15.4.5

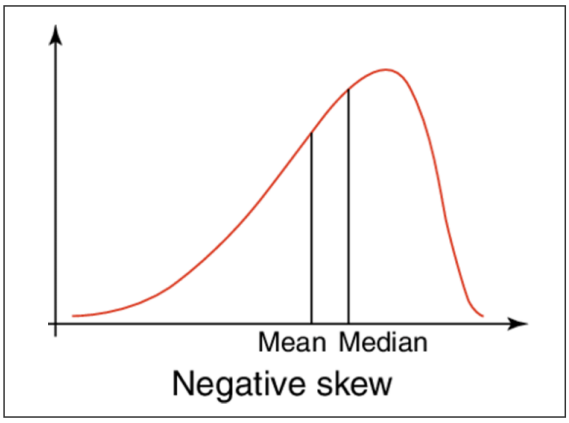
Understand Skew

Now that the team can test for normality, it's time to learn how to deal with datasets that aren't normal. Thankfully, Colleen knows how to do this and is using these mini lessons as a chance to practice for that big presentation for the CEO! So, on their next lunch break, they dig into the concept of skew.

When dealing with relatively smaller sample sizes, our data distributions are often asymmetrical. Compared to the normal distribution, where each tail of the distribution (on either side of the mean μ) mirrors one another, the asymmetrical distribution has one distribution tail that is longer than the other. This asymmetrical distribution is commonly referred to as a **skewed distribution** and there are two types—left skew and right skew.

Left Skew

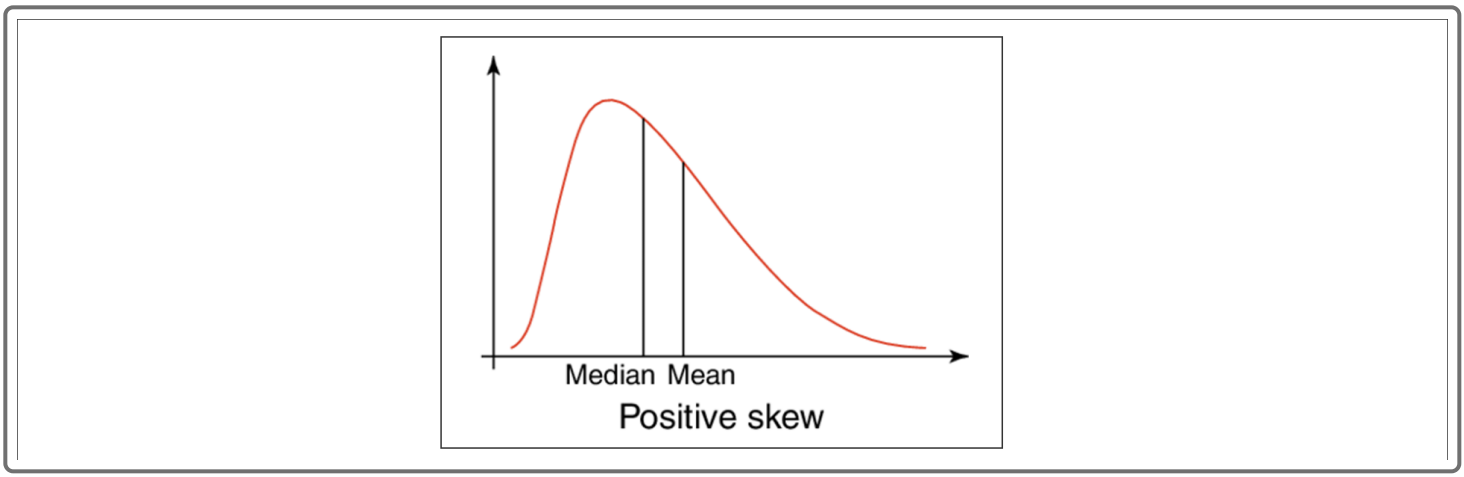
A data distribution is considered to be **left skewed,** or negative skewed, if the left tail is longer than the right, as shown below.



When data is skewed left, from the center of the distribution curve, there is a higher probability that extreme negative values exist within our dataset. When this occurs, the mean may no longer accurately reflect the central tendency of the data. Instead, we would use the median to describe the central tendency of the data. This skew is called negative skewed.

Right Skew

A data distribution is considered to be **right skewed**, or positive skewed, if the right tail is longer than the left, as shown below.



When data is skewed right, from the center of the distribution curve, there is a higher probability that extreme positive values exist within our dataset. Once again, if this occurs, we would use the median to describe the central tendency of the data. This skew is called positive skewed.

Manage Skewness

As with most problems in data analytics, we must approach skewness on a case-by-case basis. Depending on the severity of the skewness and the size of the dataset, there are multiple means of dealing (or not dealing) with skewness.

If our dataset is large, or the skewness is very subtle, we would simply point out that our data distribution shows signs of skew during reporting or presentation. In these cases, our mean and median will be roughly the same value, and there should be minimal impact to any downstream analysis.

If our dataset is smaller, or the skewness does impact the overall shape of our distribution, more action is needed. There are a few different things we can try:

- If possible, add more data points to our dataset to alleviate the effect of skew. However, this might not be possible or might not improve the distribution.
- Resample or regenerate data if we think that the data might not be representative of the original conditions or dataset.
- Transform our data values by normalization, using another numerical variable, or by transforming the data using an operator. The concept of transforming skewed data is very popular with scientists who deal with datasets where values can differ by orders of magnitude. One of the easiest means of transforming data is using a log-transform, where each value in the numeric dataset is transformed taking either natural log, or $\log_{10}$. By using a log-transformation, the effects of extreme values are reduced, and this transformation can help make each distribution tail more symmetrical.

IMPORTANT

No matter what approach is used to help reduce the skewness of a dataset, it's good practice to disclose this information in a report or to use annotations on your results. This will help the reader understand the context surrounding any results, and it will make your analysis more credible.