# Statistical analysis

Lecture 1

## Mean
## Median, and Mode

Ola ali

# Mean,
# Median, and Mode

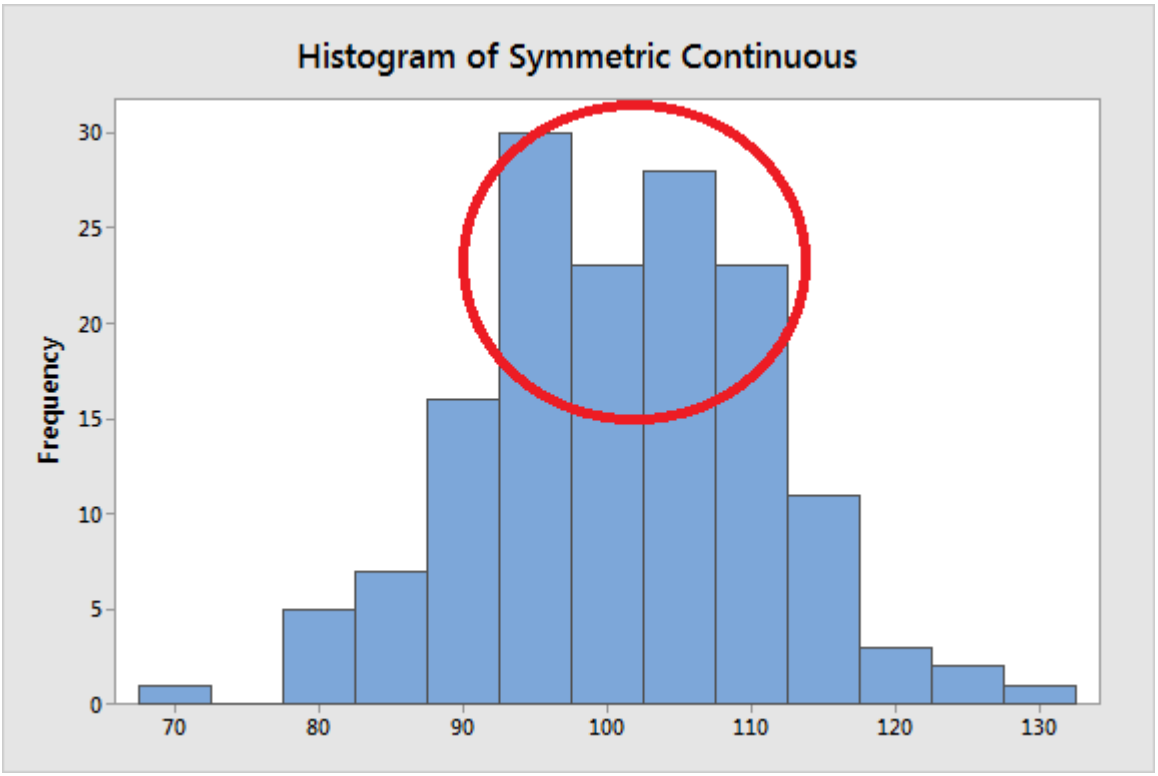
A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

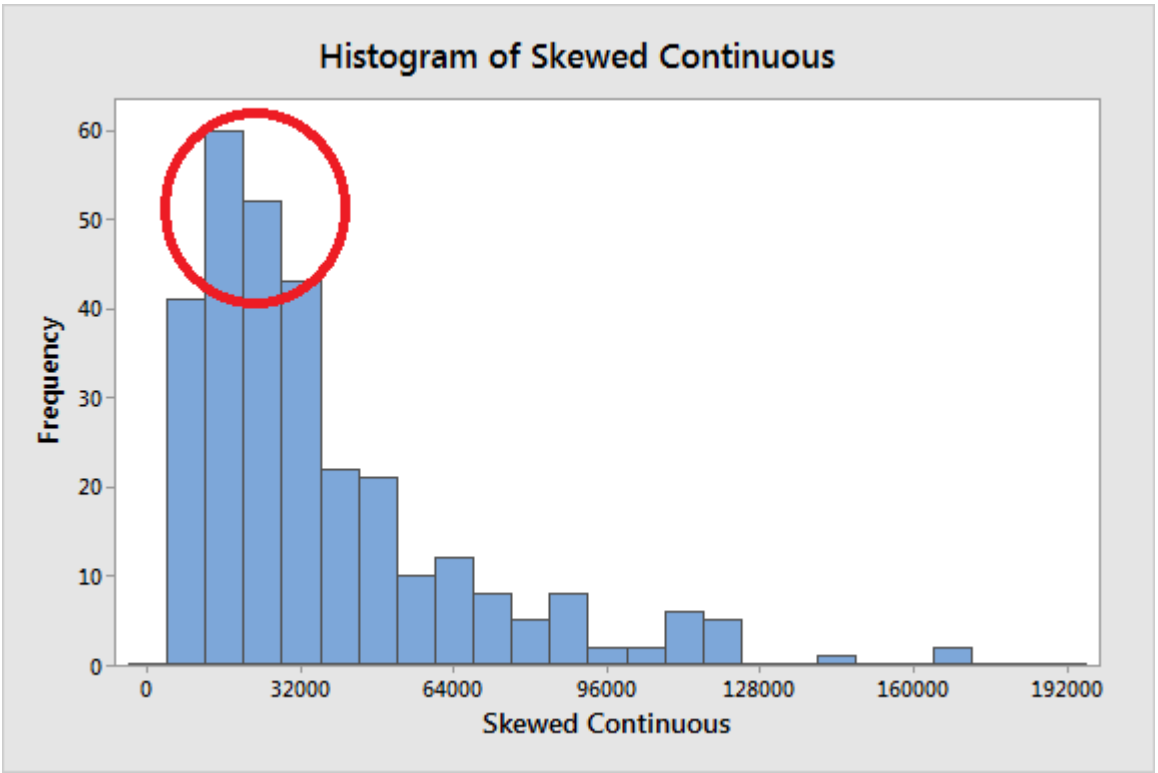
Choosing the best measure of central tendency depends on the type of data you have. In this post, I explore these measures of central tendency, show you how to calculate them, and how to determine which one is best for your data.

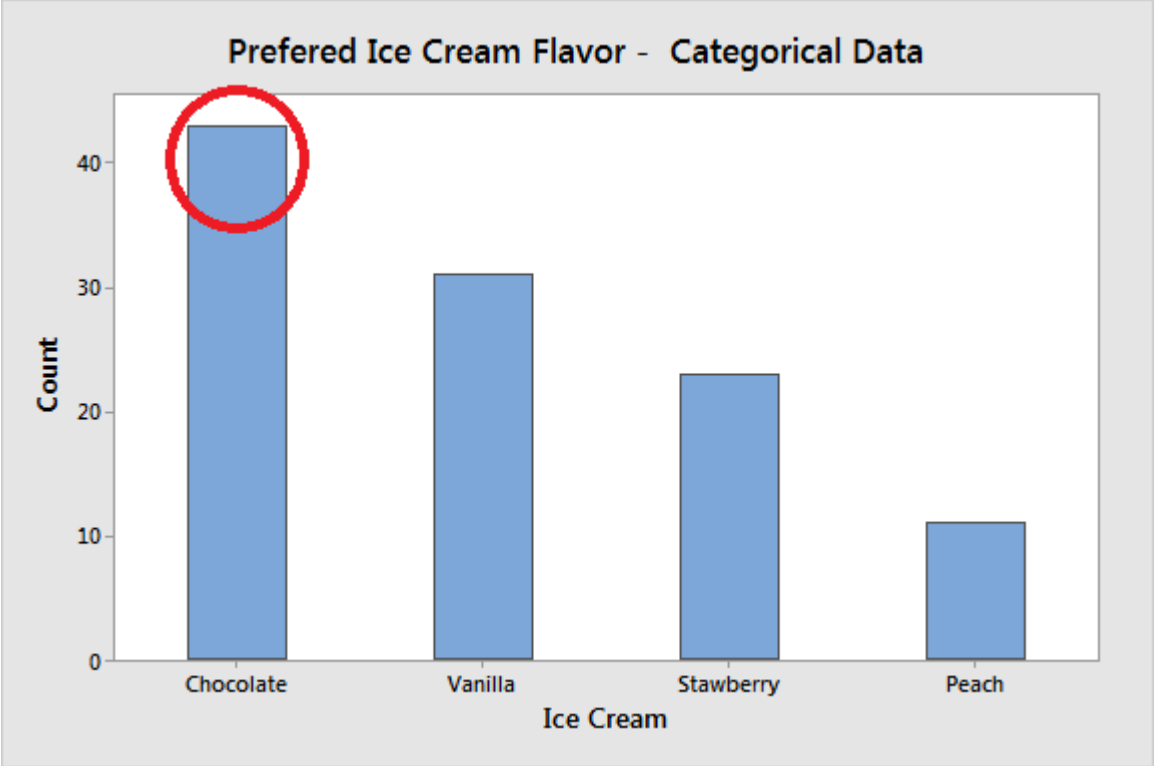
## Locating the Center of Your Data

Most articles that you'll read about the mean, median, and mode focus on how you calculate each one. I'm going to take a slightly different approach to start out. My philosophy throughout my blog is to help you intuitively grasp statistics by focusing on concepts. Consequently, I'm going to start by illustrating the central point of several datasets graphically—so you understand the goal. Then, we'll move on to choosing the best measure of central tendency for your data and the calculations.

The three distributions below represent different data conditions. In each distribution, look for the region where the most common values fall. Even though the shapes and type of data are different, you can find that central location. That's the area in the distribution where the most common values are located.


Histogram of Symmetric Continuous
Frequency
0
5
10
15
20
25
30
70
80
90
100
110
120
130

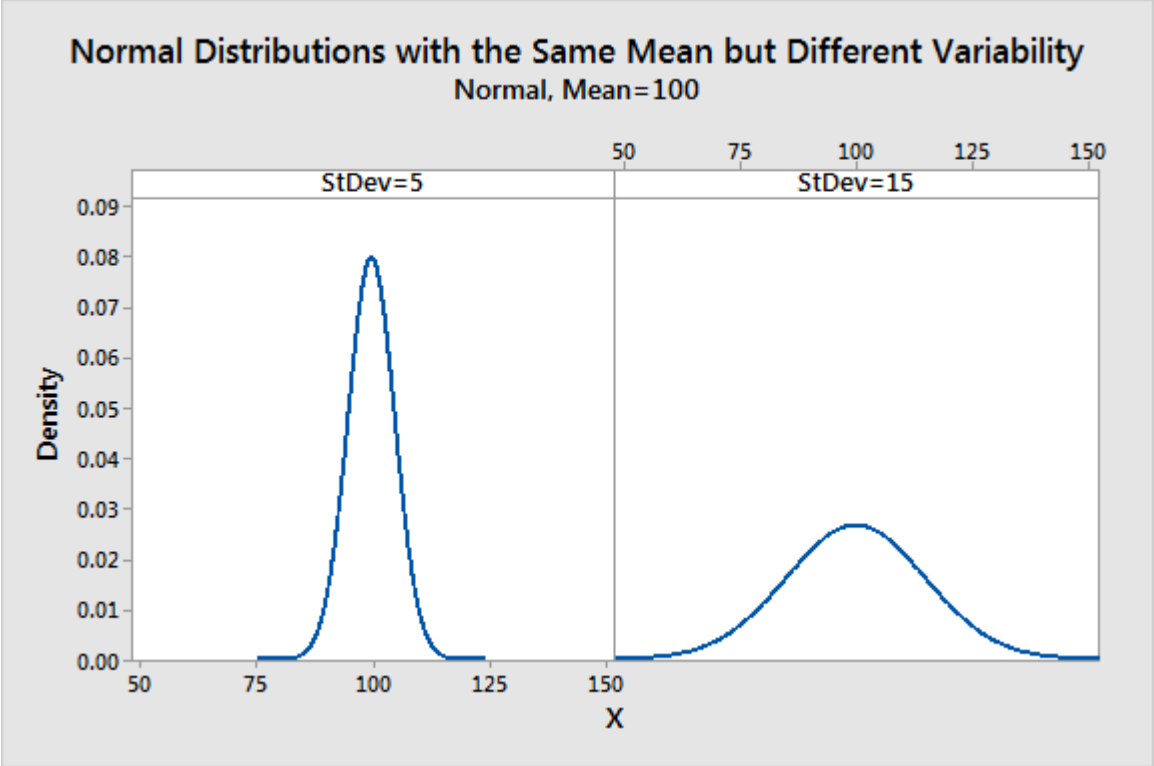

Histogram of Skewed Continuous
Frequency
0
10
20
30
40
50
60
0
32000
64000
96000
128000
160000
192000
Skewed Continuous




As the graphs highlight, you can see where most values tend to occur. That's the concept. Measures of central tendency represent this idea with a value. Coming up, you'll learn that as the distribution and kind of data changes, so does the best measure of central tendency. Consequently, you need to know the type of data you have, and graph it, before choosing a measure of central tendency!

**Related posts**: Guide to Data Types and How to Graph Them

The central tendency of a distribution represents one characteristic of a distribution. Another aspect is the variability around that central value. While measures of variability is the topic of a different article (link below), this property describes how far away the data points tend to fall from the center. The graph below shows how distributions with the same central tendency (mean = 100) can actually be quite different. The panel on the left displays a distribution that is tightly clustered around the mean, while the distribution on the right is more spread out. It is crucial to understand that the central tendency summarizes only one aspect of a distribution and that it provides an incomplete picture by itself.



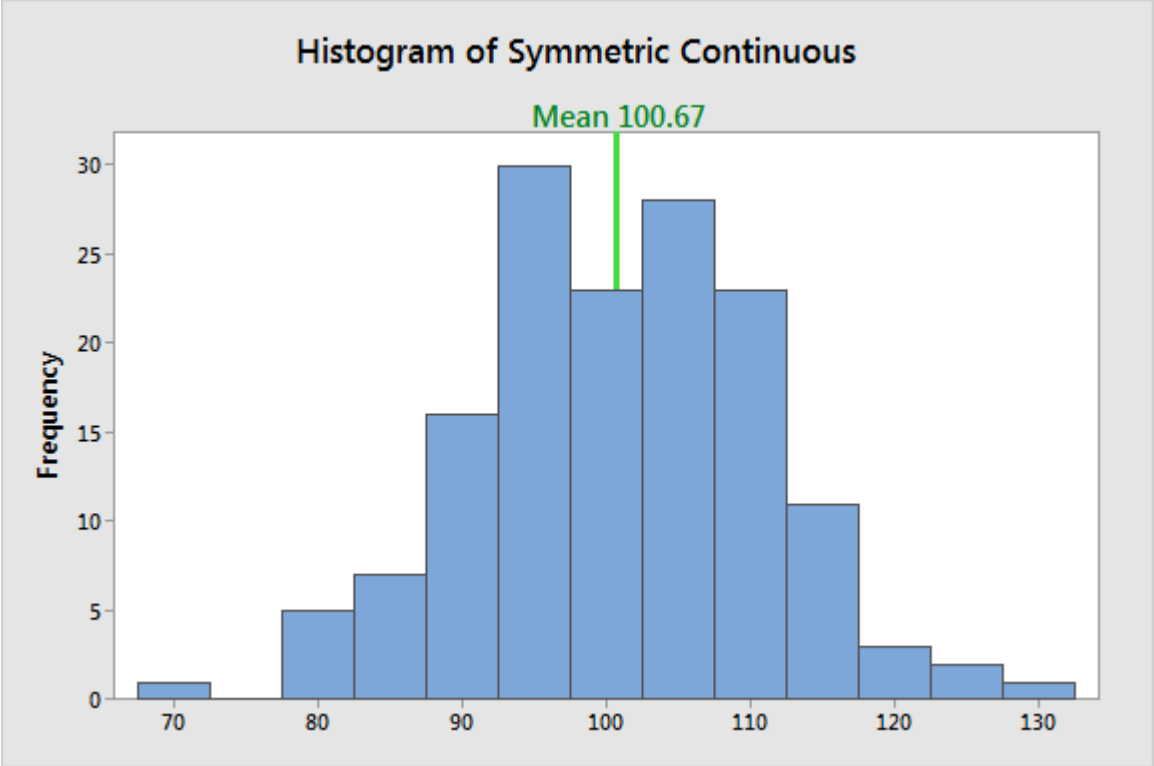

**Related post**: Measures of Variability: Range, Interquartile Range, Variance, and Standard Deviation

## Mean

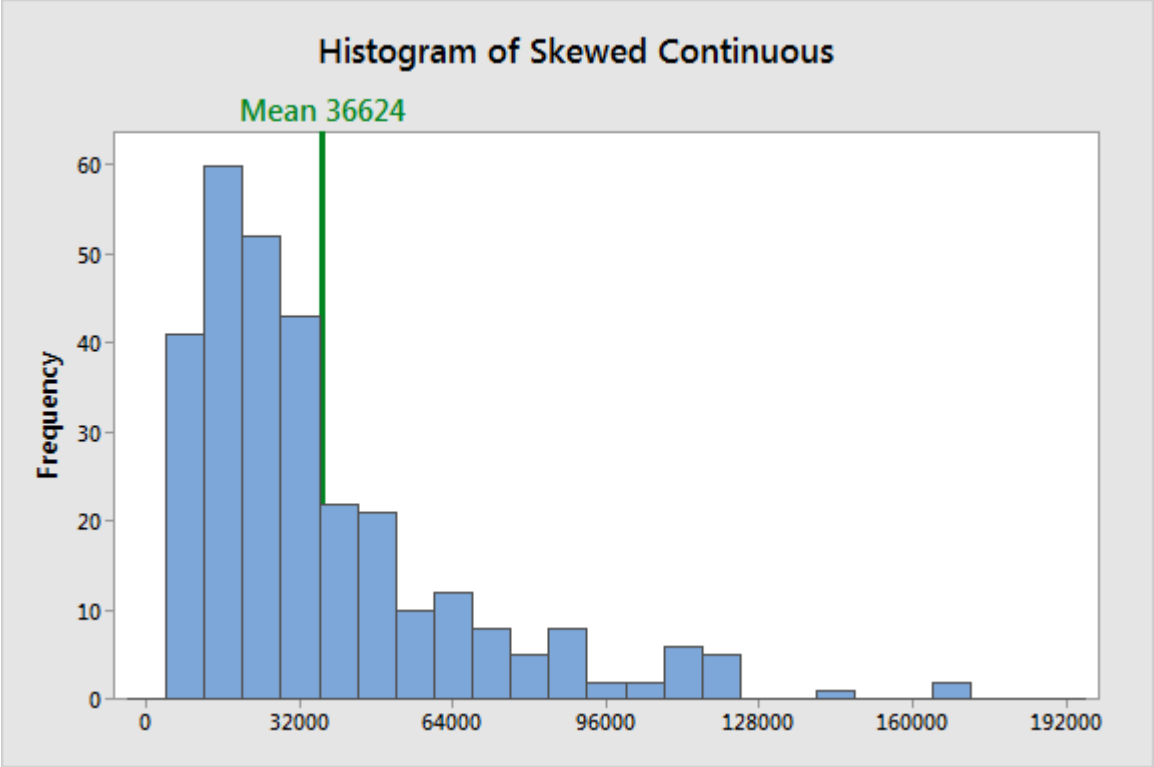
The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

The calculation of the mean incorporates all values in the data. If you change any value, the mean changes. However, the mean doesn't always locate the center of the data accurately. Observe the histograms below where I display the mean in the distributions.




In a symmetric distribution, the mean locates the center accurately.




However, in a skewed distribution, the mean can miss the mark. In the histogram above, it is starting to fall outside the central area. This problem occurs because outliers have a substantial impact on the mean. Extreme

values in an extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center. Consequently, it's best to use the mean as a measure of the central tendency when you have a symmetric distribution.

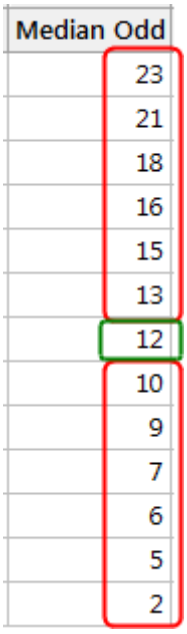
**When to use the mean**: Symmetric distribution, Continuous data

**Related post**: Using Histograms to Understand Your Data

## Median

The median is the middle value. It is the value that splits the dataset in half. To find the median, order your data from smallest to largest, and then find the data point that has an equal amount of values above it and below it. The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values. I'll show you how to find the median for both cases. In the examples below, I use whole numbers for simplicity, but you can have decimal places.

In the dataset with the odd number of observations, notice how the number 12 has six values above it and six below it. Therefore, 12 is the median of this dataset.




When there is an even number of values, you count in to the two innermost values and then take the average. The average of 27 and 29 is 28. Consequently, 28 is the median of this dataset.

| Median Even | |
|---|---|
| | 40 |
| | 38 |
| | 35 |
| | 33 |
| | 32 |
| | 30 |
| 28 | 29 |
| | 27 |
| | 26 |
| | 24 |
| | 23 |
| | 22 |
| | 19 |
| | 17 |

Outliers and skewed data have a smaller effect on the median. To understand why, imagine we have the Median dataset below and find that the median is 46. However, we discover data entry errors and need to change four values, which are shaded in the Median Fixed dataset. We'll make them all significantly higher so that we now have a skewed distribution with large outliers.

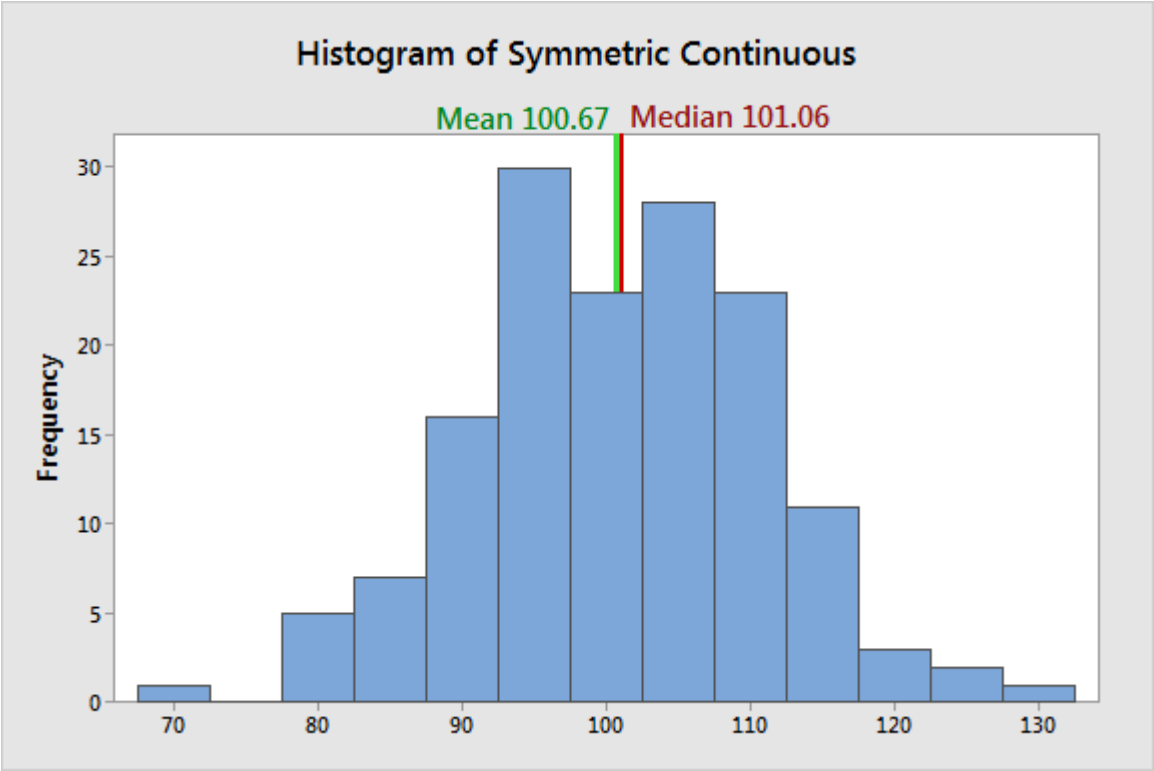
| Median | Median Fixed |
|---|---|
| 69 | 112 |
| 56 | 93 |
| 54 | 89 |
| 52 | 82 |
| 47 | 47 |
| 46 | 46 |
| 46 | 46 |
| 45 | 45 |
| 43 | 43 |
| 36 | 36 |
| 35 | 35 |
| 34 | 34 |
| 31 | 31 |

As you can see, the median doesn't change at all. It is still 46. Unlike the mean, the median value doesn't depend on all the values in the dataset. Consequently, when some of the values are more extreme, the effect on the median is smaller. Of course, with other types of changes, the median can

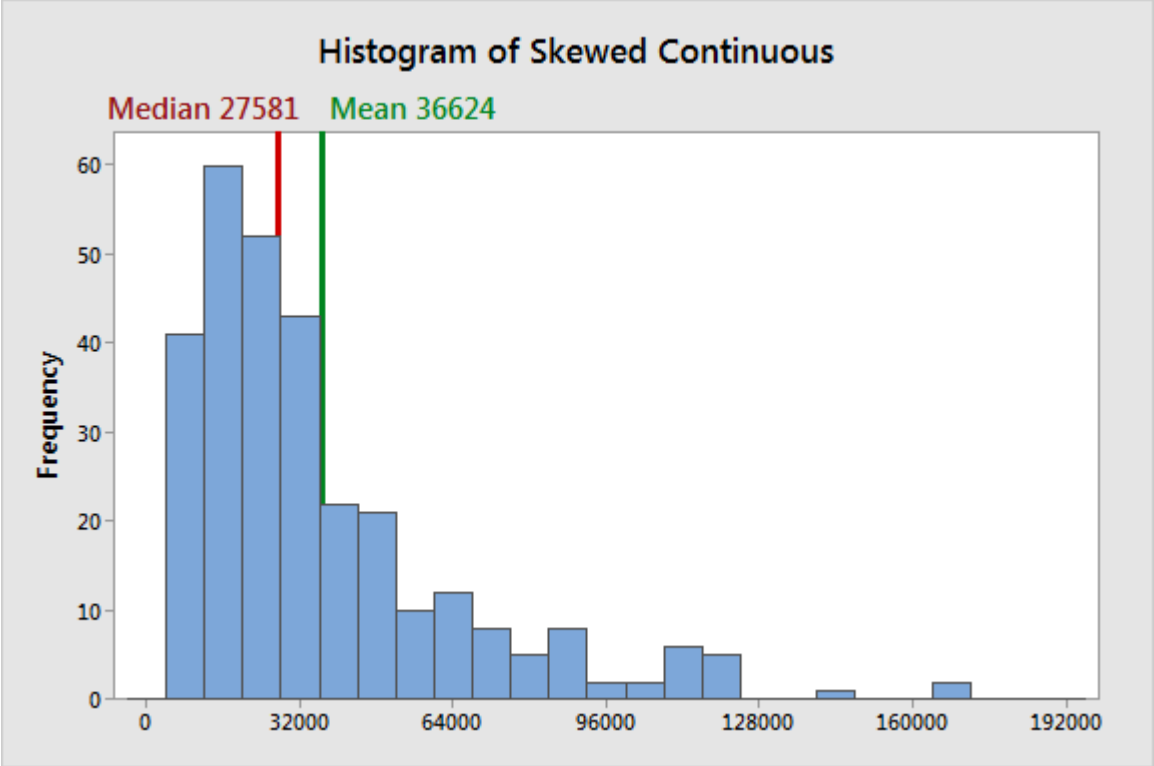
change. When you have a skewed distribution, the median is a better measure of central tendency than the mean.

## Comparing the mean and median

Now, let's test the median on the symmetrical and skewed distributions to see how it performs, and I'll include the mean on the histograms so we can make comparisons.




In a symmetric distribution, the mean and median both find the center accurately. They are approximately equal.




In a skewed distribution, the outliers in the tail pull the mean away from the center towards the longer tail. For this example, the mean and median differ by over 9000, and the median better represents the central tendency for the distribution.

These data are based on the U.S. household income for 2006. Income is the classic example of when to use the median because it tends to be skewed. The median indicates that half of all incomes fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

**When to use the median**: Skewed distribution, Continuous data, Ordinal data

## Mode

The mode is the value that occurs the most frequently in your data set. On a bar chart, the mode is the highest bar. If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution. If no value repeats, the data do not have a mode.

In the dataset below, the value 5 occurs most frequently, which makes it the mode. These data might represent a 5-point Likert scale.

| Mode |
| --- |
| 5 |
| 5 |
| 5 |
| 4 |
| 4 |
| 3 |
| 2 |
| 2 |
| 1 |

Typically, you use the mode with categorical, ordinal, and discrete data. In fact, the mode is the only measure of central tendency that you can use with categorical data—such as the most preferred flavor of ice cream. However, with categorical data, there isn't a central value because you can't order the groups. With ordinal and discrete data, the mode can be a value that is not in the center. Again, the mode represents the most common value .