# Skewness and Kurtosis

## Introduction

We have previously identified a normal distribution to be symmetrical in shape. But when you're dealing with real-world data you'll often come across asymmetric distributions as well. In this lesson, you'll learn how to measure asymmetry (or skewness) in a distribution. Additionally, you'll learn about kurtosis. Kurtosis defines whether a distribution is truly "normal" or whether it may have so-called "fatter" or "thinner" tails than you would observe when data are normally distributed.

## Objectives

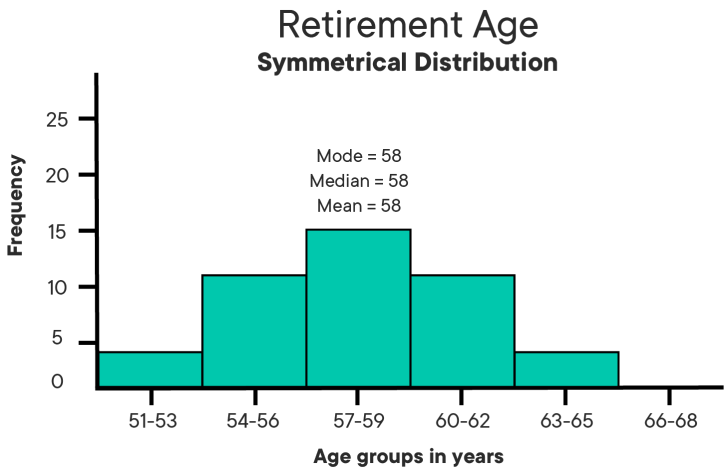
You will be able to:

- Define skewness and kurtosis and their relationship to symmetric distributions

## Symmetric Distributions

A distribution is symmetric if the relative frequency or probability of certain values are equal at equal distances from the point of symmetry. The point of symmetry for normal distributions is the mean (and at the same time median and mode!)

Have a look at following histogram:




This distribution meets all of the conditions of being symmetrical.

The most common symmetric distribution is the normal distribution, however, there are a number of other distributions that are symmetric. Here is a good article that looks into all sorts of symmetrical distributions. We'll focus on normal distributions (by far the most common group) here, and see how these can lose symmetry!

## Skewness

Skewness is the degree of distortion or deviation from the symmetrical normal distribution. Skewness can be seen as a measure to calculate the lack of symmetry in the data distribution.

Skewness helps you identify extreme values in one of the tails. Symmetrical distributions have a skewness of 0.

Distributions can be **positively** or **negatively** skewed.

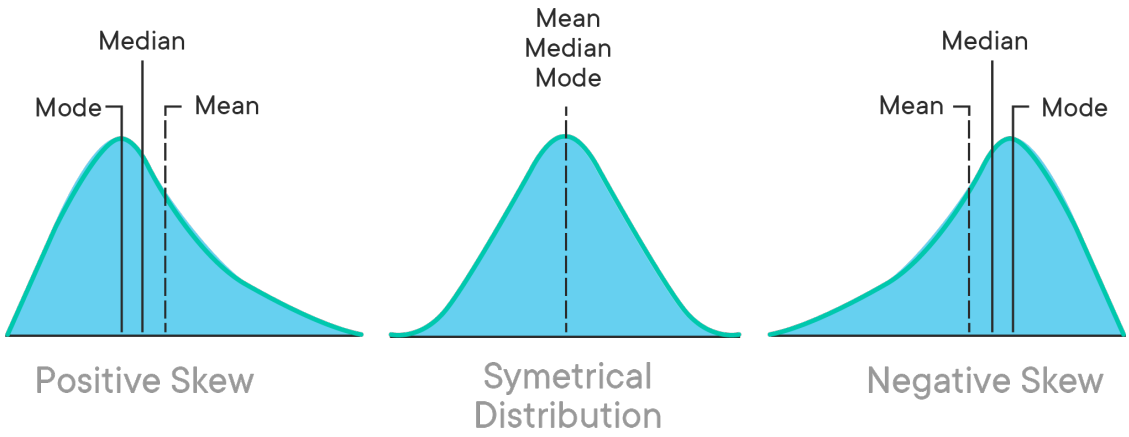
### Positive Skewness

A distribution is **positively skewed** when the tail on the right side of the distribution is longer (also often called "fatter"). When there is positive skewness, the mean and median are bigger than the mode.

### Negative Skewness

Distributions are **negatively skewed** when the tail on the left side of the distribution is longer or fatter than the tail on the right side. When there is negative skewness, the mean and median are smaller than the mode.

This behavior is shown in the images below:




Skewness can have implications for data analysis and the usage of certain models. The "normality assumption" seen before does not hold when data is skewed. When data is skewed, you'll need to transform the data first.

## Measuring Skewness

For univariate data $Y_1, Y_2, \ldots, Y_n$ the formula for skewness is:

$$\frac{\sum_{i=1}^{n}(Y_i - Y)^3}{\frac{n}{s^3}}$$

where $Y$ is the mean, $s$ is the standard deviation, and $n$ is the number of data points. This formula for skewness is referred to as the **Fisher-Pearson coefficient of skewness**. There are also other ways to calculate skewness, yet this one is the one that is used most commonly.

## Using this formula, when is data skewed?

The rule of thumb seems to be:

- A skewness between -0.5 and 0.5 means that the data are pretty symmetrical
- A skewness between -1 and -0.5 (negatively skewed) or between 0.5 and 1 (positively skewed) means that the data are moderately skewed.
- A skewness smaller than -1 (negatively skewed) or bigger than 1 (positively skewed) means that the data are highly skewed.

**Example**

Imagine you have house values ranging from 200,000 USD to 1,500,000 USD with an average of 800,000 USD.

If the peak of the distribution is left of the average value, the house prices are positively skewed. This means that more than half of the houses were sold for less than the average value 800,000 USD, and that there are a limited number of houses that were sold for a *much* higher value than 800,000 USD, leading to a long tail in the higher price ranges.

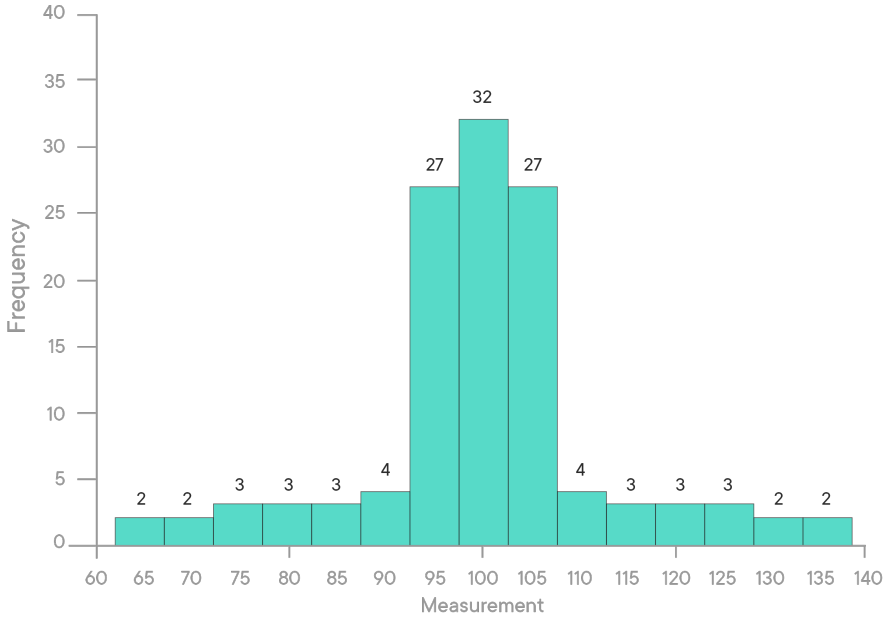
If the peak of the distributed data is on the right-hand side of the average value, this means there is negative skewness, meaning that more than half of the houses were sold for more than the average value of 800,000 USD. Additionally, this means that there is a long tail in the lower price ranges.

# Kurtosis

Kurtosis deals with the lengths of tails in the distribution.

> **Where skewness talks about extreme values in one tail versus the other, kurtosis aims at identifying extreme values in both tails at the same time!**

You can think of Kurtosis as a **measure of outliers** present in the distribution.




The distribution denoted in the image above has relatively more observations around the mean, then a steep decline and longer tails compared to the normal distribution.

## Measuring Kurtosis

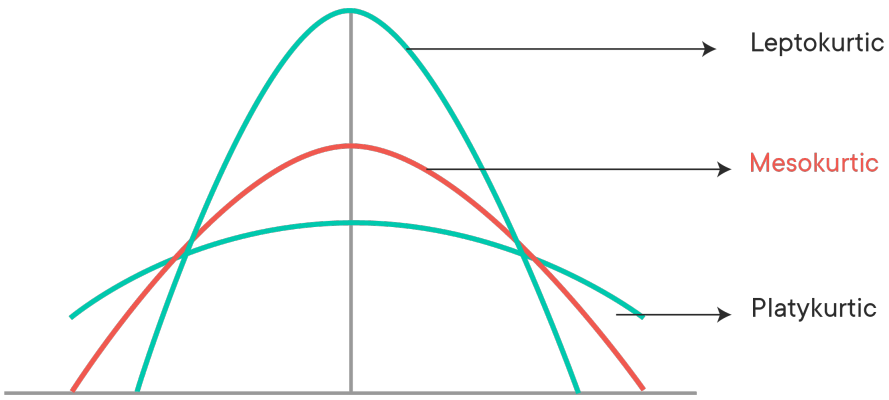
For univariate data $Y_1, Y_2, \ldots, Y_n$ the formula for kurtosis is:

$$\frac{\sum_{i=1}^{n}(Y_i - Y)^4}{\frac{n}{s^4}}$$

If there is a high kurtosis, then you may want to investigate why there are so many outliers. The presence of outliers could be indications of errors on the one hand, but they could also be some interesting observations that may need to be explored further. For banking transactions, for example, an outlier may signify fraudulent activity. How we deal with outliers mainly depends on the domain.

Low kurtosis in a data set is an indication that data has light tails or lacks outliers. If we get low kurtosis, then also we need to investigate and trim the dataset of unwanted results.

## How much kurtosis is bad kurtosis?



**Mesokurtic ($kurtosis \approx 3$)**

A mesokurtic distribution has kurtosis statistics that lie close to the ones of a normal distribution. Mesokurtic distributions have a kurtosis of around 3. According to this definition, the standard normal distribution has a kurtosis of 3.

**Platykurtic ($kurtosis < 3$):**

When a distribution is platykurtic, the distribution is shorter and tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that the tails are light and that there are fewer outliers than in a normal distribution.

**Leptokurtic ($kurtosis > 3$)**

When you have a leptokurtic distribution, you have a distribution with longer and fatter tails. The peak is higher and sharper than the peak of a normal distribution, which means that data have heavy tails and that there are more outliers.

Outliers stretch your horizontal axis of the distribution, which means that the majority of the data appear in a narrower vertical range. This is why the leptokurtic distribution looks "skinny".

## Summary

In this lesson, you learned about skewness and kurtosis. In the next lab, you'll learn how to measure skewness and kurtosis in Python.