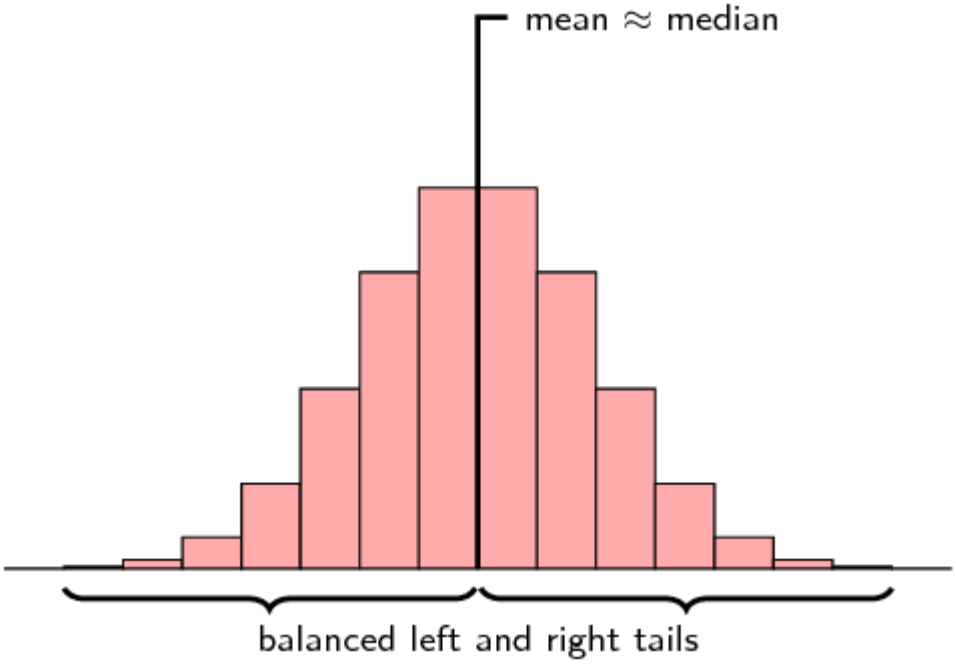
Data Distribution representation using Histogram

Normal Distribution

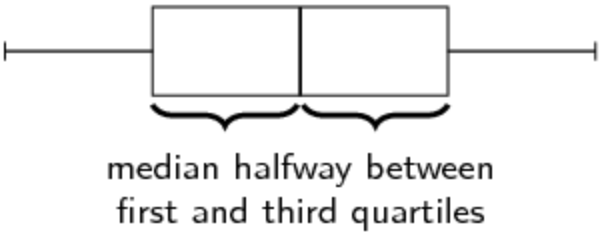
Symmetric and skewed data (EMBKD)

A symmetric distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean. The histogram below shows a typical symmetric distribution.



For symmetric distributions, the mean is approximately equal to the median. The **tails** of the distribution are the parts to the left and to the right, away from the mean. The tail is the part where the counts in the histogram become smaller. For a symmetric distribution, the left and right tails are equally balanced, meaning that they have about the same length.

The figure below shows the box and whisker diagram for a typical symmetric data set.

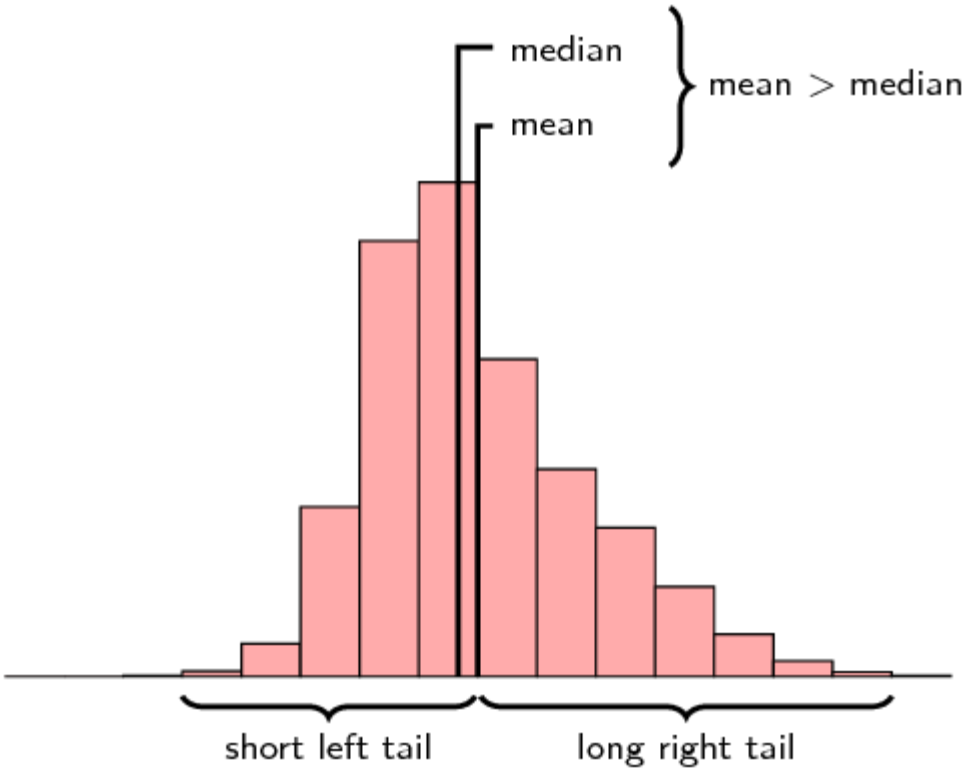


Another property of a symmetric distribution is that its median (second quartile) lies in the middle of its first and third quartiles. Note that

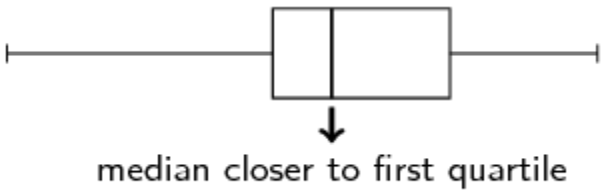
the whiskers of the plot (the minimum and maximum) do not have to be equally far away from the median. In the next section on outliers, you will see that the minimum and maximum values do not necessarily match the rest of the data distribution well.

Skewed (EMBKG)

A distribution that is **skewed right** (also known as **positively skewed**) is shown below.



Now the picture is not symmetric around the mean anymore. For a right skewed distribution, the mean is typically greater than the median. Also notice that the tail of the distribution on the right hand (positive) side is longer than on the left hand side.

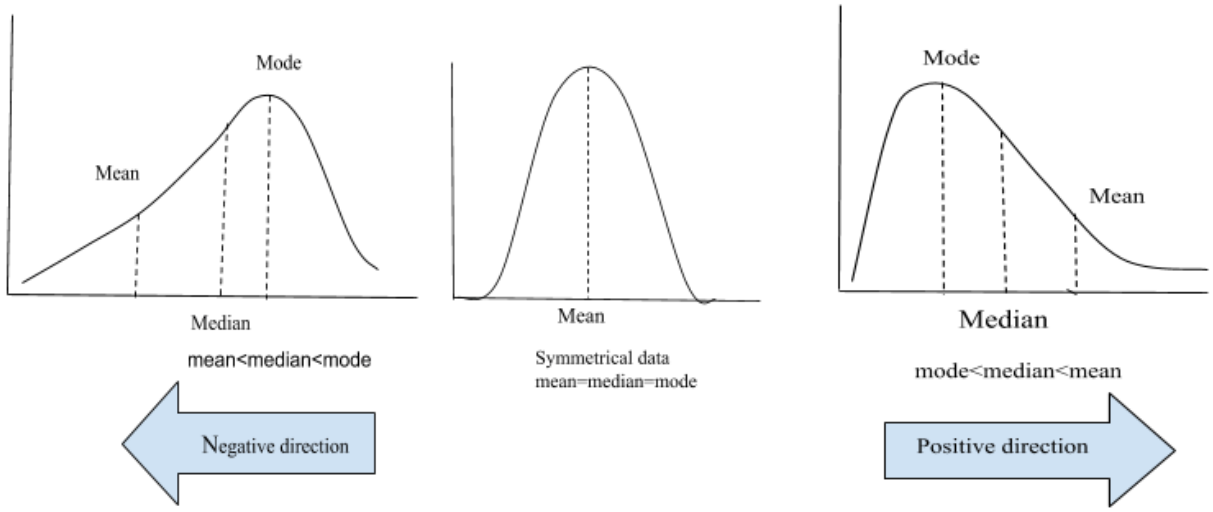


From the box and whisker diagram we can also see that the median is closer to the first quartile than the third quartile. The fact that the right hand side tail of the distribution is longer than the left can also be seen.

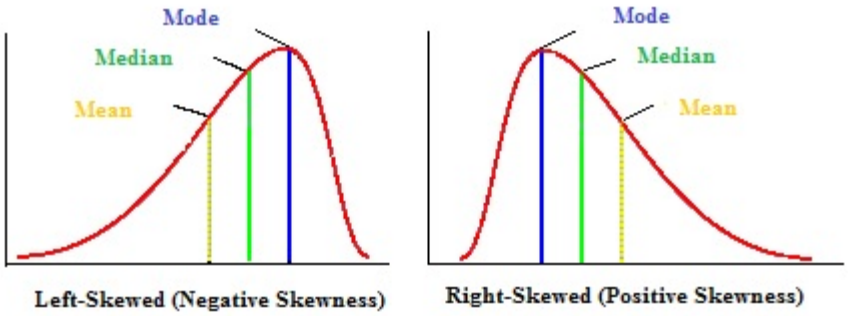
A distribution that is skewed left has exactly the opposite characteristics of one that is skewed right:

- the mean is typically less than the median;
- the tail of the distribution is longer on the left hand side than on the right hand side; and
- the median is closer to the third quartile than to the first quartile.

Skewness



- The mean, median and mode are all Centrality measures (center of a set of data).
- The skewness of the data can be determined by how these quantities are related to one another
- By studying the shape of the data we can discover the relation between the mean, median and mode
- If the mean > median it indicates that the distribution is positively skewed.
  If the mean is < median it indicates that the distribution is negatively skewed



-

- Kal Pearson's first method uses mode and it's formula $S_K = \frac{mean - mode}{standard\ deviation}$

- However, this method is not considered very stable in data's where mode is made up of too few pieces as it won't be considered a very strong measure of central tendency.

- For example in the first data set below 8 only occurs twice, so while using Pearson's first formula of skewness you have to be cautioned as it won't be a good measure of central tendency.

- Set 1 = [1, 2, 3, 4, 5, 8, 7, 8]

- However in the second set you can see that 8 appears ten times thus, you can use the Pearson's measure of skewness as you know it will give you a more stable and reliable result.

- Set 2= [1, 5, 6, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8]

- The second formula of skewness uses the median and is denoted (By Kal Pearson)

$$S_K = \frac{3(mean - median)}{standard\ deviation}$$

Karl Pearson's coefficient of skewness lies between -3 and +3.

- If SK = 0 then we can say that the frequency distribution is normal and symmetrical.
- If SK < 0 then we can say that the frequency distribution is negatively skewed.
- If SK > 0 then we can say that the frequency distribution is positively skewed.

Example

Q. The age (in years) of 6 randomly selected students from a class are:

[22, 25, 24, 23, 24, 20]

Find the Karl Pearson's coefficient of skewness.

Solution:

STEP#1

We will first find the mean.

REMEMBER: For mean we first add all the data together and then divide it by the total number of numbers.

$$\bar{X} = \frac{138}{6} = 23 \quad years$$

STEP#2

We will now find the median.

REMEMBER: For median we pick the middle value of the set $M = \frac{X1 + X2}{2}$

$$M = \frac{24 + 23}{2} = 23.5 \quad years$$

STEP#3

Find the variance first and then take its unroot for Standard deviation.

$$= \frac{1}{5}(3190 - \frac{138^2}{6})$$
$$= \frac{1}{5}(3190 - \frac{19044}{6})$$
$$= \frac{1}{5}(3190 - 3174)$$
$$= \frac{16}{5}$$
$$s_x = \sqrt{s_x^2}$$
$$= \sqrt{3.2}$$
$$= 1.7889 \quad years$$

STEP#4

Put it all into Pearson's equation to get:

As you can tell the value of SK< 0 thus we can say that the data is negatively skewed.

What Is the Interquartile Range Rule?

What Is the Interquartile Range?

Any set of data can be described by its five-number summary. These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

- The minimum or lowest value of the dataset
- The first quartile Q_1, which represents a quarter($1/4^{th}$ Position) of the way through the list of all data
- The median of the data set, which represents the midpoint of the whole list of data
- The third quartile Q_3, which represents three-quarters of the way through the list of all data ($3/4^{th}$ Position)
- The maximum or highest value of the data set.

These five numbers tell a person more about their data than looking at the numbers all at once could, or at least make this much easier.

- For example, the range, which is the minimum subtracted from the maximum, is one indicator of how spread out the data is in a set
    - note: the range is highly sensitive to outliers—if an outlier is also a minimum or maximum, the range will not be an accurate representation of the breadth of a data set
    - Range would be difficult to extrapolate otherwise.

IQR

Similar to the range but less sensitive to outliers is the interquartile range.

- The interquartile range is **calculated in much the same way as the range**.
    - All you do to find it is **subtract the first quartile from the third quartile**:

$$\text{IQR} = Q_3 - Q_1.$$

- The interquartile range shows how the data is spread about the median.
- It is less susceptible than the range to outliers and can, therefore, be more helpful.

Using the Interquartile Rule to Find Outliers

Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Note :

Remember that the interquartile rule is only a rule of thumb that generally holds but does not apply to every case. In general, you should always follow up your outlier analysis by studying the resulting outliers to see if they make sense. Any potential outlier obtained by the interquartile method should be examined in the context of the entire set of data.

IQR Example

Given Set of data: 1, 3, 4, 6, 7, 7, 8, 8, 10, 12, 17.

The five-number summary for this data set is

- minimum = 1,
- first quartile = 4,
- median = 7,
- third quartile = 10 and
- maximum = 17.

*** You may look at the data and automatically say that 17 is an outlier, but what does the interquartile range rule say?

IQR for the given Data :

- $Q_3 - Q_1 = 10 - 4 = 6$

Now

- multiply your answer by 1.5 to get 1.5 x 6 = 9.
- (1^{st} QR – 9) = 4 – 9 = -5. No data is less than this.
- (3^{rd} QR + 9) 10 + 9 =19. No data is greater than this.
- Despite the maximum value being five more than the nearest data point, the interquartile range rule shows that it should probably not be considered an outlier for this data set.