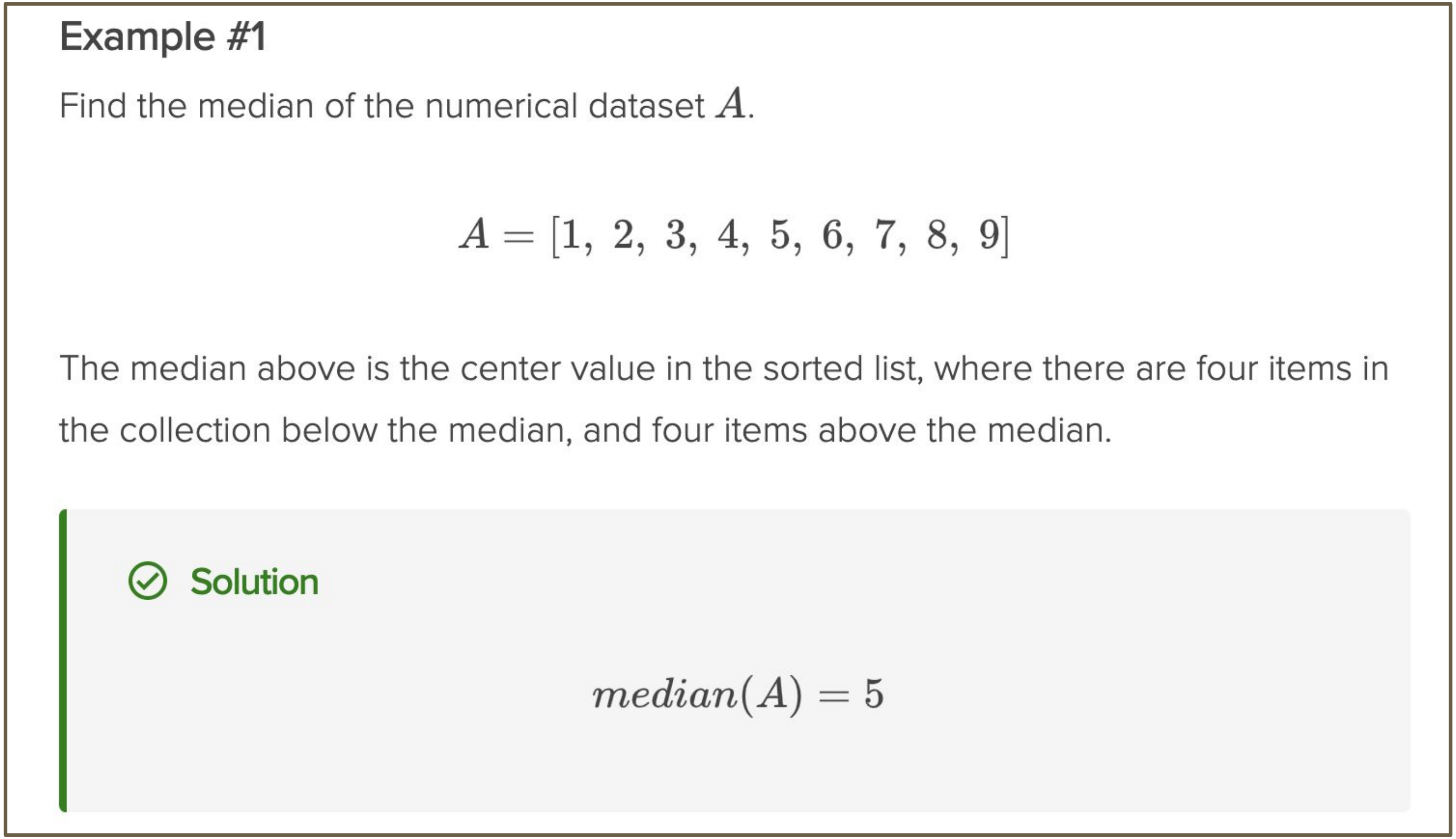
Median

- The median is another measure of central tendency. It can be considered the "middle" value of some sorted numerical collection.
- Half of the collection is equal to or lesser than the median, and half of the collection is equal to or greater than the median.
- In circumstances where a collection has extreme outliers the median can be a more robust, or superior measure to the mean.
- More comparisons between the mean and median are made in the next lesson.



Definition: Median

Denoting or relating to a value or quantity lying at the midpoint of a frequency distribution of observed values or quantities, such that there is an equal probability of falling above or below it.



Example #1

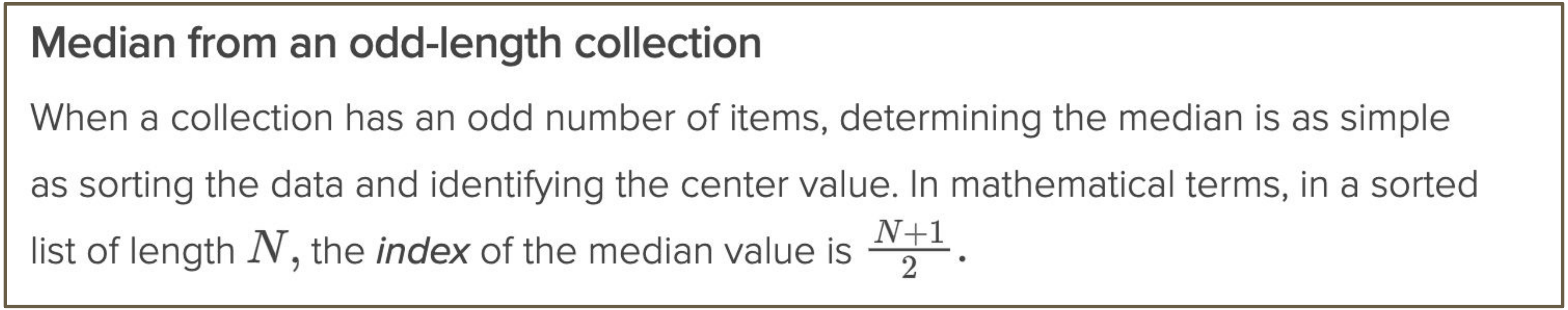
Find the median of the numerical dataset A.

$$A = [1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9]$$

The median above is the center value in the sorted list, where there are four items in the collection below the median, and four items above the median.

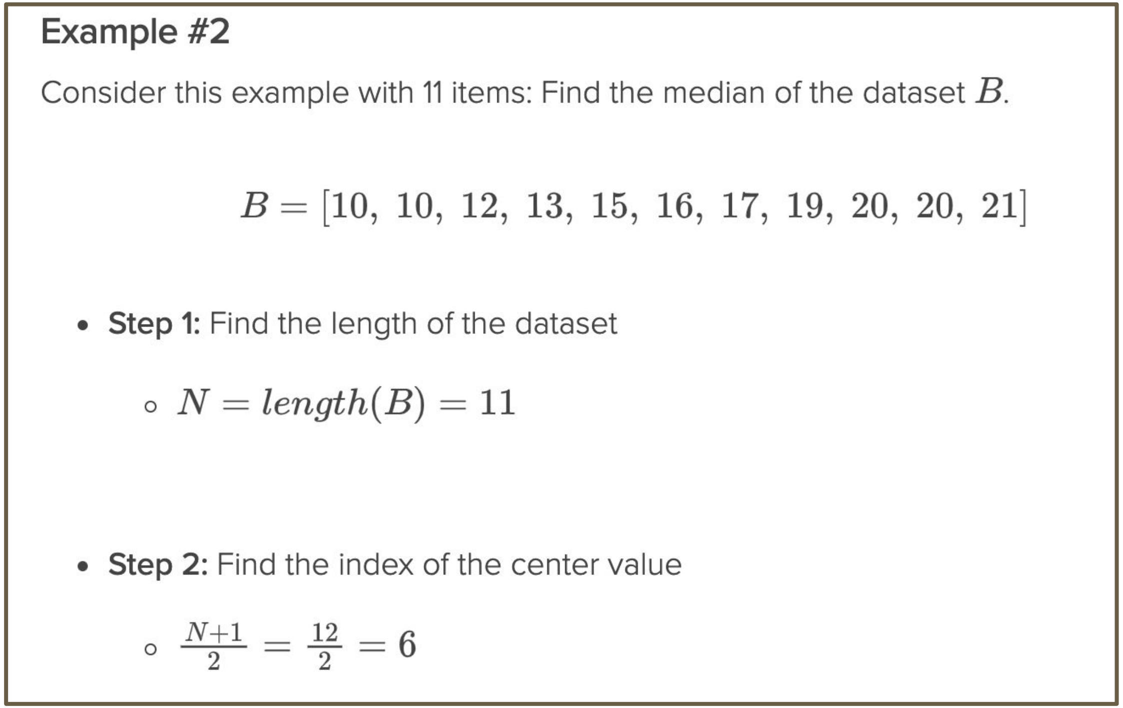
Solution

$$median(A) = 5$$



Median from an odd-length collection

When a collection has an odd number of items, determining the median is as simple as sorting the data and identifying the center value. In mathematical terms, in a sorted list of length N, the *index* of the median value is $\frac{N+1}{2}$.

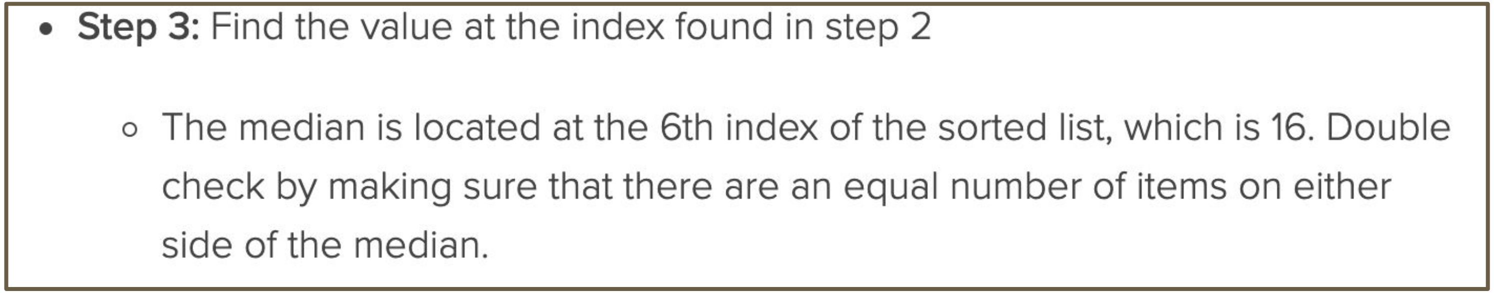


Example #2

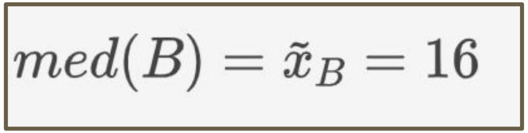
Consider this example with 11 items: Find the median of the dataset B.

$$B = [10,\ 10,\ 12,\ 13,\ 15,\ 16,\ 17,\ 19,\ 20,\ 20,\ 21]$$

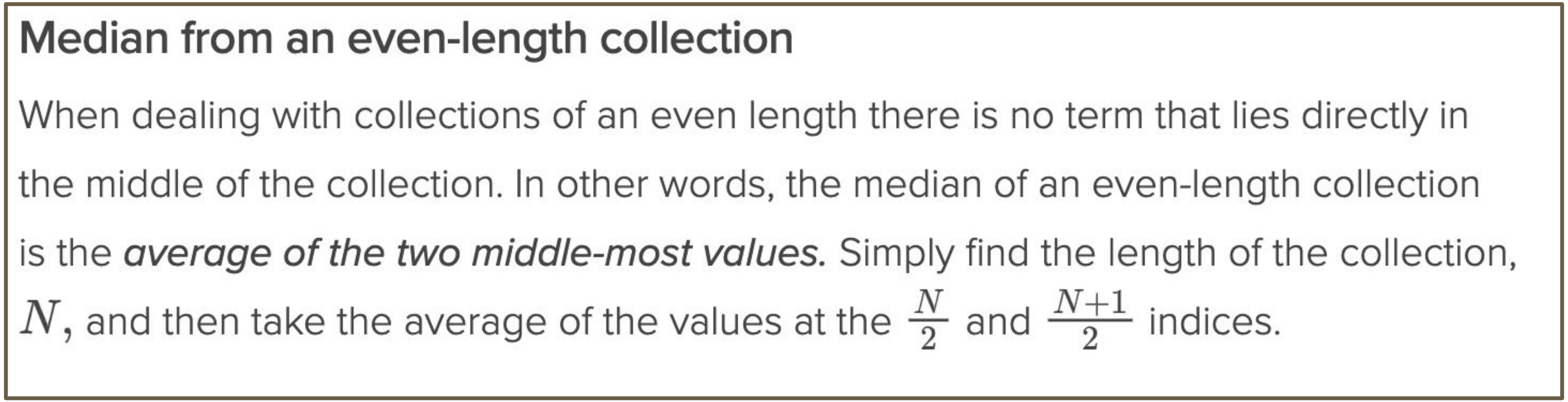
- **Step 1:** Find the length of the dataset
  - $N = length(B) = 11$
- **Step 2:** Find the index of the center value
  - $\frac{N+1}{2} = \frac{12}{2} = 6$



- **Step 3:** Find the value at the index found in step 2
  - The median is located at the 6th index of the sorted list, which is 16. Double check by making sure that there are an equal number of items on either side of the median.

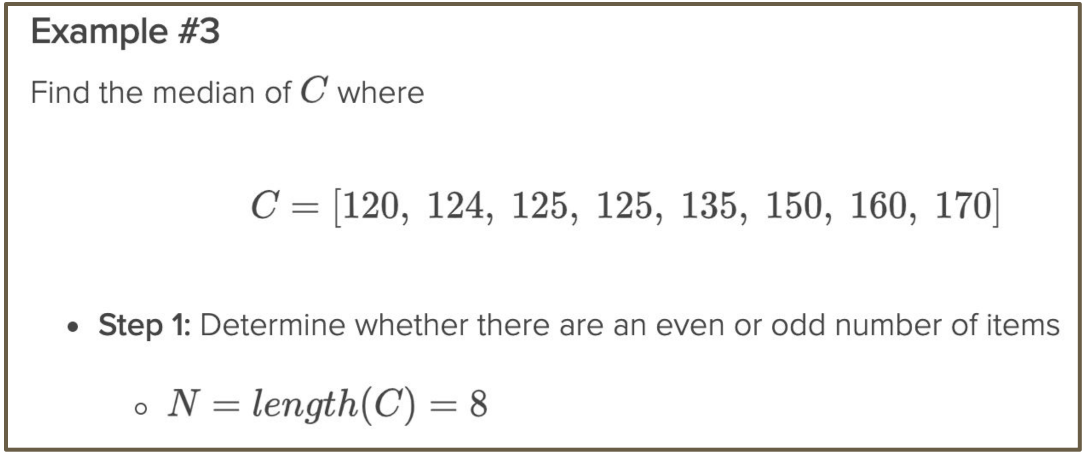


$$med(B) = \tilde{x}_B = 16$$



Median from an even-length collection

When dealing with collections of an even length there is no term that lies directly in the middle of the collection. In other words, the median of an even-length collection is the ***average of the two middle-most values.*** Simply find the length of the collection, N, and then take the average of the values at the $\frac{N}{2}$ and $\frac{N+1}{2}$ indices.

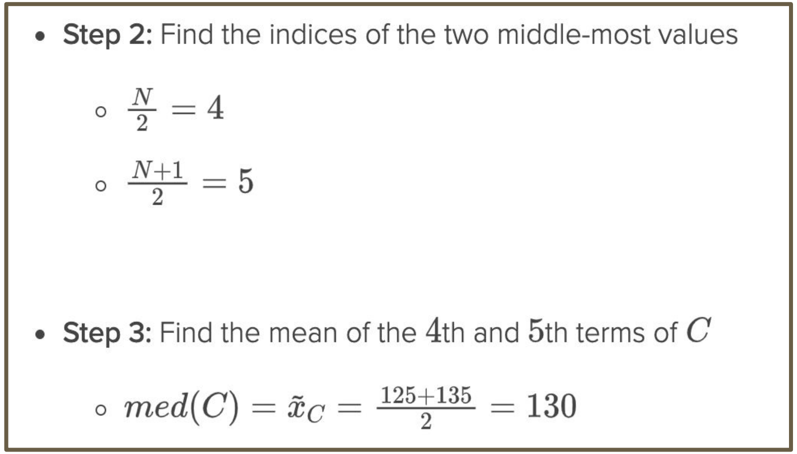


Example #3

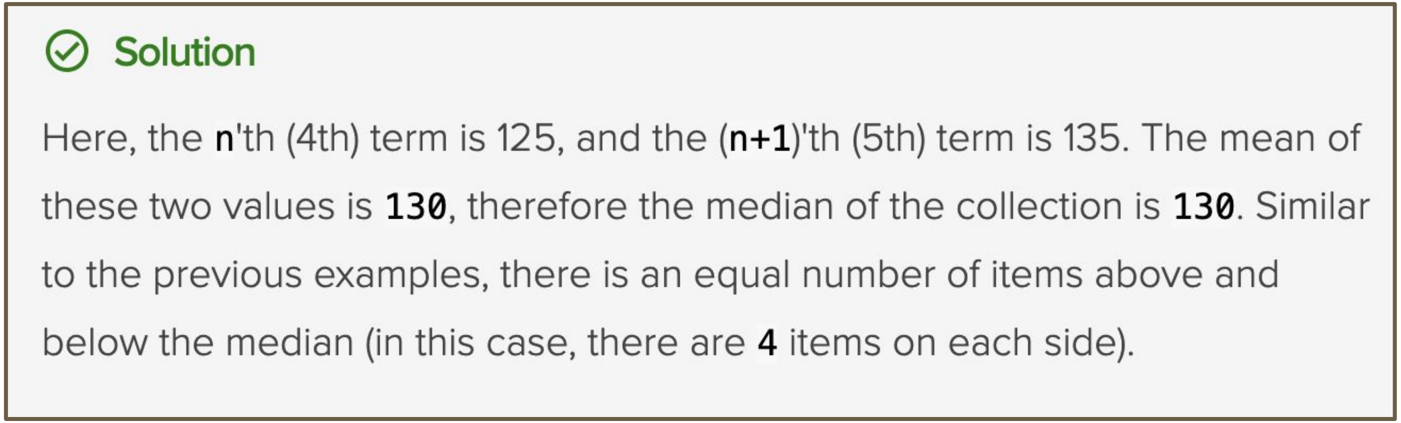
Find the median of C where

$$C = [120,\ 124,\ 125,\ 125,\ 135,\ 150,\ 160,\ 170]$$

- **Step 1:** Determine whether there are an even or odd number of items
  - $N = length(C) = 8$

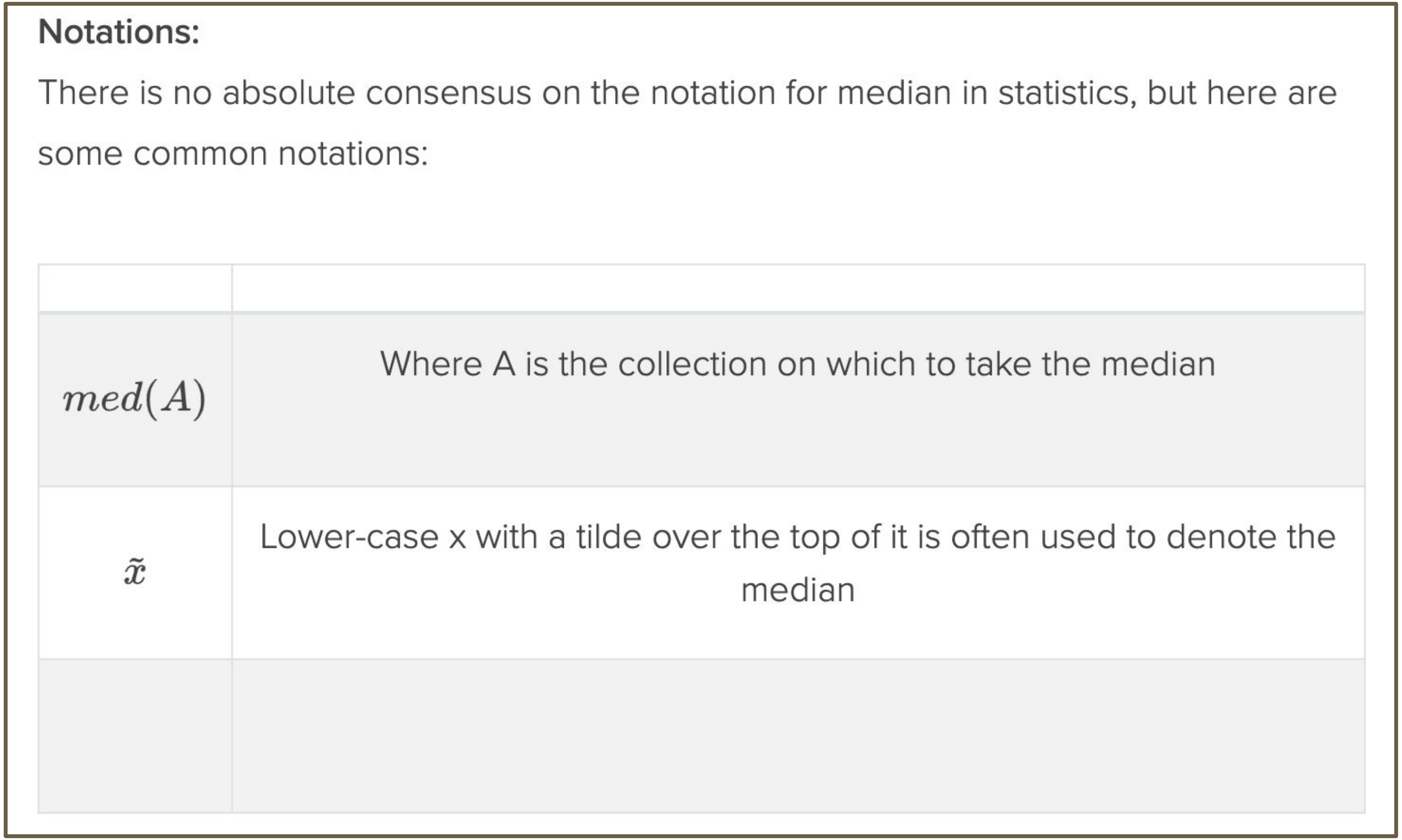


- **Step 2:** Find the indices of the two middle-most values
  - $\frac{N}{2} = 4$
  - $\frac{N+1}{2} = 5$
- **Step 3:** Find the mean of the 4th and 5th terms of C
  - $med(C) = \tilde{x}_C = \frac{125+135}{2} = 130$



Solution

Here, the `n`'th (4th) term is 125, and the (`n+1`)'th (5th) term is 135. The mean of these two values is `130`, therefore the median of the collection is `130`. Similar to the previous examples, there is an equal number of items above and below the median (in this case, there are `4` items on each side).



Notations:

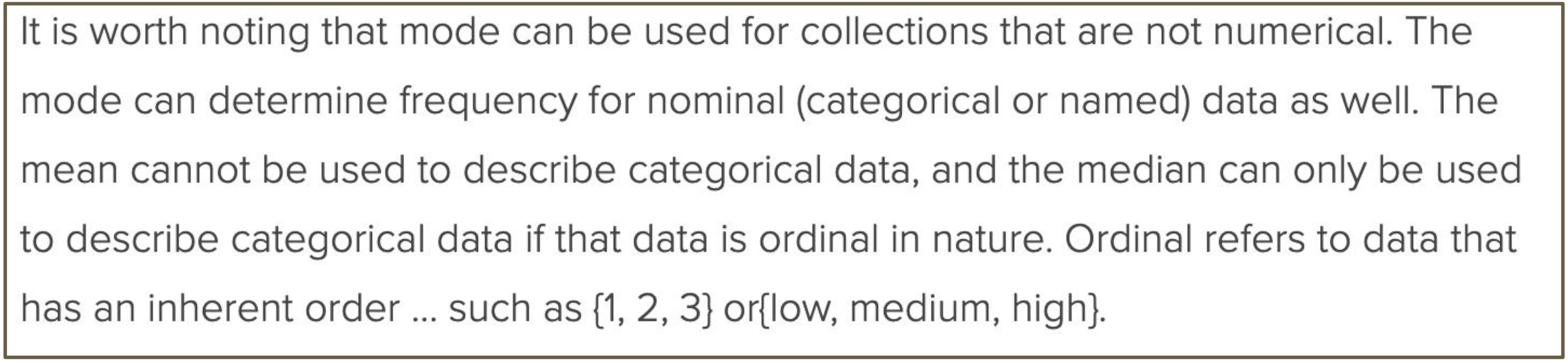
There is no absolute consensus on the notation for median in statistics, but here are some common notations:

$med(A)$	Where A is the collection on which to take the median
$\tilde{x}$	Lower-case x with a tilde over the top of it is often used to denote the median

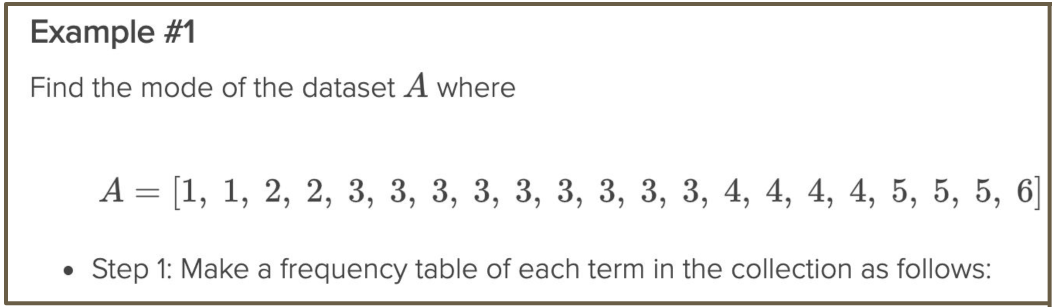
- Find the *Median* Using *Python*: **Workshop**
- *Median* related mathematical challenges: **Tutorial**

Mode

- The mode of a numerical collection is a different approach than mean or median. Instead of finding the center of a collection, the mode seeks to find the item with the greatest frequency.
- There are situations in which the mode may do a better job of describing a particular collection than a mean or median could based on the characteristics of the distribution.



It is worth noting that mode can be used for collections that are not numerical. The mode can determine frequency for nominal (categorical or named) data as well. The mean cannot be used to describe categorical data, and the median can only be used to describe categorical data if that data is ordinal in nature. Ordinal refers to data that has an inherent order … such as {1, 2, 3} or{low, medium, high}.

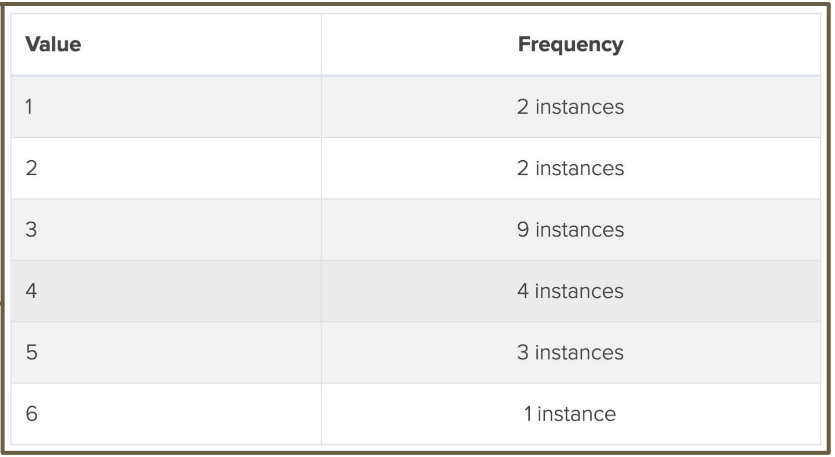


Example #1

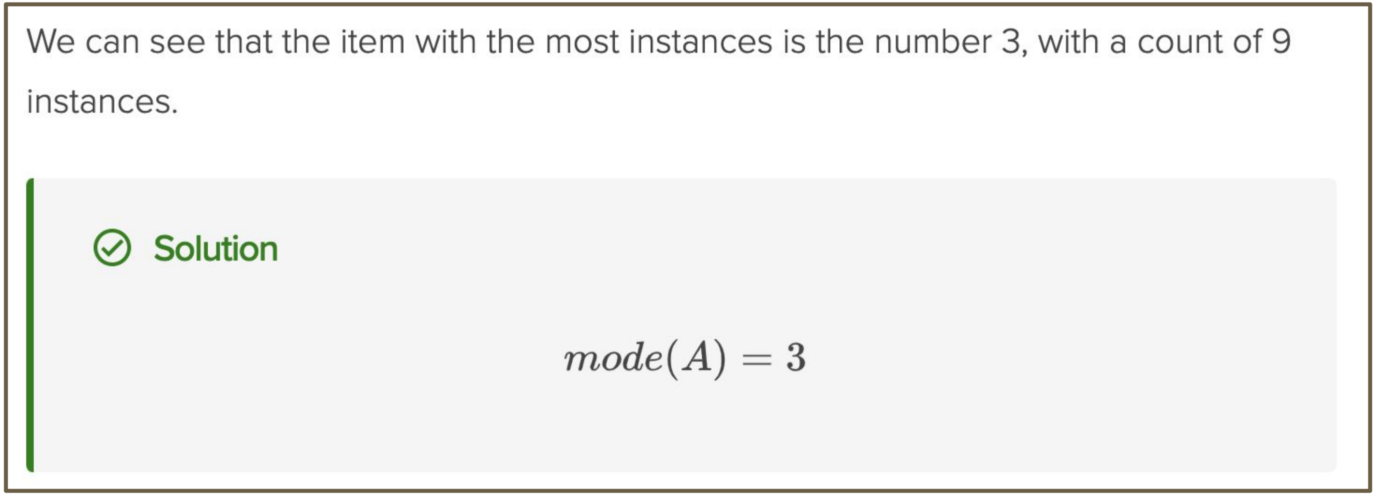
Find the mode of the dataset A where

$$A = [1,\ 1,\ 2,\ 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4,\ 4,\ 5,\ 5,\ 5,\ 6]$$

- Step 1: Make a frequency table of each term in the collection as follows:



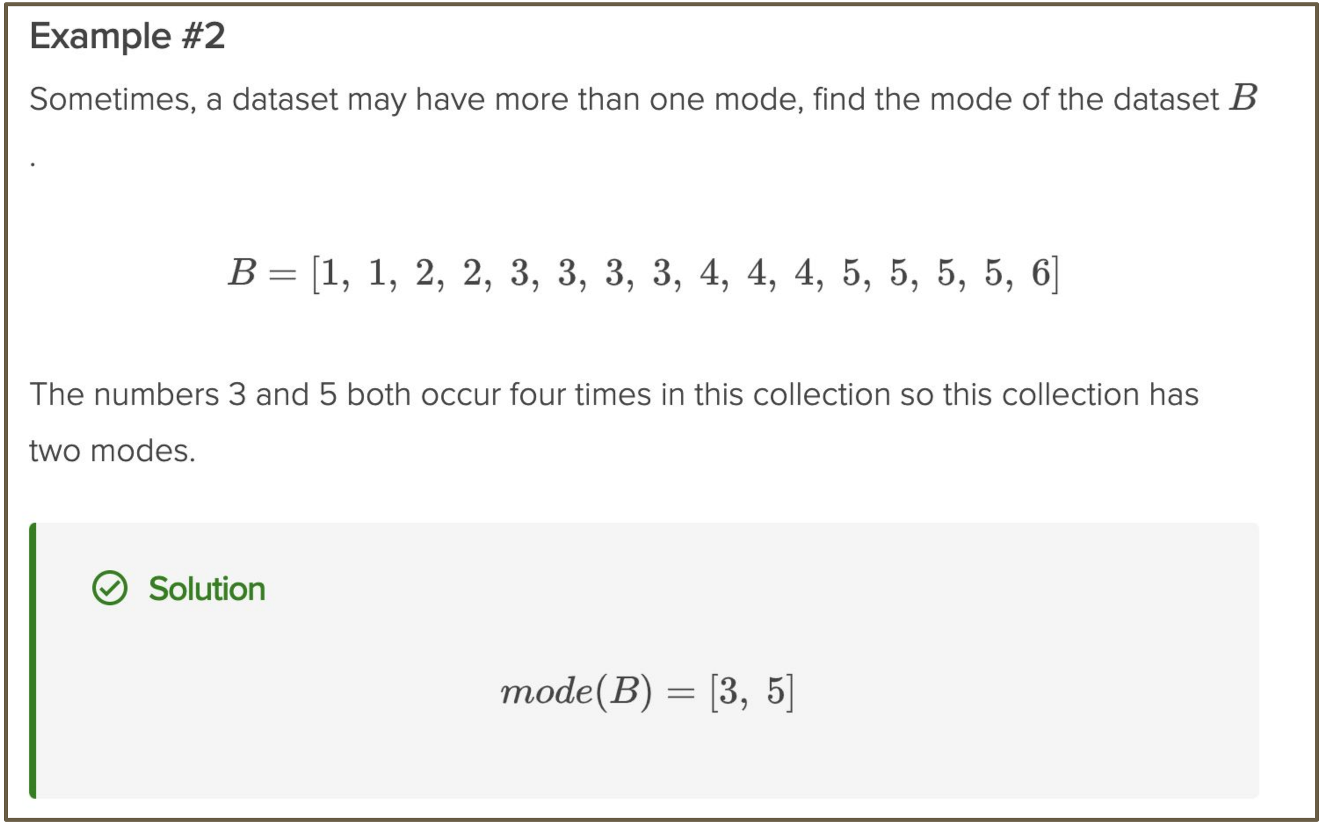
Value	Frequency
1	2 instances
2	2 instances
3	9 instances
4	4 instances
5	3 instances
6	1 instance



We can see that the item with the most instances is the number 3, with a count of 9 instances.

Solution

$$mode(A) = 3$$



Example #2

Sometimes, a dataset may have more than one mode, find the mode of the dataset B

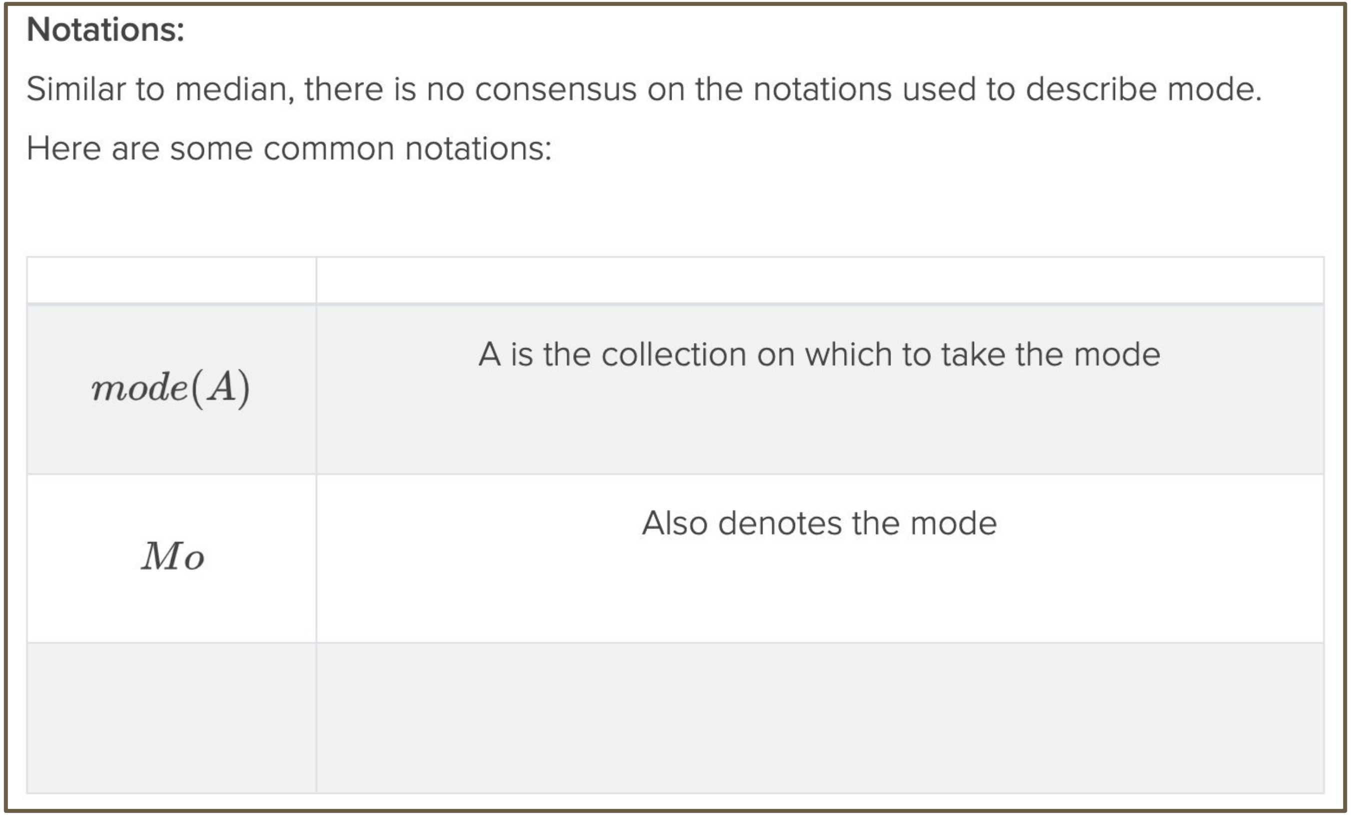
.

$$B = [1,\ 1,\ 2,\ 2,\ 3,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4,\ 5,\ 5,\ 5,\ 5,\ 6]$$

The numbers 3 and 5 both occur four times in this collection so this collection has two modes.

✓ Solution

$$mode(B) = [3,\ 5]$$



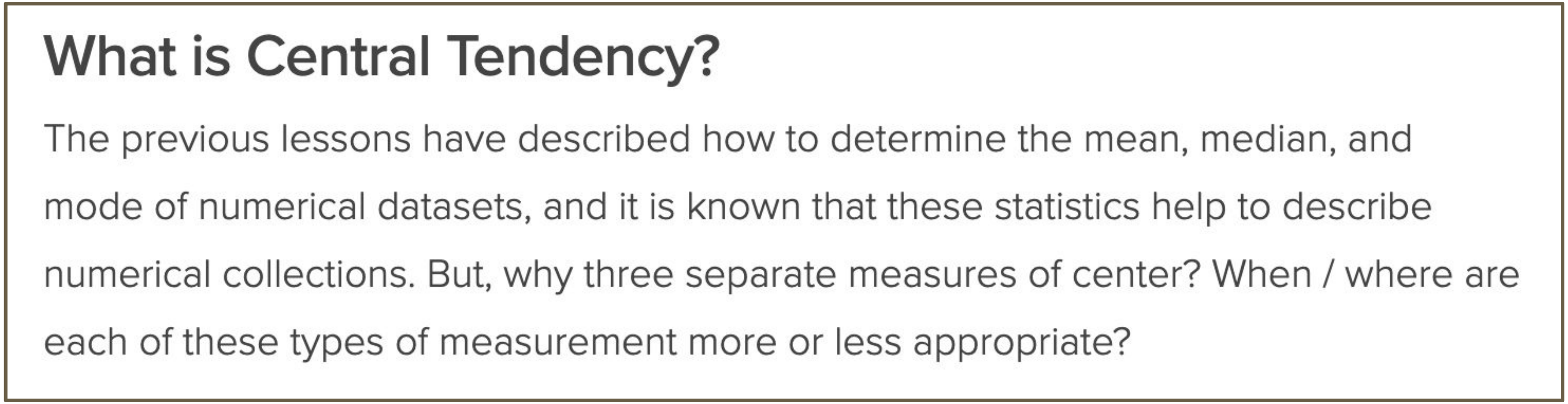
Notations:

Similar to median, there is no consensus on the notations used to describe mode.

Here are some common notations:

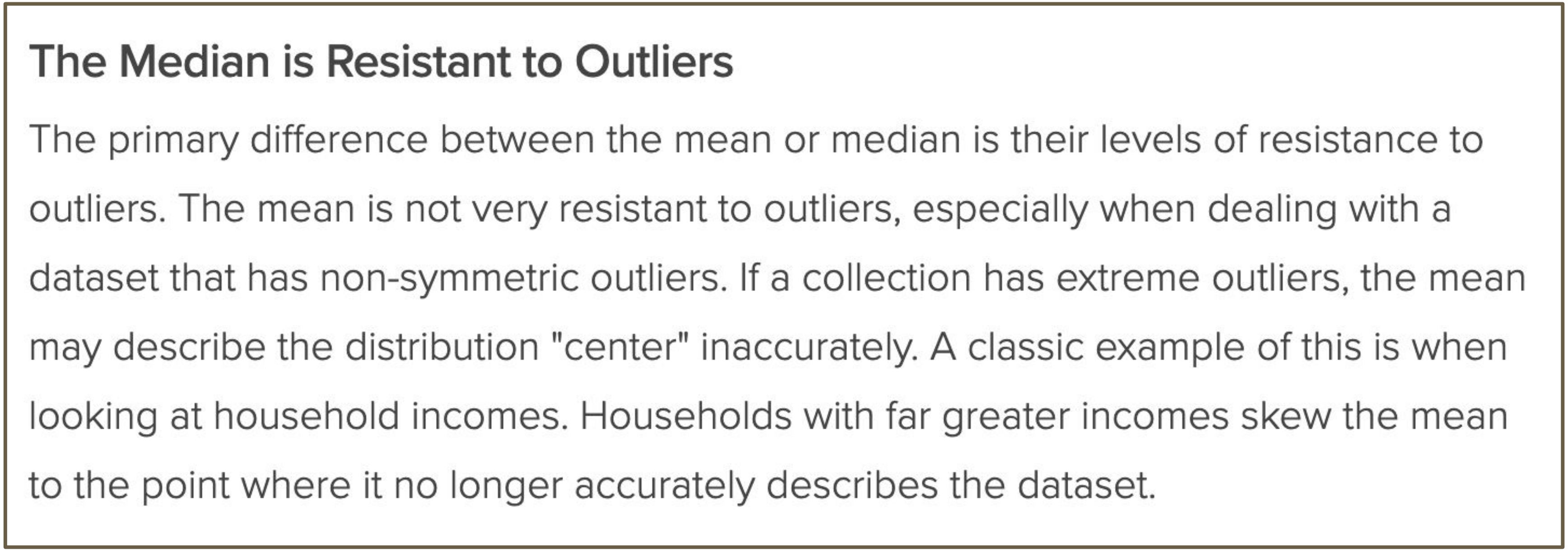
$mode(A)$	A is the collection on which to take the mode
Mo	Also denotes the mode

- Find the *Mode* Using *Python*: **Workshop**
- *Mode* related mathematical challenges: **Tutorial**



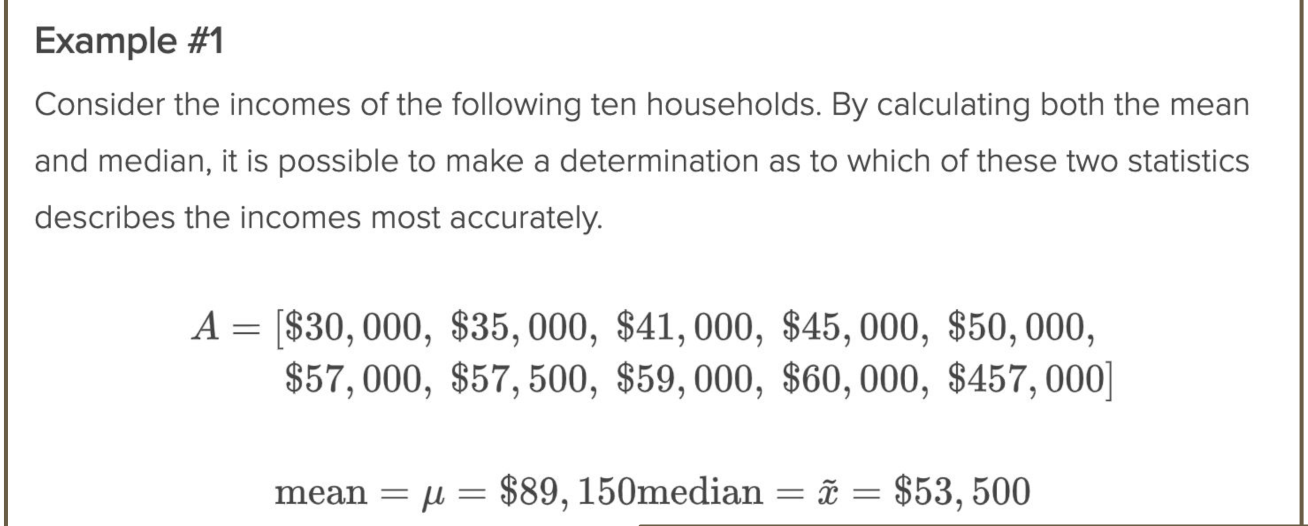
What is Central Tendency?

The previous lessons have described how to determine the mean, median, and mode of numerical datasets, and it is known that these statistics help to describe numerical collections. But, why three separate measures of center? When / where are each of these types of measurement more or less appropriate?



The Median is Resistant to Outliers

The primary difference between the mean or median is their levels of resistance to outliers. The mean is not very resistant to outliers, especially when dealing with a dataset that has non-symmetric outliers. If a collection has extreme outliers, the mean may describe the distribution "center" inaccurately. A classic example of this is when looking at household incomes. Households with far greater incomes skew the mean to the point where it no longer accurately describes the dataset.

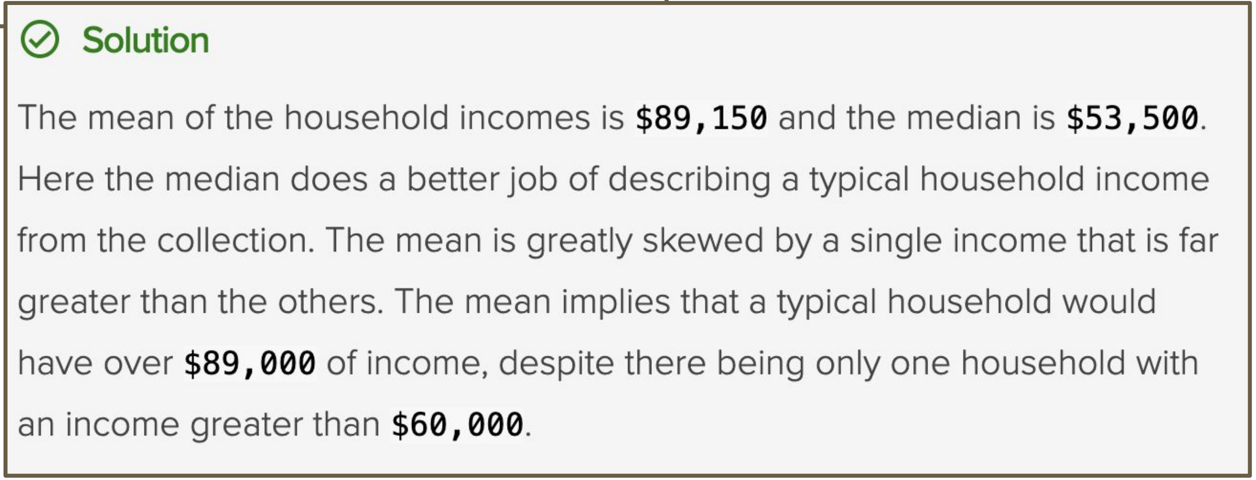


Example #1

Consider the incomes of the following ten households. By calculating both the mean and median, it is possible to make a determination as to which of these two statistics describes the incomes most accurately.

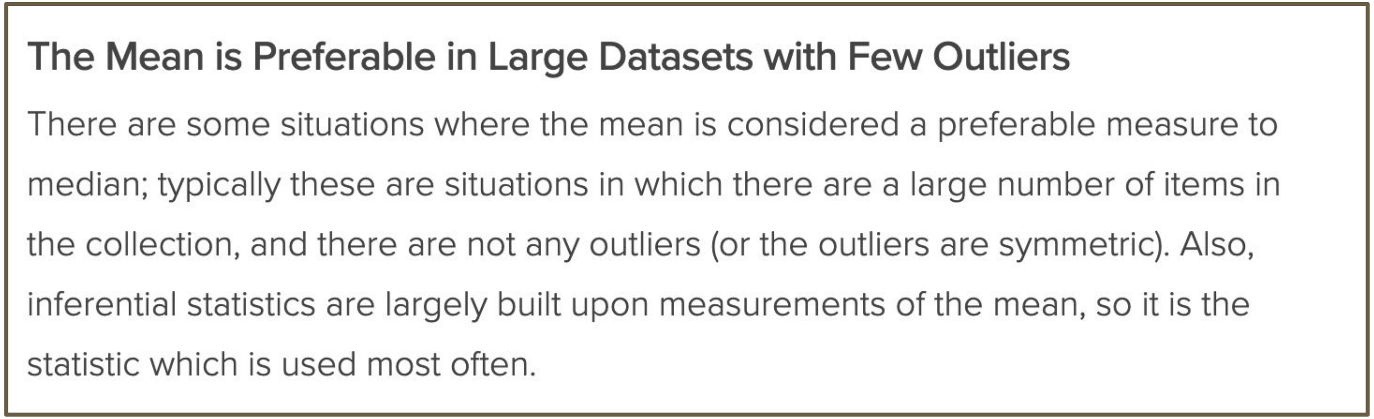
$$A = [\$30,000,\ \$35,000,\ \$41,000,\ \$45,000,\ \$50,000,\\ \$57,000,\ \$57,500,\ \$59,000,\ \$60,000,\ \$457,000]$$

$$\text{mean} = \mu = \$89,150 \text{median} = \tilde{x} = \$53,500$$



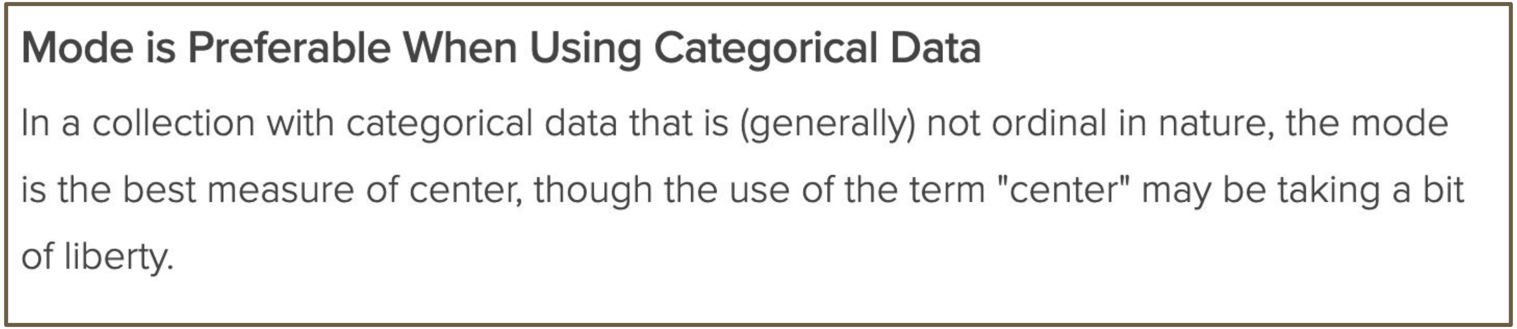
Solution

The mean of the household incomes is `$89,150` and the median is `$53,500`. Here the median does a better job of describing a typical household income from the collection. The mean is greatly skewed by a single income that is far greater than the others. The mean implies that a typical household would have over `$89,000` of income, despite there being only one household with an income greater than `$60,000`.



The Mean is Preferable in Large Datasets with Few Outliers

There are some situations where the mean is considered a preferable measure to median; typically these are situations in which there are a large number of items in the collection, and there are not any outliers (or the outliers are symmetric). Also, inferential statistics are largely built upon measurements of the mean, so it is the statistic which is used most often.

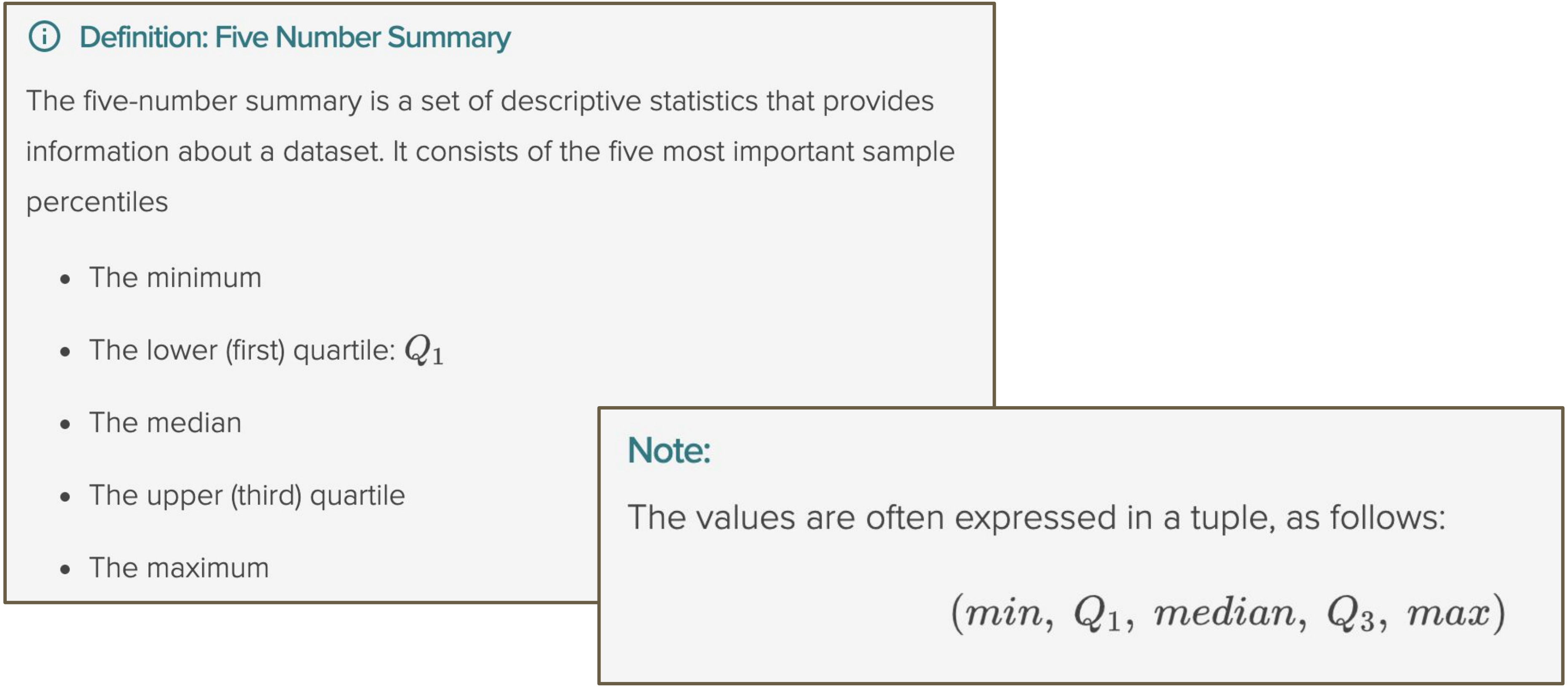


Mode is Preferable When Using Categorical Data

In a collection with categorical data that is (generally) not ordinal in nature, the mode is the best measure of center, though the use of the term "center" may be taking a bit of liberty.

Five Number Summary

- The five number summary gives a more in-depth description of a numerical collection of values.
- In addition to identifying a measure of center (median), it gives us more insight into the way the values are distributed.



ⓘ Definition: Five Number Summary

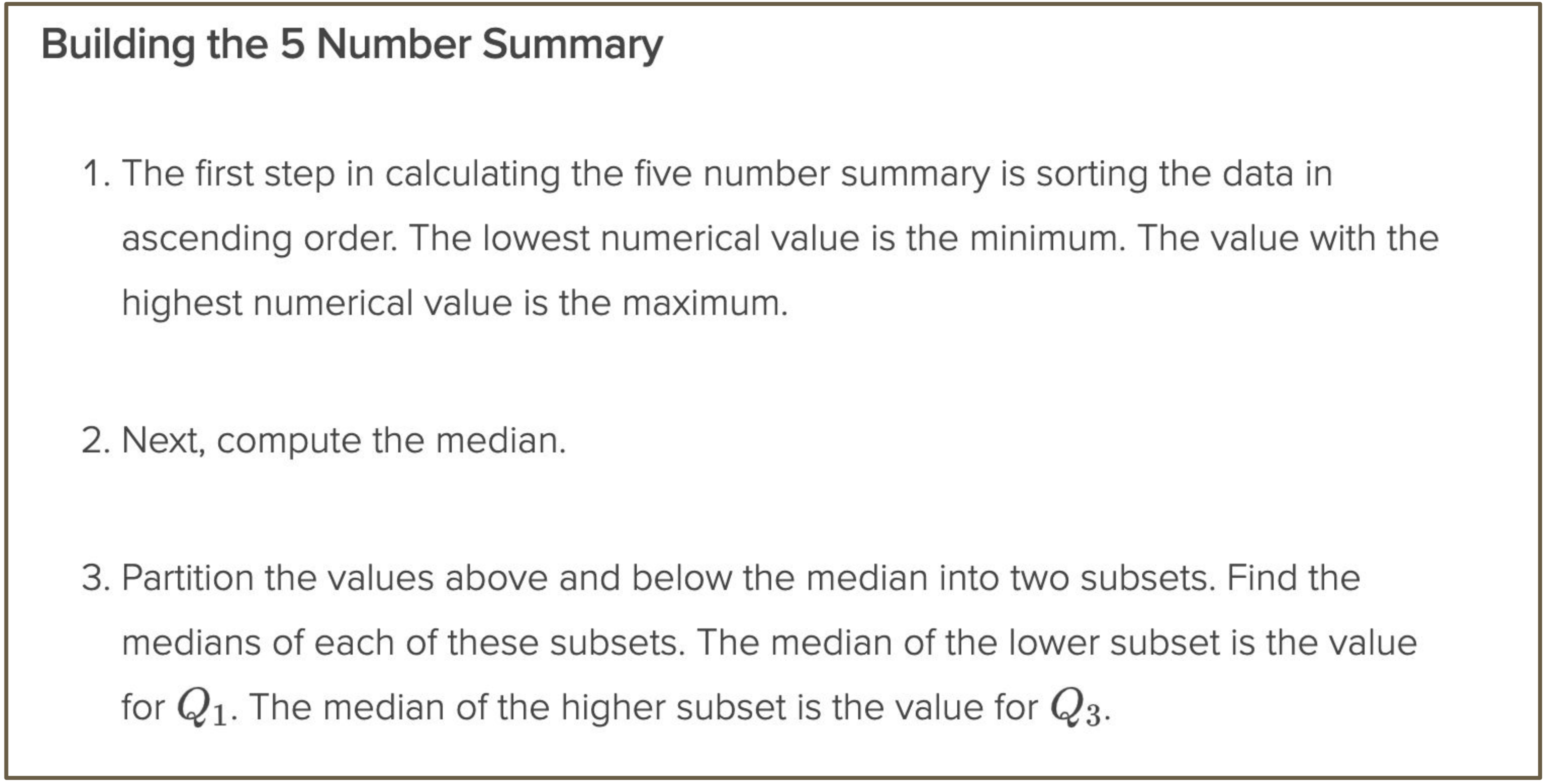
The five-number summary is a set of descriptive statistics that provides information about a dataset. It consists of the five most important sample percentiles

- The minimum
- The lower (first) quartile: Q_1
- The median
- The upper (third) quartile
- The maximum

Note:

The values are often expressed in a tuple, as follows:

$$(min,\ Q_1,\ median,\ Q_3,\ max)$$

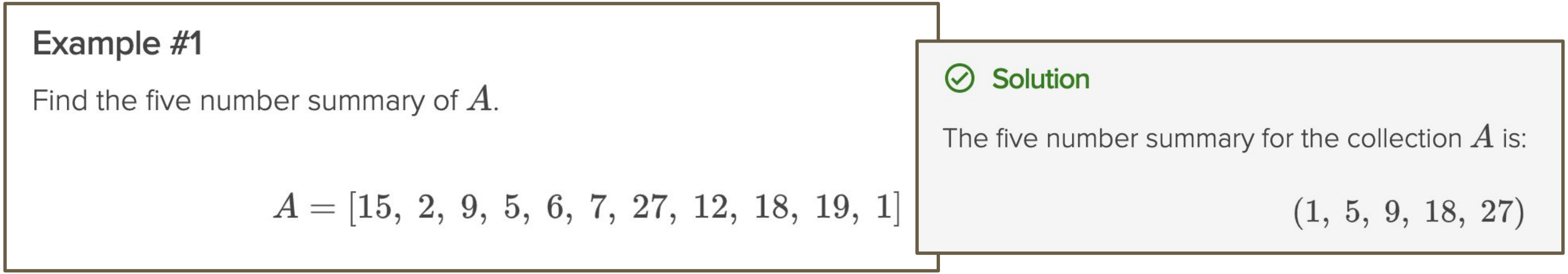


Building the 5 Number Summary

1. The first step in calculating the five number summary is sorting the data in ascending order. The lowest numerical value is the minimum. The value with the highest numerical value is the maximum.

2. Next, compute the median.

3. Partition the values above and below the median into two subsets. Find the medians of each of these subsets. The median of the lower subset is the value for Q_1. The median of the higher subset is the value for Q_3.



Example #1

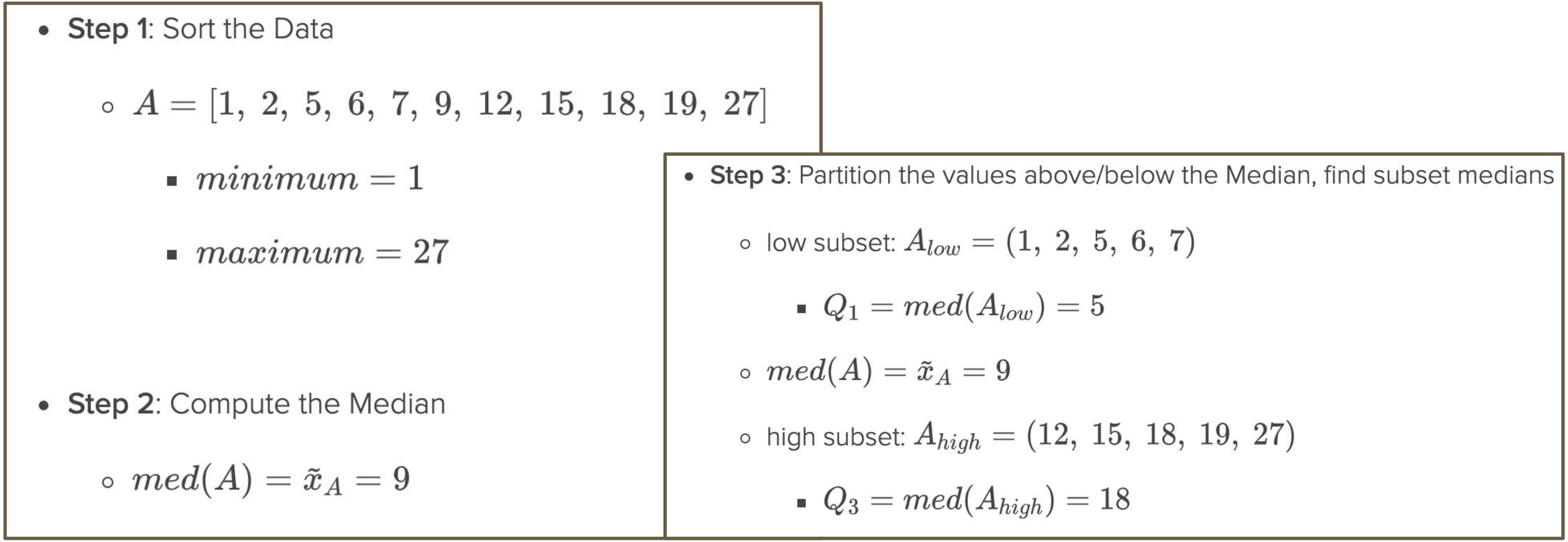
Find the five number summary of A.

$$A = [15,\ 2,\ 9,\ 5,\ 6,\ 7,\ 27,\ 12,\ 18,\ 19,\ 1]$$

✓ **Solution**

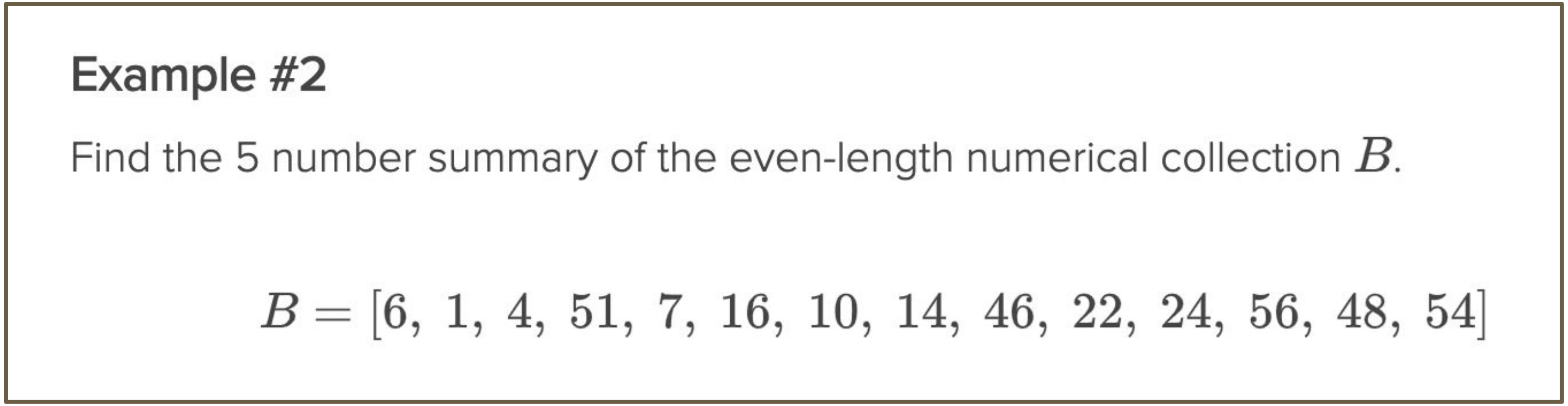
The five number summary for the collection A is:

$$(1,\ 5,\ 9,\ 18,\ 27)$$



- **Step 1**: Sort the Data
  - $A = [1,\ 2,\ 5,\ 6,\ 7,\ 9,\ 12,\ 15,\ 18,\ 19,\ 27]$
    - $minimum = 1$
    - $maximum = 27$
- **Step 2**: Compute the Median
  - $med(A) = \tilde{x}_A = 9$

- **Step 3**: Partition the values above/below the Median, find subset medians
  - low subset: $A_{low} = (1,\ 2,\ 5,\ 6,\ 7)$
    - $Q_1 = med(A_{low}) = 5$
  - $med(A) = \tilde{x}_A = 9$
  - high subset: $A_{high} = (12,\ 15,\ 18,\ 19,\ 27)$
    - $Q_3 = med(A_{high}) = 18$

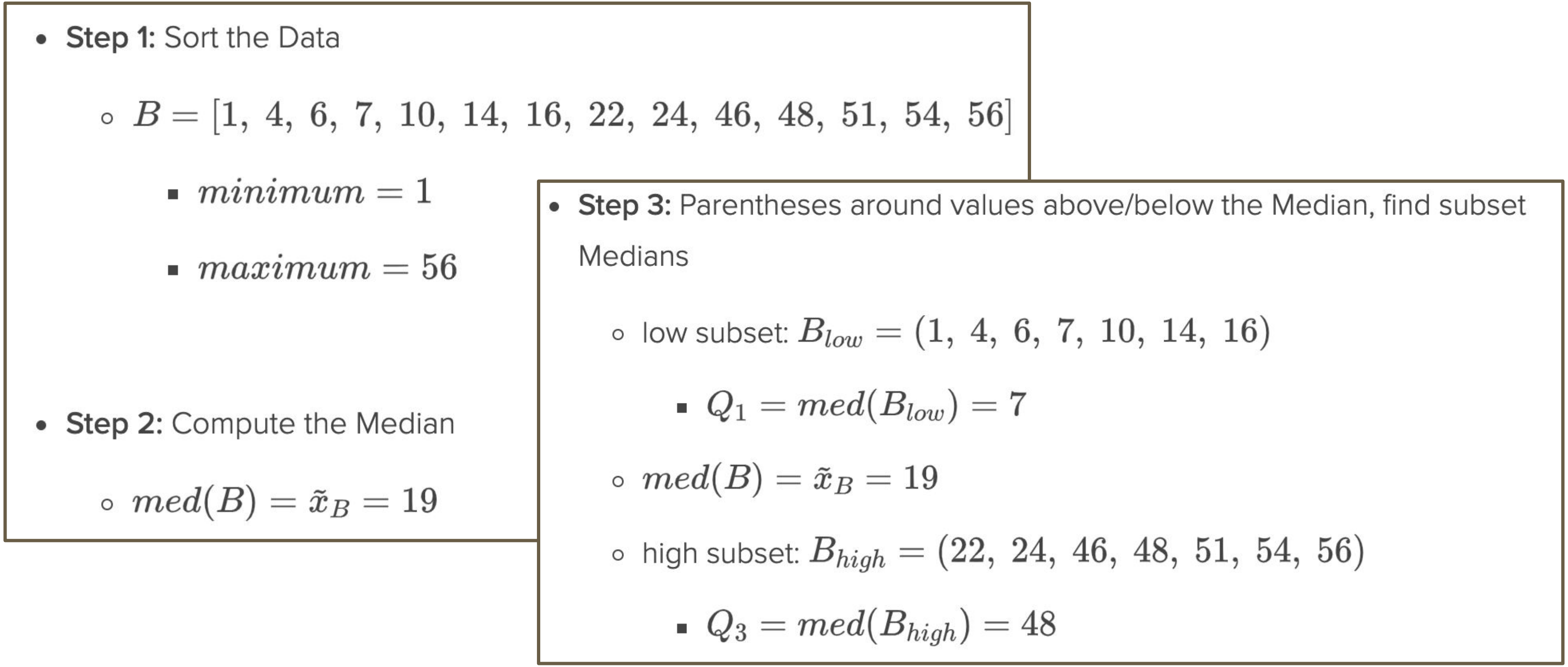


Example #2

Find the 5 number summary of the even-length numerical collection B.

$$B = [6,\ 1,\ 4,\ 51,\ 7,\ 16,\ 10,\ 14,\ 46,\ 22,\ 24,\ 56,\ 48,\ 54]$$

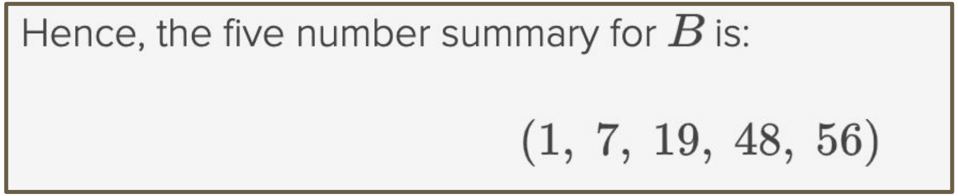
EXERCISE: SOLVE right now



- **Step 1:** Sort the Data
  - $B = [1,\ 4,\ 6,\ 7,\ 10,\ 14,\ 16,\ 22,\ 24,\ 46,\ 48,\ 51,\ 54,\ 56]$
    - $minimum = 1$
    - $maximum = 56$

- **Step 2:** Compute the Median
  - $med(B) = \tilde{x}_B = 19$

- **Step 3:** Parentheses around values above/below the Median, find subset Medians
  - low subset: $B_{low} = (1,\ 4,\ 6,\ 7,\ 10,\ 14,\ 16)$
    - $Q_1 = med(B_{low}) = 7$
  - $med(B) = \tilde{x}_B = 19$
  - high subset: $B_{high} = (22,\ 24,\ 46,\ 48,\ 51,\ 54,\ 56)$
    - $Q_3 = med(B_{high}) = 48$



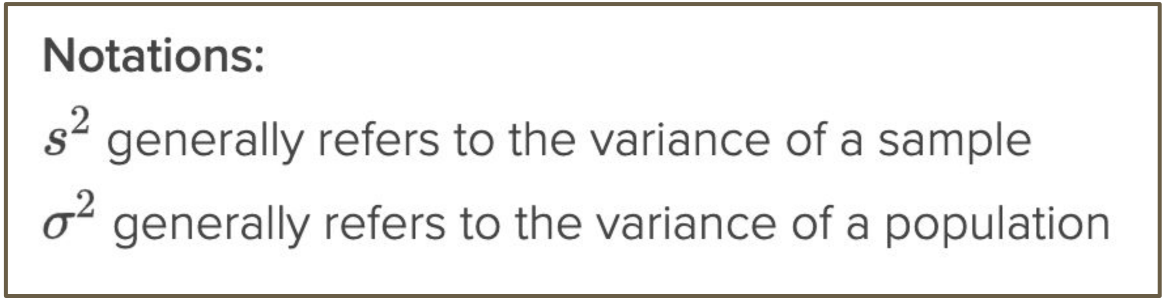
Hence, the five number summary for B is:

$$(1,\ 7,\ 19,\ 48,\ 56)$$

- Find the *Five Number Summary* Using *Python*: **Workshop**
- *Five Number Summary* related mathematical challenges: **Tutorial**

Variance and Standard Deviation

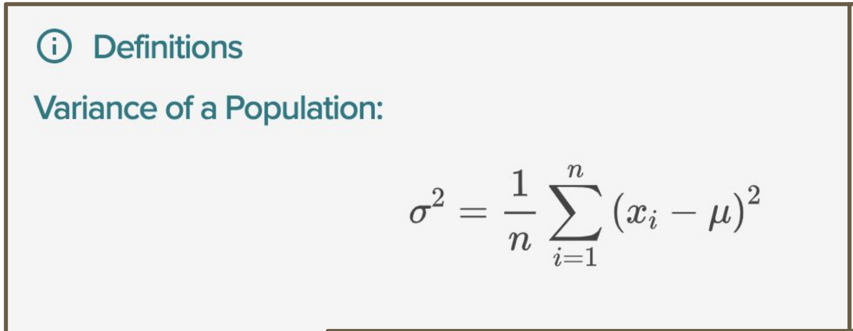
- The purpose of both the variance and standard deviation statistics are to express an easily interpretable measure of spread in a collection.
- The variance can be interpreted as the average squared deviations of each number from the mean, and it is calculated as such. The reason why we square the deviations is so we can deal with only positive values, If we didn't square the values our variation would end up being zero for every distribution.
- This approach would not yield a meaningful measure of a collection's spread.



Notations:

s^2 generally refers to the variance of a sample

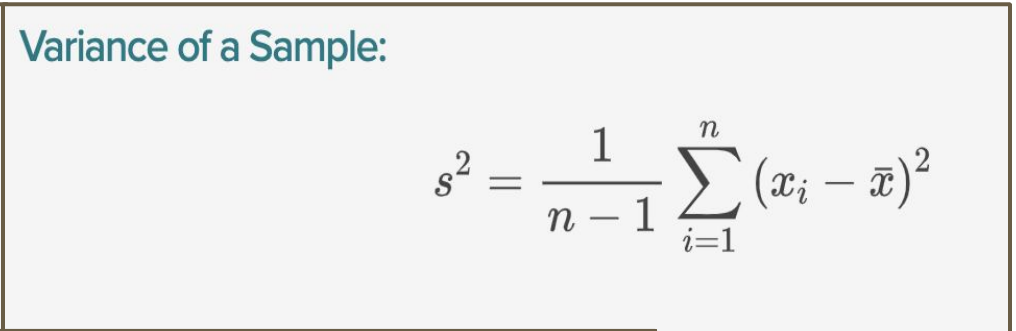
σ^2 generally refers to the variance of a population



ⓘ Definitions

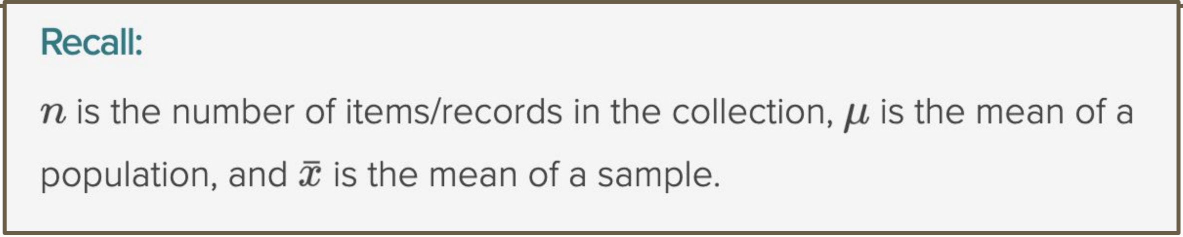
Variance of a Population:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$



Variance of a Sample:

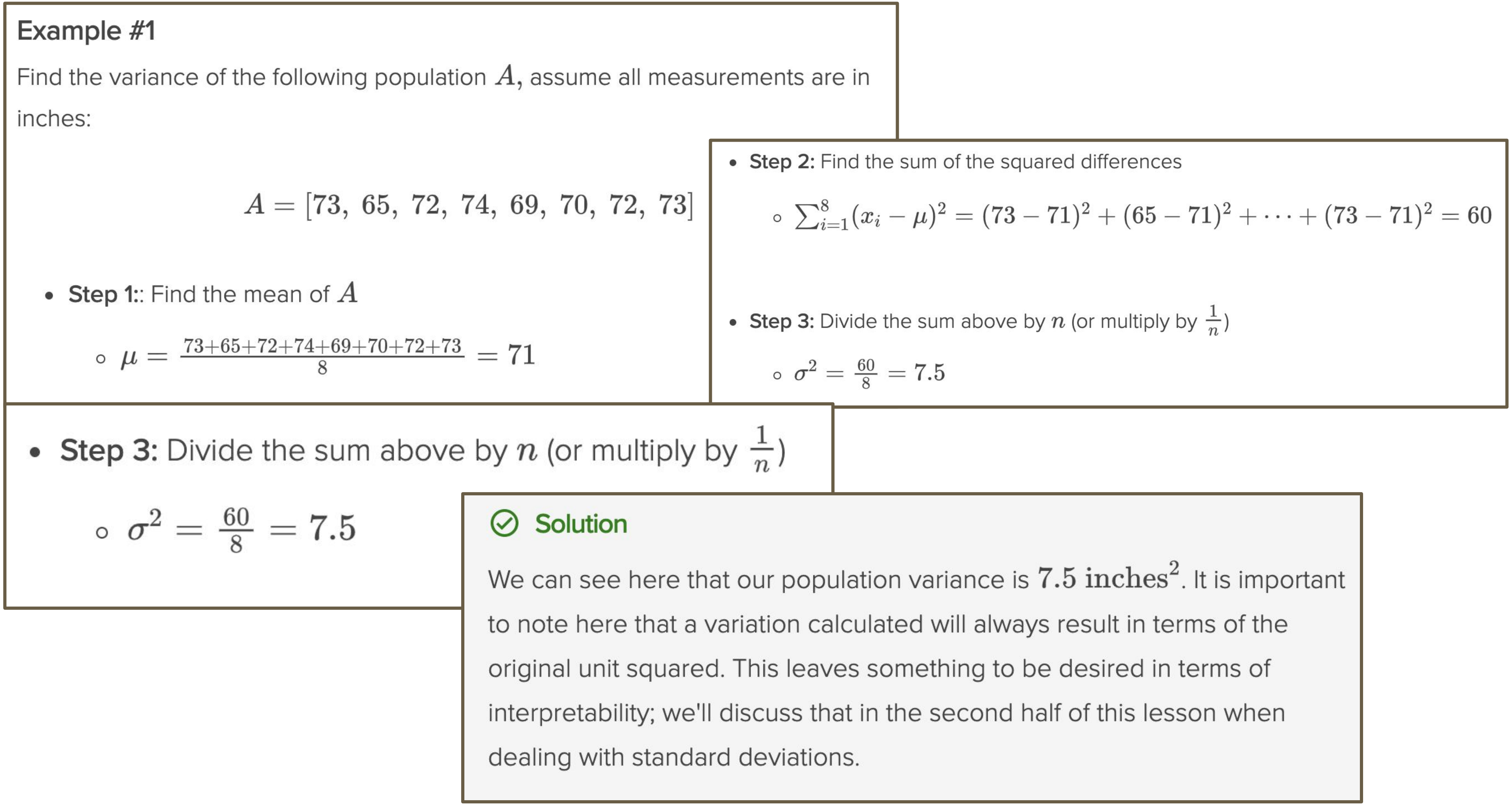
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



Recall:

n is the number of items/records in the collection, μ is the mean of a population, and $\bar{x}$ is the mean of a sample.

You can see the two formulas for variation are very similar, the primary difference being that the population variance is averaged by dividing by **n**; in the computation of a sample standard deviation, we use **n - 1**. This is called the **Bessel's correction**. This correction is made, because it partially corrects the bias in the estimation of a population variance.



Example #1

Find the variance of the following population A, assume all measurements are in inches:

$$A = [73,\ 65,\ 72,\ 74,\ 69,\ 70,\ 72,\ 73]$$

- **Step 1:**: Find the mean of A
  - $\mu = \frac{73+65+72+74+69+70+72+73}{8} = 71$

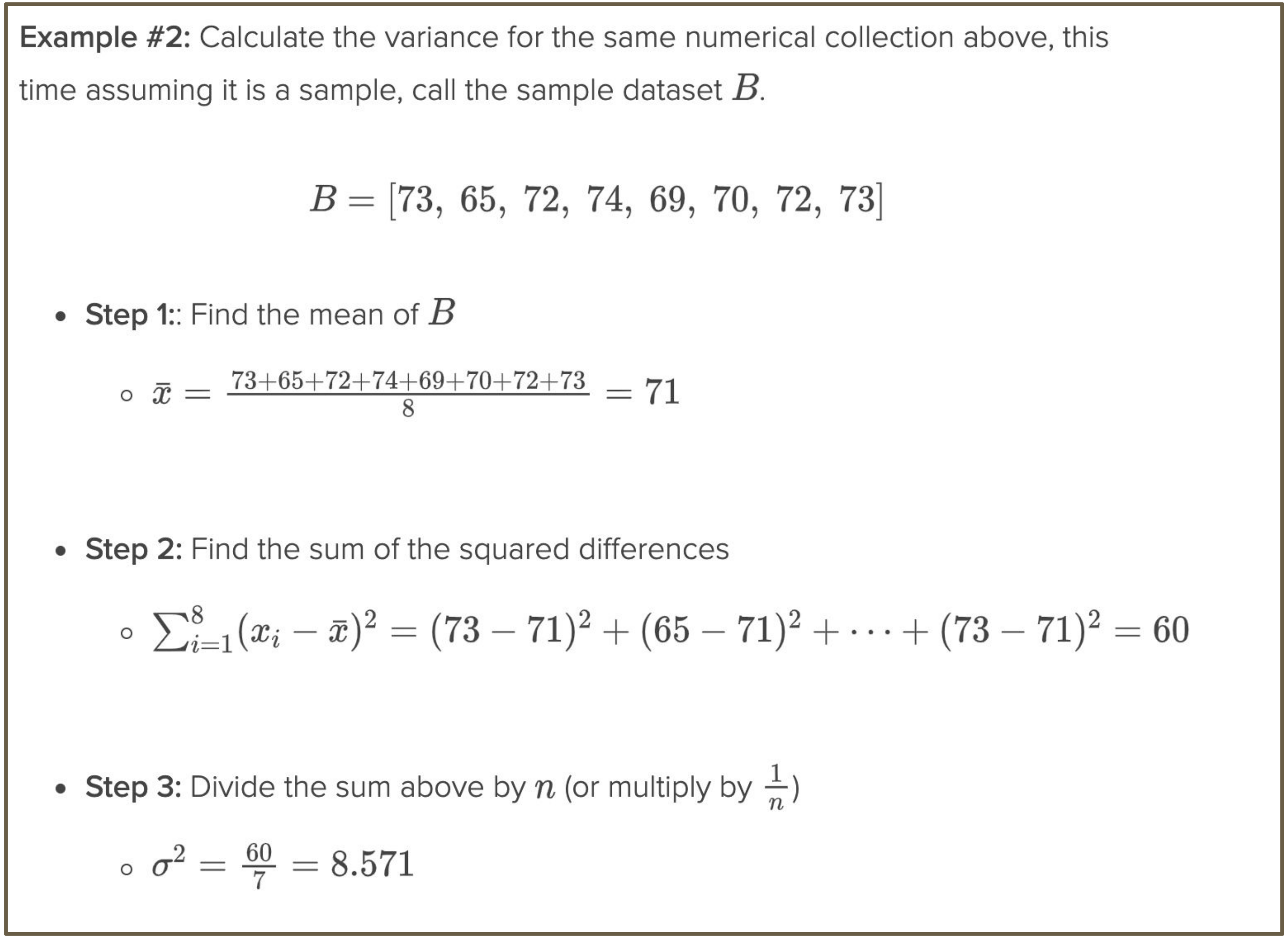
- **Step 2:** Find the sum of the squared differences
  - $\sum_{i=1}^{8}(x_i - \mu)^2 = (73-71)^2 + (65-71)^2 + \cdots + (73-71)^2 = 60$

- **Step 3:** Divide the sum above by n (or multiply by $\frac{1}{n}$)
  - $\sigma^2 = \frac{60}{8} = 7.5$

- **Step 3:** Divide the sum above by n (or multiply by $\frac{1}{n}$)
  - $\sigma^2 = \frac{60}{8} = 7.5$

✓ **Solution**

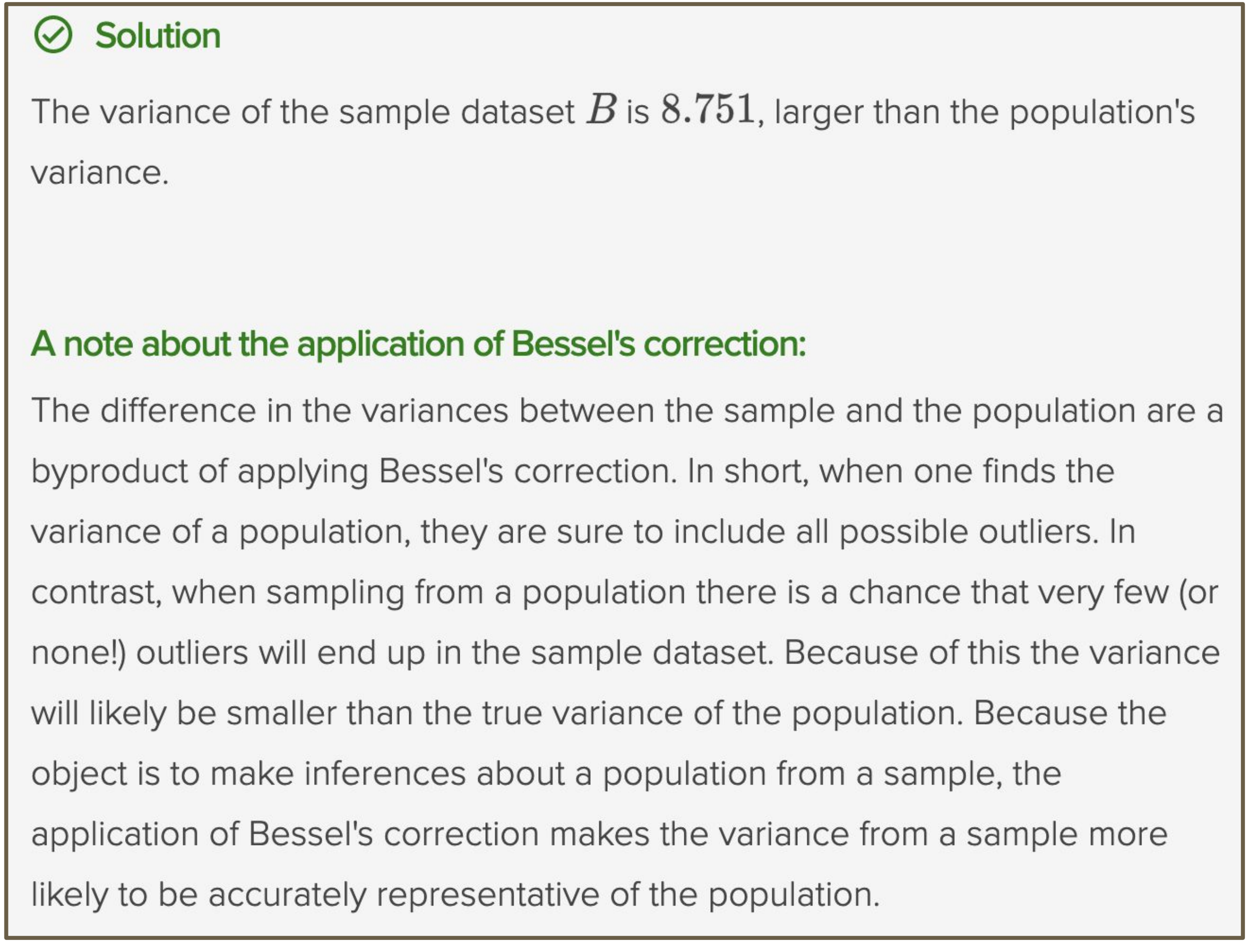
We can see here that our population variance is 7.5 inches^2. It is important to note here that a variation calculated will always result in terms of the original unit squared. This leaves something to be desired in terms of interpretability; we'll discuss that in the second half of this lesson when dealing with standard deviations.



Example #2: Calculate the variance for the same numerical collection above, this time assuming it is a sample, call the sample dataset B.

$$B = [73,\ 65,\ 72,\ 74,\ 69,\ 70,\ 72,\ 73]$$

- **Step 1:**: Find the mean of B
  - $\bar{x} = \frac{73+65+72+74+69+70+72+73}{8} = 71$
- **Step 2:** Find the sum of the squared differences
  - $\sum_{i=1}^{8}(x_i - \bar{x})^2 = (73-71)^2 + (65-71)^2 + \cdots + (73-71)^2 = 60$
- **Step 3:** Divide the sum above by n (or multiply by $\frac{1}{n}$)
  - $\sigma^2 = \frac{60}{7} = 8.571$



Solution

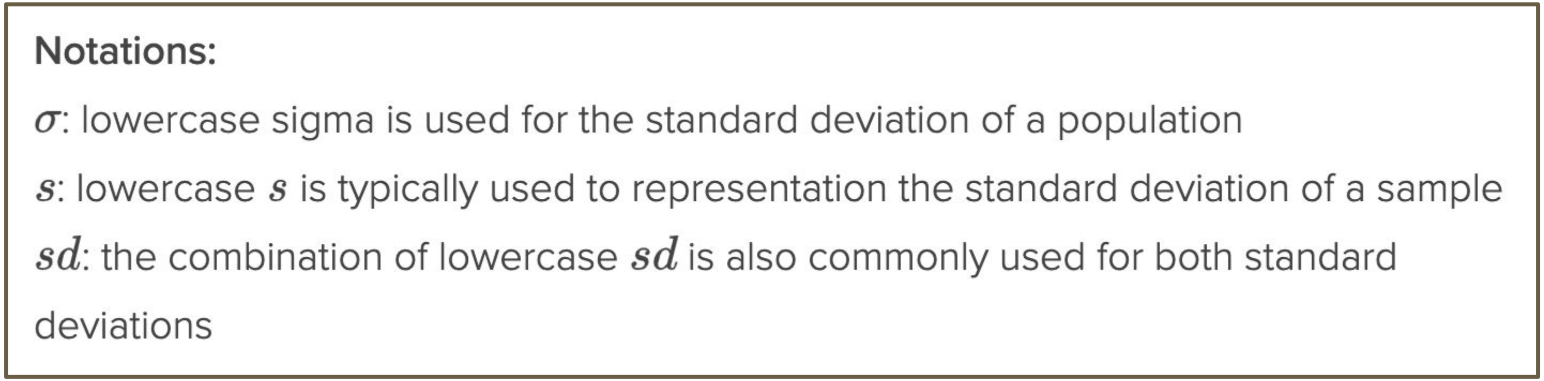
The variance of the sample dataset B is 8.751, larger than the population's variance.

A note about the application of Bessel's correction:

The difference in the variances between the sample and the population are a byproduct of applying Bessel's correction. In short, when one finds the variance of a population, they are sure to include all possible outliers. In contrast, when sampling from a population there is a chance that very few (or none!) outliers will end up in the sample dataset. Because of this the variance will likely be smaller than the true variance of the population. Because the object is to make inferences about a population from a sample, the application of Bessel's correction makes the variance from a sample more likely to be accurately representative of the population.

Standard Deviation (Population & Sample)

- The variance does a good job of describing the spread of a population or sample. However imagining the average spread in terms of the original units squared can be difficult to interpret.
- We typically take the square root of our variance, this yields us a standard deviation. The standard deviation ends up being in the same units as the original data.

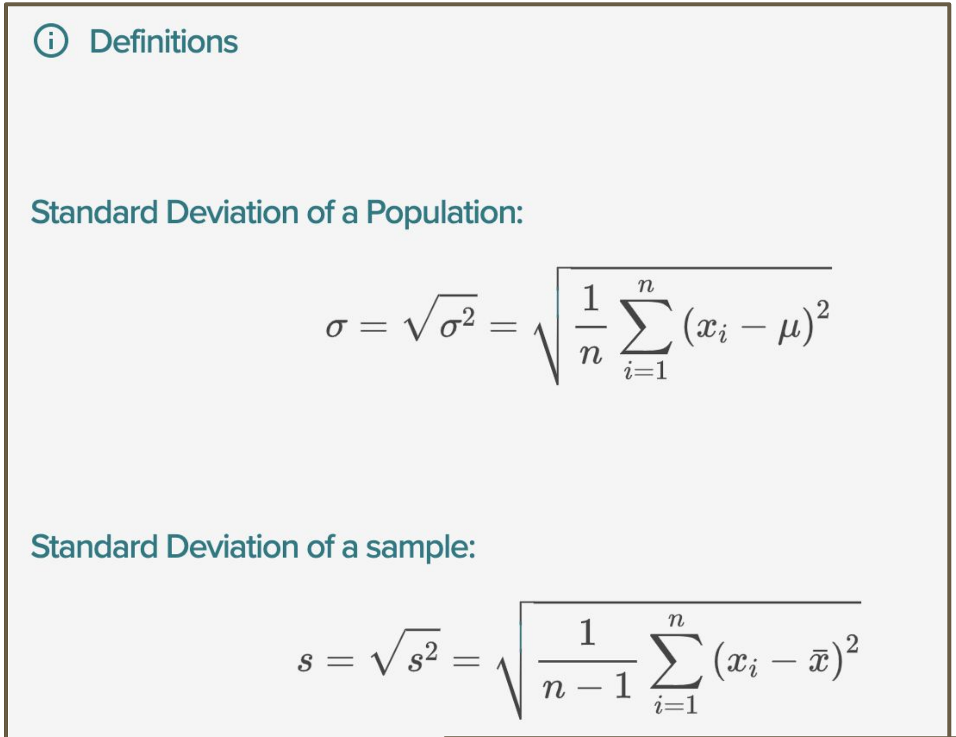


Notations:

σ: lowercase sigma is used for the standard deviation of a population

s: lowercase s is typically used to representation the standard deviation of a sample

sd: the combination of lowercase sd is also commonly used for both standard deviations



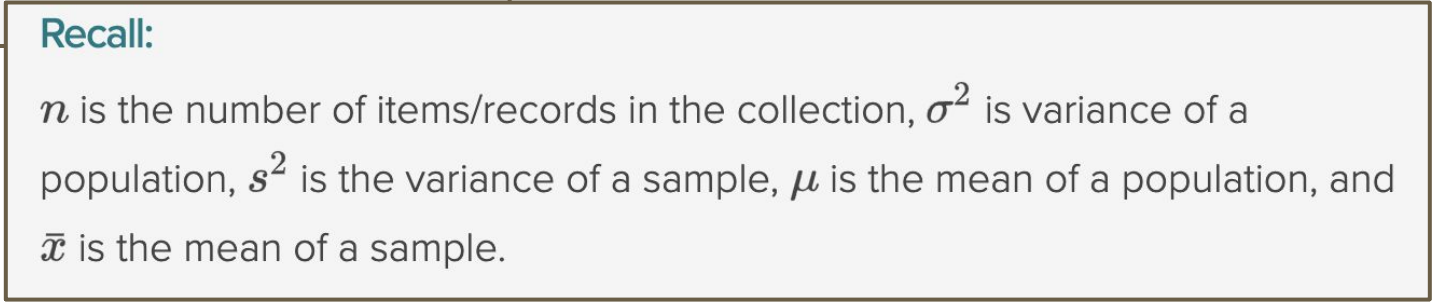
Definitions

Standard Deviation of a Population:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Standard Deviation of a sample:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Recall:

n is the number of items/records in the collection, σ^2 is variance of a population, s^2 is the variance of a sample, μ is the mean of a population, and $\bar{x}$ is the mean of a sample.

- Find the *Standard Deviation, Variance* Using *Python*: **Workshop**
- *Standard Deviation, Variance* related mathematical challenges: **Tutorial**

Thank You !