Why is the mean considered an important statistic in data analysis?

Authored by

stats writer

June 30, 2024

The mean is a central measure of tendency in a set of data and is often referred to as the average. It is calculated by adding all the values in a data set and dividing the sum by the total number of values. The mean is considered an important statistic in data analysis for several reasons. Firstly, it provides a single numerical value that summarizes the entire data set, making it easier to understand and interpret. Additionally, the mean is commonly used in statistical tests and calculations, such as standard deviation and correlation, to further analyze the data. Moreover, the mean is less affected by extreme values or outliers compared to other measures of central tendency, making it a more reliable representation of the data. This makes the mean a crucial tool for making informed decisions and drawing meaningful conclusions in various fields such as economics, medicine, and social sciences. Overall, the mean is an essential statistic in data analysis as it helps to simplify complex data and provides valuable insights for decision-making.

Why is the Mean Important in Statistics?

The mean of a dataset represents the average value of the dataset. It is calculated as:

Mean = Σxi / n

where:

Σ: A symbol that means "sum"xi: The ith observation in a datasetn: The total number of observations in the dataset

For example, suppose we have the following dataset with 11 observations:

Dataset: 3, 4, 4, 6, 7, 8, 12, 13, 15, 16, 17

The mean of the dataset is calculated as:

Mean = (3+4+4+6+7+8+12+13+15+16+17) / 11 = 9.54

In statistics, the mean is important for the following reasons:

1. The mean gives us an idea of where the "center" of a dataset is located.

2. Because of how it's calculated, the mean carries a piece of information from *every* in a dataset.

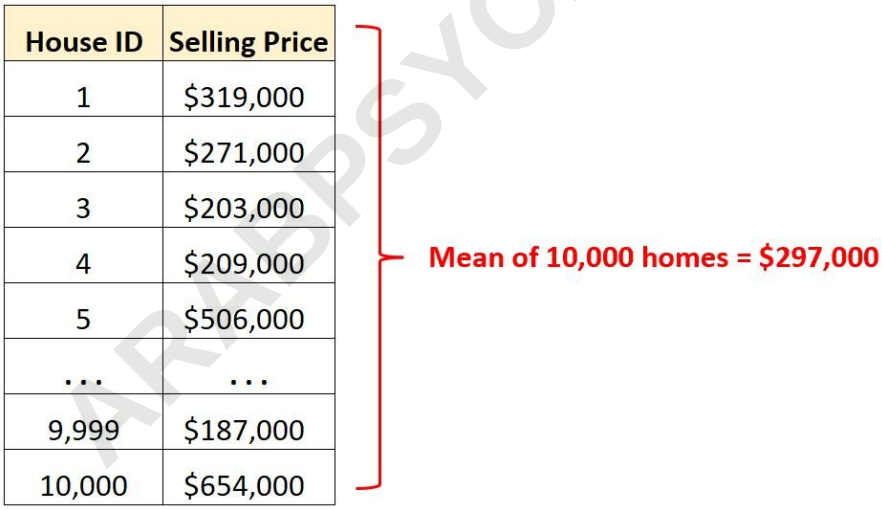
The following example illustrates both of these reasons.

Example: Calculating the Mean of a Dataset

Suppose we have a dataset that contains the selling price of 10,000 different homes in a certain city.

House ID	Selling Price
1	$319,000
2	$271,000
3	$203,000
4	$209,000
5	$506,000
. . .	. . .
9,999	$187,000
10,000	$654,000

Instead of staring at thousands of rows of , we can calculate the mean value to quickly understand the average selling price of homes in this city.



This single value for the mean is much easier to interpret compared to staring at all of the rows of raw data.

And since every single house selling price was used to calculate the mean, we could multiply the average selling price by the total number of houses to find the total selling price of all houses in this city:

Total selling price of all houses = Average selling price * Number of housesTotal selling price of all houses = $297,000 * 10,000Total selling price of all houses = $2,970,000,000

We can see that the total selling price of all houses in this city is $2.97 billion.

When to Use the Mean

When analyzing datasets, we're often interested in understanding where the center value is located.

In statistics, there are two common metrics that we use to measure the center of a dataset:

Mean: The average value in a datasetMedian: The middle value in a dataset

The mean is the most common way to measure the center of a dataset, but it can actually be misleading in

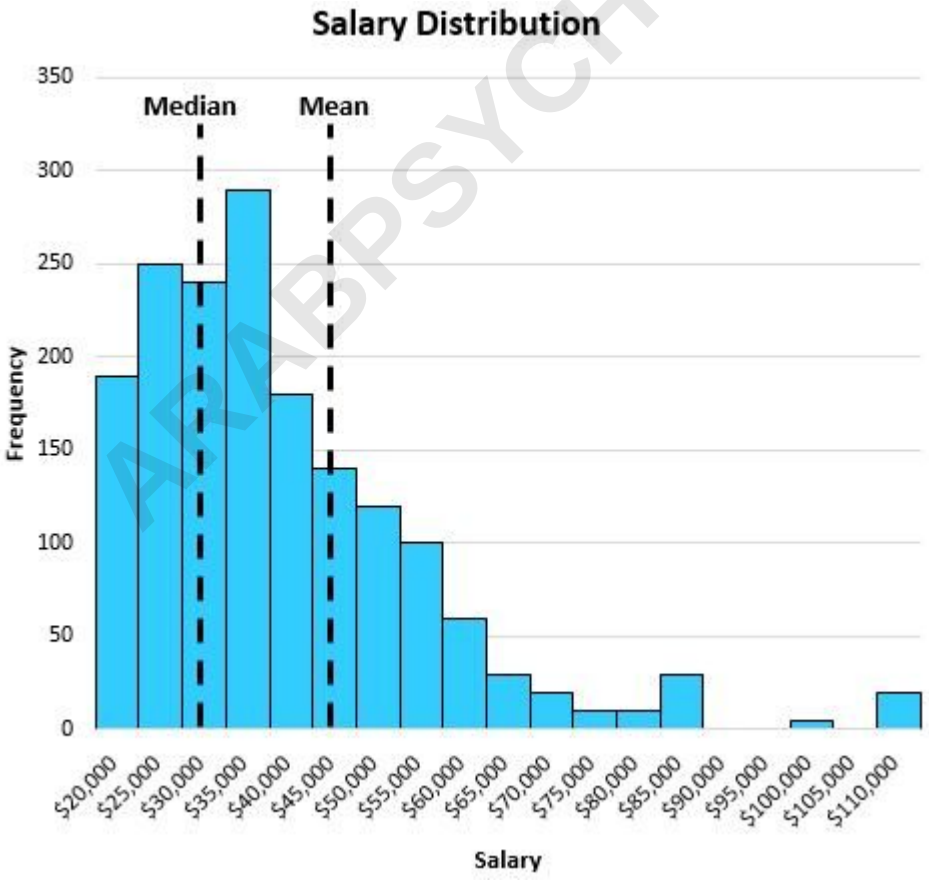
the following situations:

When the distribution is .When the distribution contains .

To illustrate this, consider the following two examples.

Example 1: Calculating the Mean of a Skewed Distribution

Consider the following distribution of salaries for residents in a certain city:



The large salaries on the right side of the distribution pull the mean away from the center of the distribution.

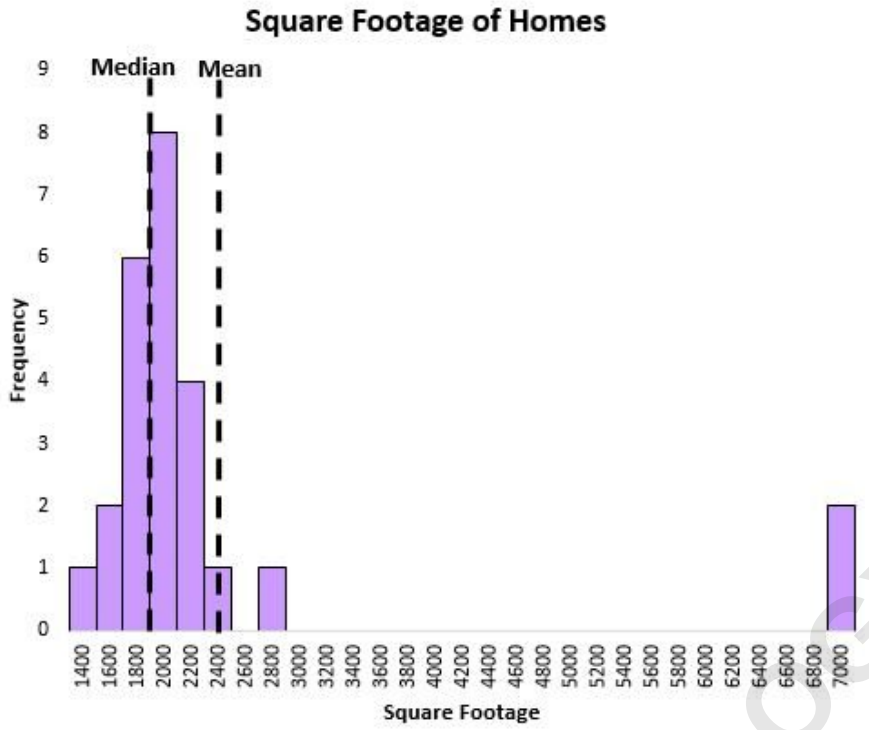
Thus, the median does a better job of capturing the "typical" salary of a resident than the mean because the distribution is right-skewed.

In this particular example, the mean salary is $47,000 while the median salary is $32,000.

Thus, the median is much more representative of the typical salary in this city.

Example 2: Calculating the Mean When Outliers Are Present

Consider the following chart that shows the square footage of houses on a certain street:



The mean is heavily influenced by a couple extremely large houses, while the median is not.

We can see that the median does a better job of capturing the "typical" square footage of a house on this street compared to the mean because it isn't influenced by the extreme outlier values.

Summary

Here's a quick summary of the main takeaways from this article:

The mean represents the average value in a dataset.The

mean is important because it gives us an idea of where the center value is located in a dataset.The mean is also important because it carries a piece of information from *every* observation in a dataset.The mean can be misleading when a dataset is skewed or contains outlies. In these scenarios, the median provides a more accurate idea of where the "center" of a dataset is located.

Additional Resources

The following tutorials provide additional information about other descriptive statistics: