



The central limit theorem

Let's have a look at the things we are about to discuss in this article:

- What is a central limit theorem and why is it important?
- Measures of descriptive statistics

What is a central limit theorem and why is it important?

The central limit theorem in statistics asserts that, given a big enough sample size, the sampling distribution of a variable's mean will approximate a normal distribution, regardless of the distribution of that variable in the population.

The Central Limit Theorem is significant in statistics because it allows us to assume that the mean sample distribution will be normal in the vast majority of circumstances. This means we can use statistical techniques based on the assumption of a normal distribution.

Measures of descriptive statistics

In the previous article we have covered that there are three measures of descriptive statistics. They are:

1. Measures of Central Tendency – Mean, Median and Mode
2. Measures of Dispersion – Standard Deviation, Variance, Range, IQR (Inter Quartile Range)
3. Measure of Symmetricity – Skewness and Kurtosis

1. Measure of Central Tendency

A measure of central tendency is a summary statistic that represents the centre 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.

Learnvista Pvt Ltd.

2nd Floor, 147, 5th Main Rd, Rajiv Gandhi Nagar HSR Sector 7, Near Salarpuria Serenity, Bengaluru, Karnataka 560102

Mob:- +91 779568798, Email:- contacts@learnbay.co



● Mean

Average value of the set of Numbers. Mean is a number around which a whole data is spread out. Denoted by population mean and for sample mean.

Ex: Find the mean of 5,5,2,6,3,8,9?

Mean is $(5+5+2+6+3+8+9)/7 = 38/7 = 5.43$

● Median

Median is the value which divides the data in 2 equal parts i.e. number of terms on the right side of it is the same as number of terms on the left side of it when data is arranged in either ascending or descending order.

(Note: If you sort data in descending order, it won't affect median but IQR will be negative.)

Ex: Find the Median of 5,5,2,6,3,8,9?

Putting it in ascending order = 2,3,5,5,6,8,9. Hence, Median = Mid Number = 5.

(Note: Median of an even set of numbers can be found by taking the average of the 2 middle numbers.)

● Mode

Mode is the term appearing maximum time in a data set i.e. the term that has the highest frequency.

Example: Find the Median of 5,5,2,6,3,8,9?

Mode = Maximum number of repetition in dataset = 5. Hence, Mode = 5.

(Note: If there is no repetition of data then mode is not present.)

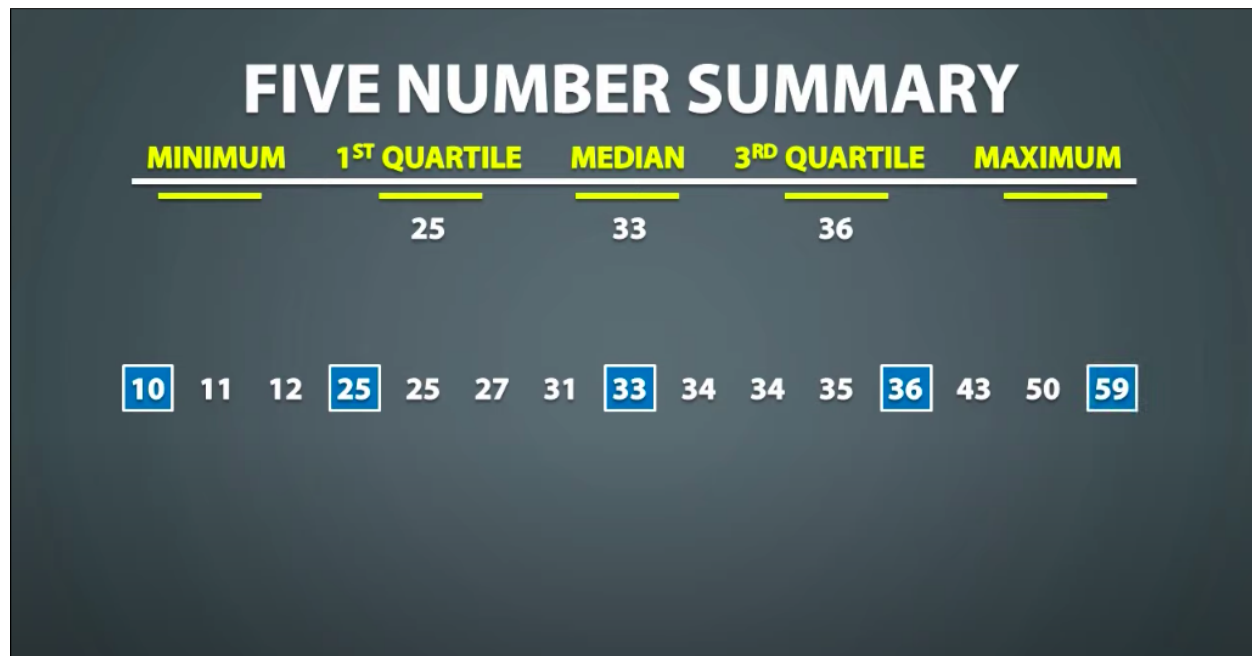
Let's build some concepts before going ahead.

1. What is Minimum and Maximum value?

It is the minimum and Maximum values of the dataset respectively.

2. What is the 1st and 3rd Quartile?

Also called the lower and upper quartile respectively. When we divide the dataset into two groups while calculating median (sorted in ascending order), then the median of first half is 1st Quartile and median of second half is 3rd Quartile.



Let's look at an example,

Given the ages of people registered for a webinar, calculate the 5 point summary (5 number summary) of the ages of the participants?

19, 26, 25, 37, 32, 28, 22, 23, 29, 34, 39, 31

Step 1: Sort in ascending order

19, 22, 23, 25, 26, 28, 29, 31, 32, 34, 37, 39

Step 2: Find the median

Since the value of numbers in the list is 12 and it is an even number the median is the average of the 6th and 7th number $(28+29)/2=28.5$

Step 3: Find the first quartile

Median of the first half is the first quartile, so $(23+25)/2=24$.

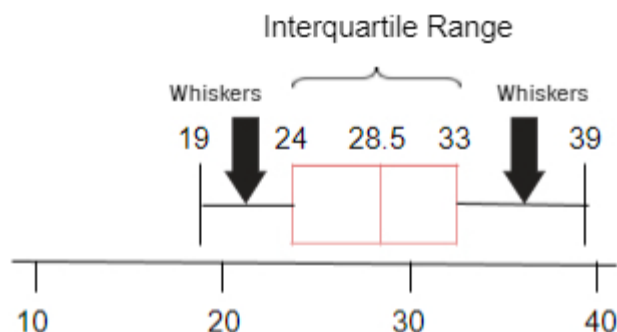
Step 4: Find the 3rd quartile

Median of the second half is the third quartile, so $(32+34)/2 = 33$.

Step 5: Pick minimum and maximum from the list

The minimum value in the list is 19 and the maximum in the list is 39.

Let's draw a boxplot that visually represents the five number summary.



2. Measure of Spread / Dispersion

1. Standard deviation

Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

$$\text{S.D.} = \sqrt{\frac{1}{n} \sum_{i=0}^n (x - \mu)^2}$$

SD of Population(σ)

Population STD = `pstdev()`

In Python :

$$\text{S.D.} = \sqrt{\frac{1}{n-1} \sum_{i=0}^n (x - \bar{x})^2}$$

SD of Sample(s)

Sample STD = `stdev()`

2. Variance

Variance is a square of average distance between each quantity and mean. That is, it is a square of standard deviation.



$$\text{VAR} = \frac{1}{n} \sum_{i=0}^n (x - \mu)^2$$

Population Variance(σ^2)

Population Var = pvariance()

In Python :

$$\text{VAR} = \frac{1}{n-1} \sum_{i=0}^n (x - \bar{x})^2$$

Sample Variance(S^2)

Sample Variance = variance()

3. Range

Range is one of the simplest techniques of descriptive statistics. It is the difference between lowest and highest value.

Range = Maximum - Minimum

4. IQR (Interquartile Range)

In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in an ascending order.

IQR = Q3 – Q1

Steps to find out the IQR

1. Order the data from least to greatest
2. Find the median
3. The left side of the median is the lower half and the right side of the data is the upper half.
4. Calculate the median of both the lower and upper half of the data (Called Q1 and Q3 respectively)
5. The IQR is the difference between the upper and lower medians

(Note: When we write down Minimum, Maximum, Q1, Q2 (Median) and Q3, this is called 5-point summary or 5 number summary)

Let's solve some questions to find IQR.

1. Below are the weights of 5 persons. Calculate Mean, Standard Deviation :



105, 156, 145, 172, 100

Mean = $(105 + 156 + 145 + 172 + 100) / 5 = 678 / 5 = 135.6$

$$\text{S.D.} = \sqrt{\frac{1}{n-1} \sum_{i=0}^n (x - \bar{x})^2}$$

Standard deviation =

$$\begin{aligned} \sum (x - \bar{x})^2 &= \\ (105 - 135.6)^2 + (156 - 135.6)^2 + (145 - 135.6)^2 + (172 - 135.6)^2 + (100 - 135.6)^2 \\ &= 1362.72 \\ n &= 5 \end{aligned}$$

$$\text{S.D.} = \sqrt{\frac{1}{n-1} \sum_{i=0}^n (x - \bar{x})^2}$$

= 18.45

2. Suppose each one of them gained extra 5 Kg. weight during winters. Can you calculate the new Mean and Standard deviation?

Since they gained 5kg weight during winter the new sample is

110, 161, 150, 177, 105

The new mean is

$(110 + 161 + 150 + 177 + 105) / 5 = 140.6$

$$\text{S.D.} = \sqrt{\frac{1}{n-1} \sum_{i=0}^n (x - \bar{x})^2}$$

Standard deviation =

$$\begin{aligned} \sum (x - \bar{x})^2 &= \\ (110 - 140.6)^2 + (161 - 140.6)^2 + (150 - 140.6)^2 + (177 - 140.6)^2 + (105 - 140.6)^2 \\ &= 4033.2 \\ n &= 5 \end{aligned}$$



$$\text{S.D.} = \sqrt{\frac{1}{n-1} \sum_{i=0}^n (x - \bar{x})^2}$$

=31.75

Data transformation Guidelines

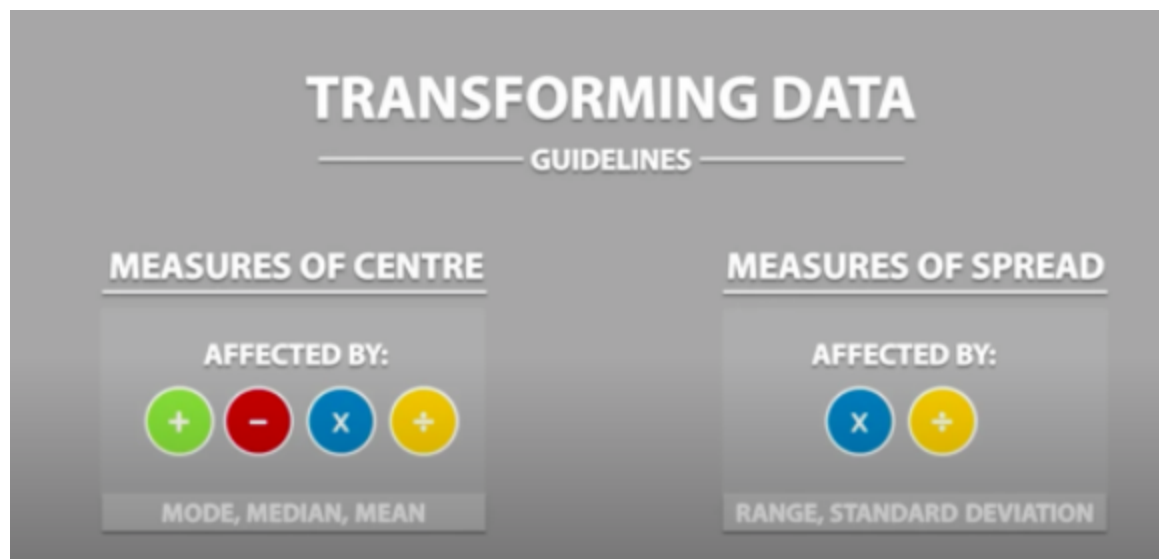
Let us list the factors that affect the measures of center and measures of spread.

Factors that affect the measures of center:

- Mode
- Median
- Mean

Factors that affect the measures of spread:

- Range
- Standard Deviation





3. Measure of symmetry – Skewness and Kurtosis

1. Skewness

Skewness is usually described as a measure of a dataset's symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0. The normal distribution has a skewness of 0. Skewness is calculated as:

```
import numpy as np
```

```
from scipy.stats import skew
```

```
x = np.random.normal(0, 2, 10000) # create random values based on a normal distribution
```

```
print(skew(x))
```

Mathematically:

$$a_3 = \sum \frac{(X_i - \bar{X})^3}{ns^3}$$

where n is the sample size, X_i is the i^{th} X value, \bar{X} is the average and s is the sample standard deviation. Note the exponent in the summation. It is “3”. The skewness is referred to as the “third standardized central moment for the probability model.”

So, when is the skewness too much?

The rule of thumb seems to be:

1. If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.



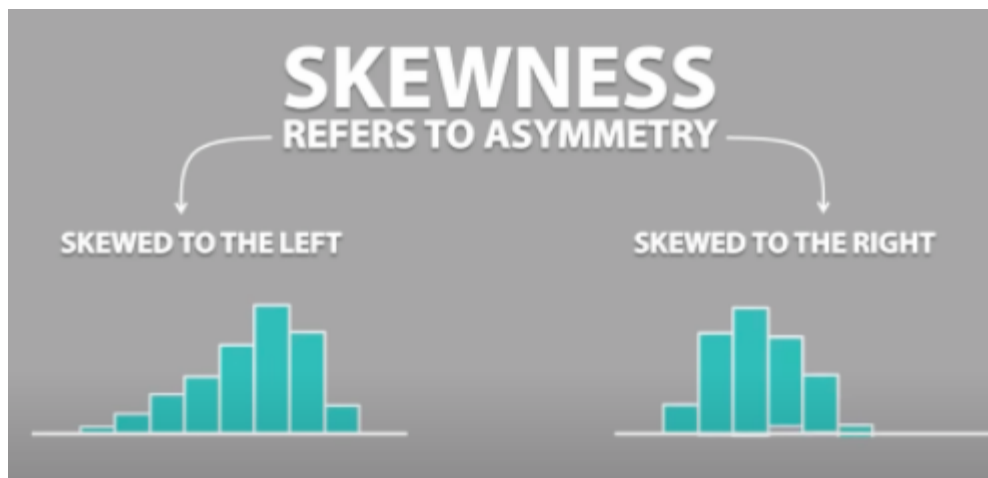
2. If the skewness is between -1 and -0.5 or between 0.5 and 1, the data are moderately skewed.
3. If the skewness is less than -1 or greater than 1, the data are highly skewed.

Importance of Skewness:

Measures of asymmetry like skewness are the link between central tendency measures and probability theory, which ultimately allows us to get a more complete understanding of the data we are working with.

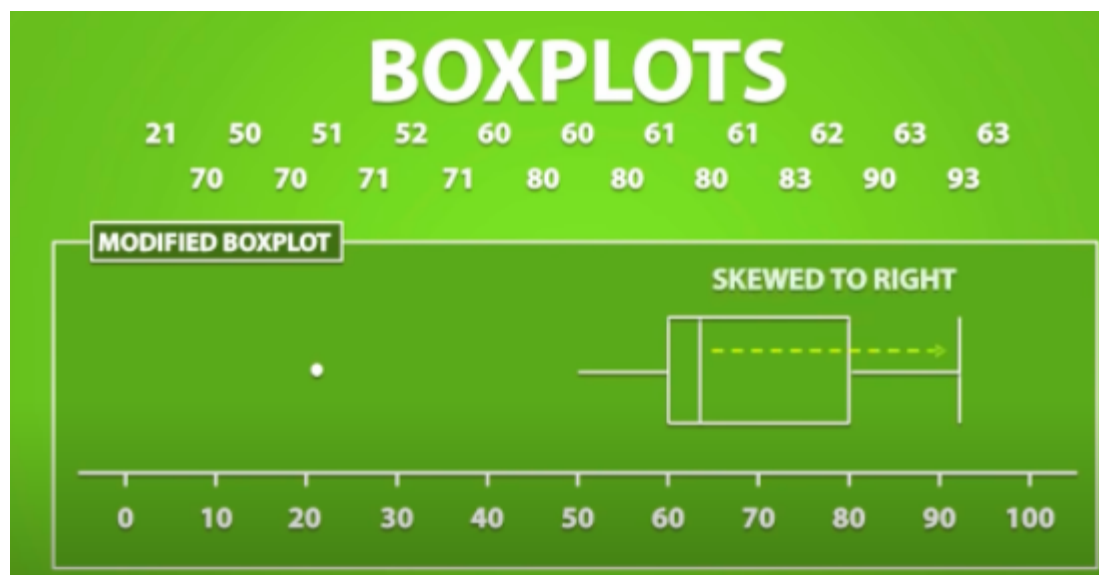
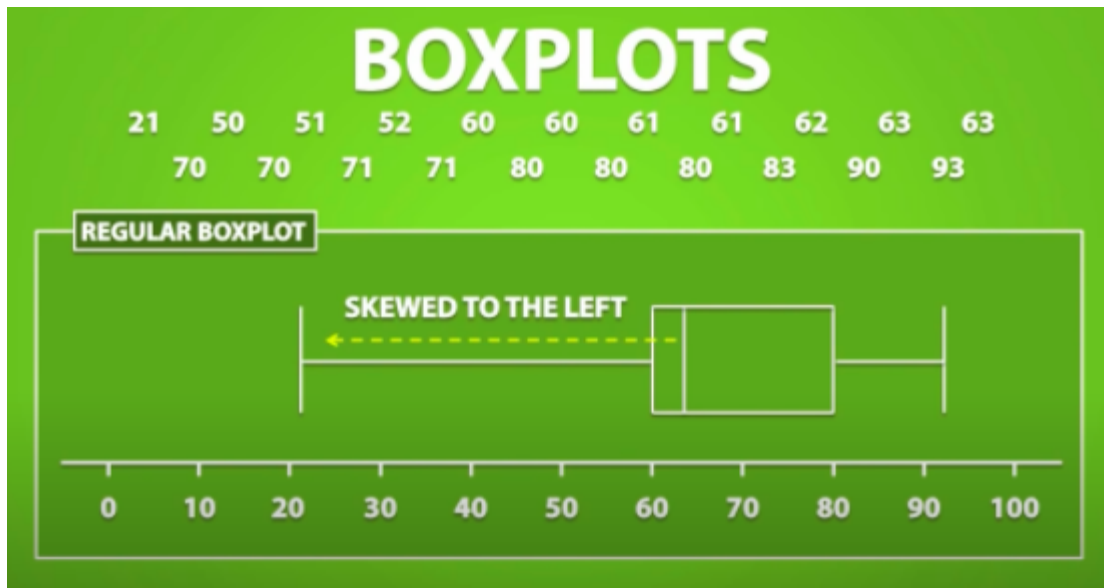
Knowing that the market has a 70% probability of going up and a 30% probability of going down may appear helpful if you rely on normal distributions. However, if you were told that if the market goes up, it will go up 2% and if it goes down, it will go down 10%, then you could see the skewed returns and make a better informed decision.

$$E(r) = 0.7 \cdot 0.02 + 0.3 \cdot -0.1 = -0.014$$





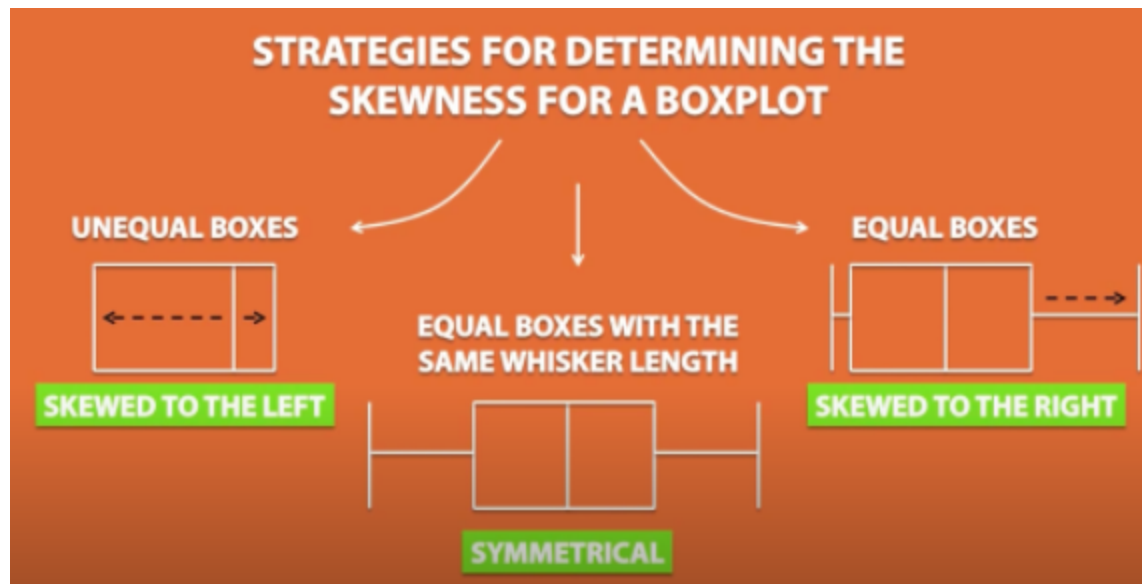
Let's take a look at how the boxplots are for the skewness.



Strategies for determining the skewness for a boxplot

There are 3 strategies for determining the skewness for a boxplot:

- Unequal boxes
- Equal boxes
- Equal boxes with the same whisker length.



2. Kurtosis

Kurtosis is all about the tails of the distribution – not the peakness or flatness. It measures the tail-heaviness of the distribution. Kurtosis is calculated as:

```
import numpy as np
```

```
from scipy.stats import kurtosis
```

```
x = np.random.normal(0, 2, 10000) # create random values based on a normal distribution
print(kurtosis(x))
```

Mathematically:

$$a_4 = \sum \frac{(X_i - \bar{X})^4}{ns^4}$$

where n is the sample size, X_i is the i^{th} X value, \bar{X} is the average and s is the sample standard deviation. Note the exponent in the summation. It is “4”. The kurtosis is referred to as the “fourth standardized central moment for the probability model.”

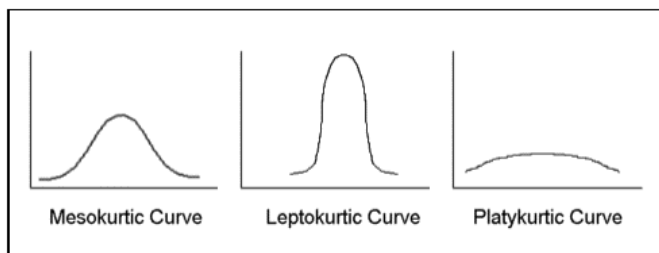
Note: Kurtosis calculated by Excel or through Python/R is actually excess kurtosis, which is (Kurtosis – 3)



What does the value of Kurtosis tell about the shape?

The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the excess kurtosis is presented: excess kurtosis is simply $\text{kurtosis} - 3$. For example, the “kurtosis” reported by Excel or any statistical library is actually the excess kurtosis.

1. A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈ 3 (excess ≈ 0) is called mesokurtic.
2. A distribution with kurtosis < 3 (excess kurtosis < 0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
3. A distribution with kurtosis > 3 (excess kurtosis > 0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.



Uses of Kurtosis:

1. Depicts the shape of the distribution - specially tails.
2. Outlier Detection : Large Kurtosis suggests there could be outliers in the data.
3. With high kurtosis, there is a chance of high variance and hence tests on Mean could lead to bad results. Hence, in that case, we would need to choose a more robust option – like a test on Median.
4. Financial Risk: E.g. The return of your asset can be farther from the mean.

3. Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal.

Common Causes of Outliers

1. Data entry errors (human errors)
2. Measurement errors (instrument errors)

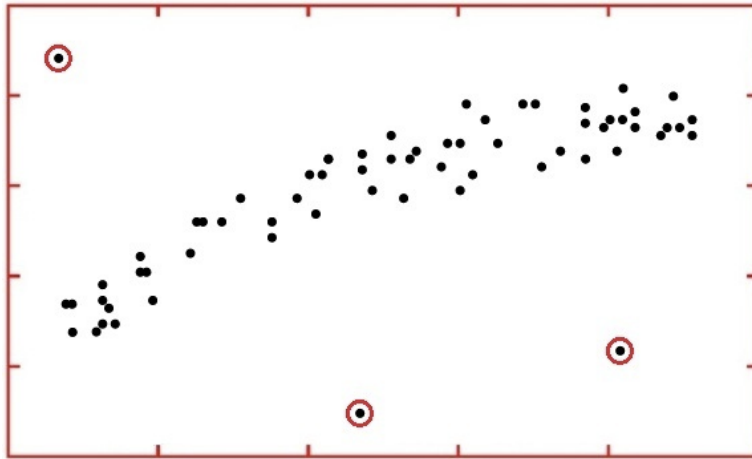
Learnvista Pvt Ltd.

2nd Floor, 147, 5th Main Rd, Rajiv Gandhi Nagar HSR Sector 7, Near Salarpuria Serenity, Bengaluru, Karnataka 560102

Mob:- +91 779568798, Email:- contacts@learnbay.co

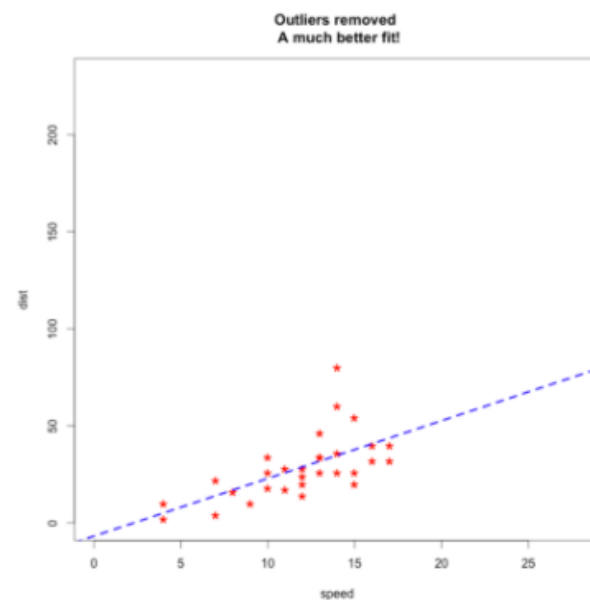
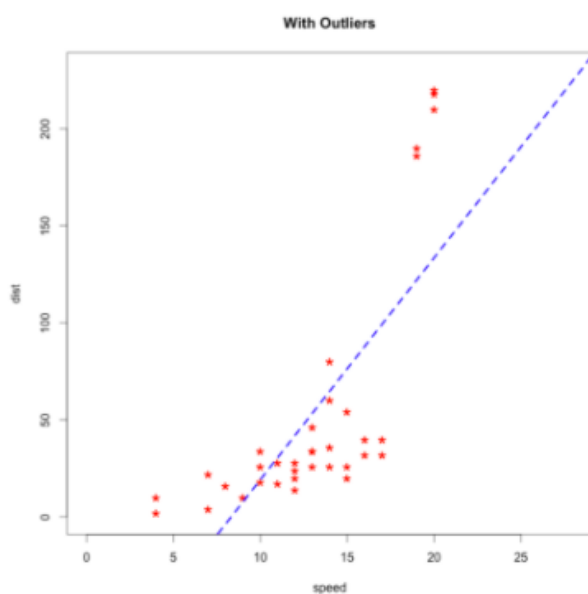


3. Experimental errors (data extraction or experiment planning/executing errors)
4. Intentional (dummy outliers made to test detection methods)
5. Data processing errors (data manipulation or data set unintended mutations)
6. Sampling errors (extracting or mixing data from wrong or various sources)
7. Natural (not an error, novelties in data)



Common methods of determining an Outlier

1. Sort the data and see for the extreme values
2. Plotting – Boxplot, Scatterplot
3. IQR Method
4. Z-Score Method



Learnvista Pvt Ltd.

2nd Floor, 147, 5th Main Rd, Rajiv Gandhi Nagar HSR Sector 7, Near Salarpuria Serenity, Bengaluru, Karnataka 560102

Mob:- +91 779568798, Email:- contacts@learnbay.co




Why do we need to treat outliers?


Outliers can impact the results of our analysis and statistical modelling in a drastic way.

IQR Method

A DATA VALUE IS CONSIDERED TO BE AN OUTLIER IF..

DATA VALUE  $Q1 - 1.5(IQR)$

OR

DATA VALUE  $Q3 + 1.5(IQR)$

Q. Can you identify the outliers from the below dataset, using the IQR method?

26.0 °C , 15.0 °C , 20.5 °C , 31 °C , -350.0 °C , 31.0 °C , 30.5 °C

In ascending order,

-350,15,20.5,26,30.5,31,31

The median of this dataset is 26

Now to find the quartile1 find the median of the series before the number 26.

The median here is 15. So the value of quartile1 is 15.

Similarly, to find the quartile3 find the median of the series after 26.

The median here is 31. So the value of quartile3 is 31.

We already know $IQR = Q3 - Q1$,

So, $IQR = 31 - 15 = 16$.

Let's find the value of

$Q1 - 1.5(IQR) = 15 - 1.5(16) = -9$

$Q3 + 1.5(IQR) = 31 + 1.5(16) = 55$

Now comparing these values to the dataset we can clearly identify the outlier is -350.



4. Z-Score Method

Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls, assuming a Normal distribution.

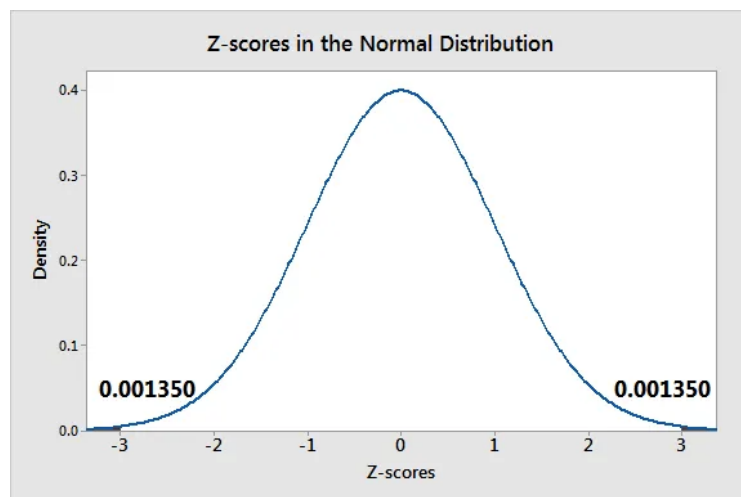
For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean.

Z-Score Formula?

$$Z = \frac{X - \mu}{\sigma}$$

Point to understand:

The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of +/-3 or further from zero. In a population that follows the normal distribution, Z-score values more extreme than +/- 3 have a probability of 0.0027 (2 * 0.00135), which is about 1 in 370 observations.



Example: Consider the below dataset. Find out the outlier using the Z-score method.

1, 2, 2, 2, 3, 1, 1, 15, 2, 2, 2, 3, 1, 1, 2

Solution:



Mean = 2.66

Std = 3.36

$Z(1) = (1 - 2.66)/3.36 = -0.49405$

$Z(2) = (2 - 2.66)/3.36 = -0.19643$

$Z(3) = (3 - 2.66)/3.36 = 0.10119$

$Z(15) = (15 - 2.66)/3.36 = 3.67262$

We will term the point outlier if it has a z-score of 3 or above (in any side - positive or negative).
Hence, here the outlier is 15.