Calculating the Mean of a Data Set

The **mean**, or **average**, is a fundamental statistical measure that represents the center of a dataset by balancing values above and below it. It is widely used in fields like business, healthcare, and education. This section explains the mean, its formulas, and how to calculate it manually and with technology. Through examples, we will see how the mean helps summarize and interpret data efficiently.

Mean

What is the Mean?

The _____, also commonly known as the **average**, is the sum of all data values divided by the number of values. The sample mean is often denoted as $\bar{x}$, while the population mean is denoted as μ. The mathematical formulas for calculating the mean are:

$$\textbf{Population Mean:} \qquad \mu = \frac{\sum x}{N}$$

$$\textbf{Sample Mean:} \qquad \bar{x} = \frac{\sum x}{n}$$

What do these symbols mean?

Although the primary focus of this text is interpretation, it is still a math textbook, so we will encounter mathematical symbols and formulas throughout. To aid our understanding, we will explain these symbols as they appear, especially since many of these will be used repeatedly throughout this text.

- The symbol $\sum$ means "sum" or "add everything up."
- The symbol x represents individual data values.
- N denotes the total number of values in a population and is often referred to as the **population size**.
- n denotes the number of values in a sample and is often referred to as the **sample size.**

Why do we have two formulas for mean?

Populations and samples each have their own formulas for related concepts, such as the mean. In this case, the formulas are functionally identical, but as we explore other topics later in this chapter, we will see that some formulas differ between populations and samples.
Additionally, note that μ represents the _____ and is classified as a parameter, while $\overline{x}$ represents the _____ and is classified as a statistic.
But why do we use μ (pronounced "mew" and written in English as "mu") instead of a more familiar letter? By convention, parameters (which describe populations) are often represented by Greek letters, whereas statistics (which describe samples) are typically denoted using more familiar Latin letters from the English alphabet.

Now that we know the formulas for mean and how to interpret them, let's do a quick example to make sure we understand how to perform a calculation.

Example 1

Consider the following data representing test scores of five students on their first exam: 75, 80, 85, 90, 95. Use this data to calculate the average exam score for this sample.

Test Scores of Five Students

Score
75
80
85
90
95

Solution

Now that we understand how to calculate the mean, let's focus on what this number actually represents. One way to think about the mean is in terms of wealth redistribution. In society, some people have more money than others. The mean, or average, represents the amount each person would have if we could redistribute wealth so that everyone had exactly the same amount.

We can see this concept clearly using our previous example. Notice that the score of 75 is 10 points below the mean, while the score of 95 is 10 points above the mean. If we take 10 points from the person who scored 95 and give them to the person who scored 75, both students would now have 85 points. Similarly, since the score of 80 is 5 points below the mean and the score of 90 is 5 points above the mean, we can transfer 5 points from the student who scored 90 to the student who scored 80, so that both also end up with 85. After these adjustments, every student has a score of 85.
This illustrates what the mean represents—it balances out values above and below average to give a single number that evenly distributes the data across all individuals in the sample or population.

Of course, we don't actually redistribute scores or money in this way. The purpose of the mean is to help us understand the central tendency of a dataset—the balance point between those with the highest values and those with the lowest.

Now that we understand what a mean is and how to calculate a mean, we need to see how to calculate the mean using a technology since many datasets number in the hundreds and thousands. Manually calculating large datasets is time-consuming and prone to errors; in these circumstances, it is okay to let the technology do the heavy lifting.

Example 2

The following Law School Admission Test (LSAT) scores for a sample of 50 students are given below. Find the mean of the sample using the Summary Statistics Calculator.

Copy Data to Clipboard

Sample of 50 LSAT Scores

LSAT Scores									
174	172	169	176	169	170	175	171	168	177
165	180	173	166	178	170	174	167	179	172
163	181	171	164	177	169	175	168	180	170
162	182	170	165	176	168	174	166	178	171
161	183	169	167	175	167	173	165	177	172

Solution

Conclusion

The mean provides a simple yet powerful way to understand the central tendency of a dataset. It balances values above and below it, making it a key tool for data analysis. While calculating the mean manually is useful for small datasets, technology is essential for handling larger ones efficiently. Understanding the mean is a crucial step in mastering statistical analysis as we will be repeatedly using the mean throughout this entire text.

Calculating the Median of a Data Set

The median is a measure of central tendency that represents the middle value of an ordered dataset. In this section, we will define the median, discuss how to calculate it, and explore examples both by hand and using technology.

Median

What is the Median?

The _____ of an ordered dataset is the value that separates the lower 50% from the upper 50%. This number may or may not be part of the dataset. Unlike the mean, the median does not have a universally accepted notation, but many people represent it as M.

How do I calculate the Median?

Let n be the sample size of your data.

- **Step 1:** Order the data from smallest to largest. Ensure all repeated values are included.
- **Step 2:** Determine whether n is even or odd:
  - If n is **odd**, the median is the exact middle data value in the ordered list.
  - If n is **even**, the median is the _____ of the two middle data values in the ordered list.

While the procedure for calculating the median seems hard on the surface, it is in fact very easy to calculate, especially for small data sets. The next two examples will demonstrate the process step by step. The first example illustrates how to find the median with an odd sample size.

Example 3

The following **weekly hours spent studying** for a sample of **7 students** are recorded below. Find the median number of study hours by hand.

Sample of Weekly Study Hours (in hours)

Study Hours Per Week						
12	15	10	18	14	11	16

Solution

Finding the median for an even sample size follows the same steps, except for the final calculation.

Example 4

The cholesterol levels (mg/dL) of a sample of 10 people are recorded below. Find the median cholesterol level by hand.

Sample of 10 Cholesterol Levels

Cholesterol Level (mg/dL)				
154	240	171	188	235
203	184	173	181	275

Solution

Note

What does it mean to separate the upper 50% from the lower 50% of the data? The median is the average of the two central numbers, so it may not be an actual data point, even if it matches a value in the list. It lies exactly between the two central numbers, dividing the dataset into two equal parts. In this case, the lower half consists of 154, 171, 173, 181, and 184, while the upper half consists of 188, 203, 235, 240, and 275.

Just like the mean, we often compute the median from large data sets. Our next example uses our Summary Statistics Calculator to compute the median.

Example 5

The following LSAT scores for a sample of 50 students are given below. Find the **median** of the sample using the Summary Statistics Calculator.

Copy Data to Clipboard

Sample of 50 LSAT Scores

LSAT Scores									
174	172	169	176	169	170	175	171	168	177
165	180	173	166	178	170	174	167	179	172
163	181	171	164	177	169	175	168	180	170

Solution

Conclusion

The median is a valuable measure of central tendency because it is resistant to outliers. By understanding how to calculate the median manually and using technology, we can analyze data more effectively in various contexts.

Calculating the Mode of a Dataset

The mode is a measure of central tendency that identifies the most frequently occurring value(s) in a dataset. Unlike the mean and median, the mode does not require numerical calculations but instead focuses on how often values appear. In this section, we will define the mode, discuss its characteristics, and explore examples of how to determine and interpret it.

Mode

Definition: Mode

The _____ of a dataset is the value(s) that occur most frequently.

What Changed in This Definition?

By removing the term "occur locally," we ensure that all identified modes have the same frequency. This revised definition makes it easier to identify all the modes at a glance from frequency distributions.

Example 6

Find the mode of the **2016-2017 tuition and fees** (in thousands of dollars) for the top 14 universities in the U.S. by hand.

2016-2017 Tuition and Fees (in $1000s)

Tuition and Fees						
45	47	52	49	55	48	48
51	51	50	51	48	51	51

Solution

Example 7

The following **LSAT scores** for a sample of 50 students are recorded below. Use the Summary Statistics Calculator to determine if a mode exists, and if it does, identify the mode(s). Also, classify the dataset as **unimodal**, **multimodal**, or having **no mode**.

Copy Data to Clipboard

Sample of 50 LSAT Scores

LSAT Scores									
174	172	169	176	169	170	175	171	168	177
165	180	173	166	178	170	174	167	179	172
163	181	171	164	177	169	175	168	180	170

Solution

The mode is particularly useful to summarize certain types of qualitative data. Unfortunately, the **Summary Statistics Calculator** does not handle qualitative data. Therefore, we will use the Frequency Distribution Tool to make a frequency distribution of the data and then find the data point(s) with the highest frequency.

Example 8

The following dataset represents the size of shirts sold over the last 30 days at a clothing retailer. Use the Frequency Distribution Tool to determine the mode of the dataset.

Copy Data to Clipboard

Shirt Sizes Sold Over the Last 30 Days

Shirt Sizes									
Small	Medium	Large	X-Large	X-Large	Medium	Large	X-Large	Small	X-Large
Medium	X-Large	X-Large	Large	Small	X-Large	Medium	X-Large	Large	X-Large
X-Large	X-Large	Large	Small	Medium	X-Large	X-Large	X-Large	Large	Medium
X-Large	Large	Medium	X-Large	X-Large	Small	Medium	Large	X-Large	X-Large
Large	X-Large	Medium	Small	X-Large	X-Large	X-Large	Large	Small	Medium

Solution

Conclusion

The mode helps identify the most common values in a dataset, making it useful for analyzing both quantitative and qualitative data. Unlike the mean and median, the mode may not always exist and can have multiple values. By understanding how to determine the mode and classify distributions as unimodal, multimodal, or having no mode, we gain deeper insights into data patterns.

How Skewness Affects the Mean, Median, and Mode of a Dataset

Measures of central tendency—mean, median, and mode—help us describe datasets. However, their values are influenced by the shape of the distribution and the presence of outliers. In this section, we will define outliers, explore how outliers impact these measures, and how their relationships change based on the skewness of a dataset.

Outliers

What is an Outlier?

An **outlier** is a data point that does not follow the overall pattern or shape of a distribution. These are data points that are typically much larger or much smaller than other points in the data set.

Let's look at an example to see how outliers affect the mean, median and mode.

Example 1

A list of 5 exam scores are given below.

$$78 \qquad 82 \qquad 84 \qquad 91 \qquad 91$$

- **Part A:** Find the mean, median, and mode of this distribution.
- **Part B:**Another student scored a 41 on the exam, which is an outlier compared to the other five scores. Revaluate the mean, median, and mode.
- **Part C:** Which measure of center changed the least? Which changed the most? Did any remain unchanged?

Solution

Since an outlier affects each one of these measurements differently, we get the following definition.

What is resistance to outliers?

A statistic is **resistant to outliers** if extreme values cause little to no change in its value.
In the Example 1, notice that

- the mean is _____ to outliers.
- the median and mode are resistant to outliers.

We will use this idea to describe the relative locations of the mean, median, and mode in unimodal distributions.

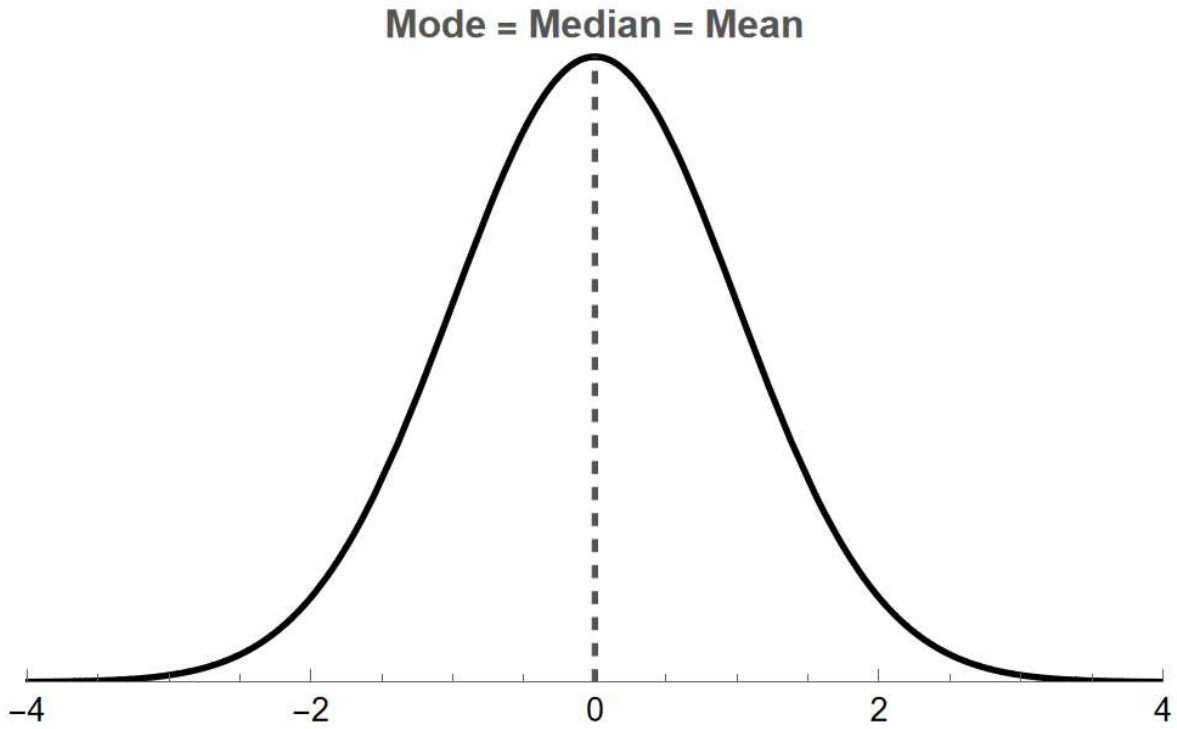
The Mean, Median, and Mode in Unimodal Distributions

Normal Distribution

In a normal distribution, we have that

$$\text{mean} = \text{median} = \text{mode}.$$

Since many real-world datasets follow a normal distribution, this is one reason we often focus on the mean rather than the median or mode. The following figure demonstrates the positions of the mean, median, and median by marking their location on top of a normal distribution.

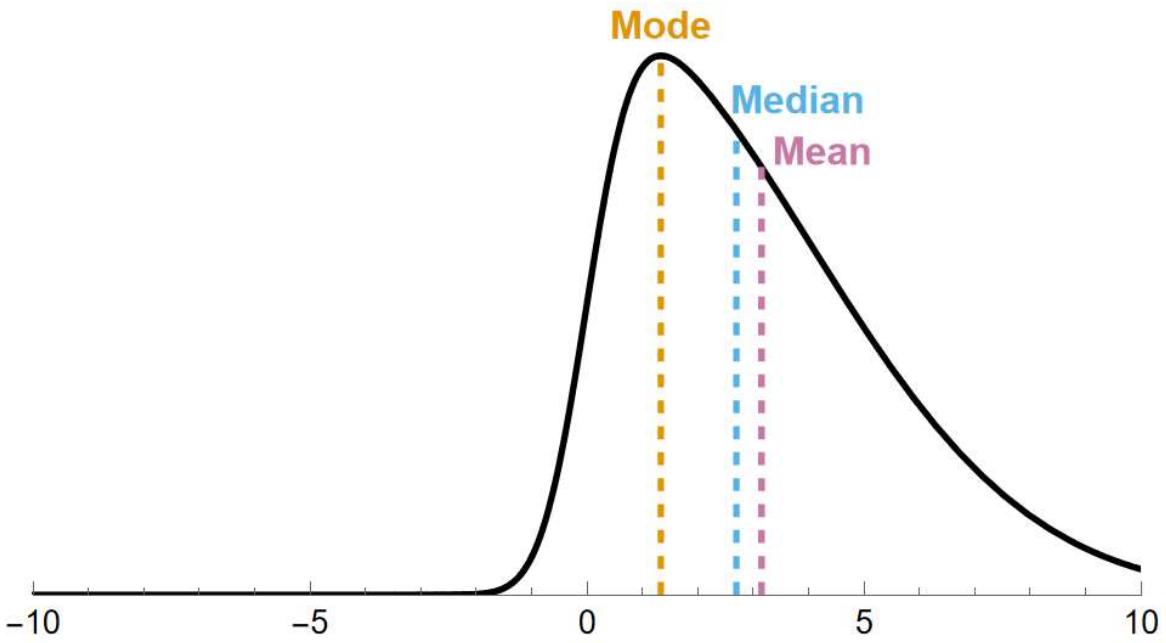


Skew-Right Distribution

In a skew-right distribution, part of the reason we get a skew to the right is because there is unusually large data in the distribution. Unusually large data will make the value of the mean and median shift to the right, with the mean being pulled further to the right than the median. The mode remains unchanged in this case. Overall, this change from a normal distribution can be summarized for skew-right distributions as

$$\text{mode} < \text{median} < \text{mean}$$

and visually as

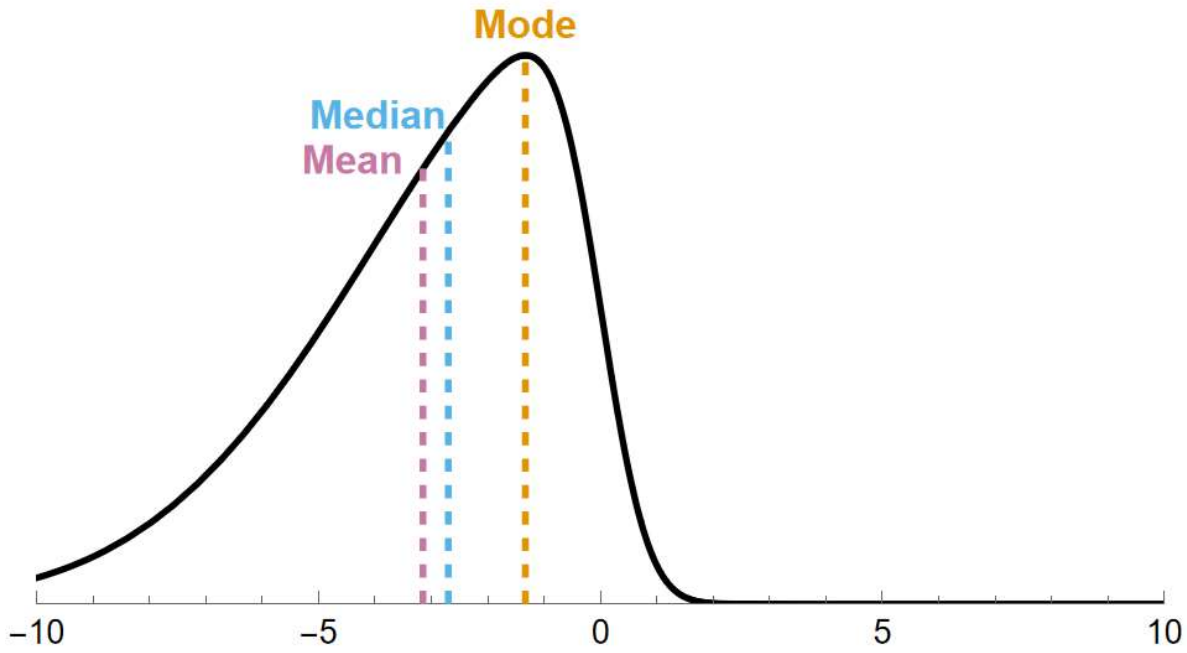


Skew-Left Distribution

In a skew-left distribution, part of the reason we get a skew to the left is because there is unusually small data in the distribution. Unusually small data will make the value of the mean and median shift to the left, with the mean being pulled further to the left than the median. The mode remains unchanged in this case. Overall, this change from a normal distribution can be summarized for skew-right distributions as

$$\text{mean} < \text{median} < \text{mode}$$

and visually as



Conclusion

The mean, median, and mode each describe the center of a dataset but respond differently to outliers and skewness. The mean is sensitive to extreme values and shifts in the direction of skewness, while the median is resistant to outliers and better represents center in skewed distributions. The mode remains unchanged by outliers and identifies the most frequently occurring value(s). Understanding these

differences helps in choosing the most appropriate measure for analyzing real-world data. We will explore that issue in more depth later in this chapter.