Intermediate Stats

Quartile vs Quantile vs Percentile

If you get confused between quartile vs quantile vs percentile
1 quartile = 0.25 quantile = 25 percentile.

- ★ A quartile is a statistical term that describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations.
- ★ A quantile defines a particular part of a data set, i.e. a quantile determines how many values in a distribution are above or below a certain limit.
- ★ A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

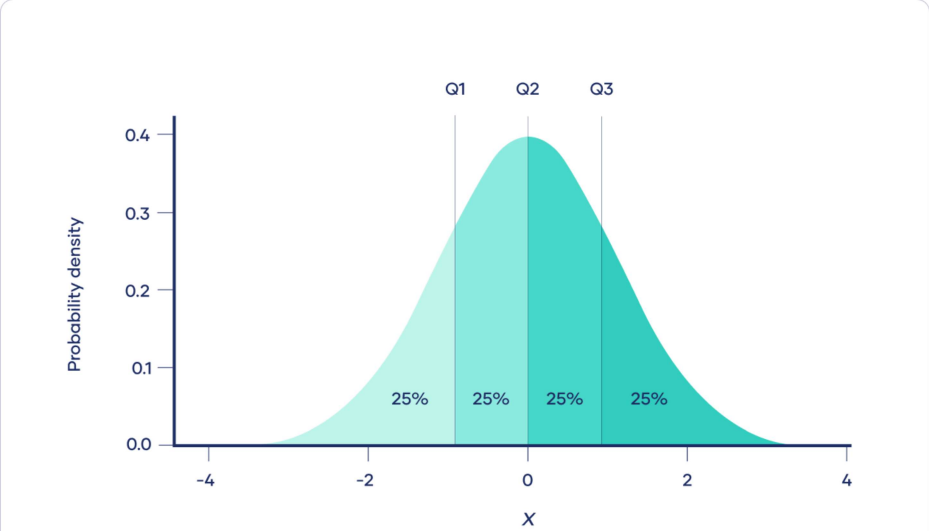
Quartiles

Quartiles are a set of descriptive statistics. They summarize the central tendency and variability of a dataset or distribution. Quartiles are a type of percentile. A percentile is a value with a certain percentage of the data falling below it. In general terms, k% of the data falls below the kth percentile.

- ➢ The **first quartile** (Q1, or the lowest quartile) is the 25th percentile, meaning that 25% of the data falls below the first quartile.
- ➢ The **second quartile** (Q2, or the median) is the 50th percentile, meaning that 50% of the data falls below the second quartile.
- ➢ The **third quartile** (Q3, or the upper quartile) is the 75th percentile, meaning that 75% of the data falls below the third quartile.

By splitting the data at the 25th, 50th, and 75th percentiles, the quartiles divide the data into four equal parts.

- In a **Sample or dataset**, the quartiles divide the data into four groups with equal numbers of observations.
- In a **Probability distribution**, the quartiles divide the distribution's range into four intervals with equal probability.



Quantiles

A quartile is a type of quantile. Quantiles are values that split sorted data or a probability distribution into equal parts. In general terms, a q-quantile divides sorted data into q parts. The most commonly used quantiles have special names:

- **Quartiles (4-quantiles):** Three quartiles split the data into four parts.
- **Deciles (10-quantiles):** Nine deciles split the data into 10 parts.
- **Percentiles (100-quantiles):** 99 percentiles split the data into 100 parts.

Percentiles

A value on a scale of one hundred that indicates the percent of a distribution that is equal to or below it. A percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular score surpassed.

For example, if you score 75 points on a test, and are ranked in the 85th percentile, it means that the score 75 is higher than 85% of the scores.

Data Set:

2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

What is the percentile ranking of '10'?

$$Percentile\ Rank\ of\ x = \frac{\#\ of\ values\ below\ x}{n} * 100$$

$$Percentile\ Rank\ of\ '10' = \frac{16}{20} * 100 = 80\%$$

Five Number Summary

Descriptive Statistics involves understanding the distribution and nature of the data. Five number summary is a part of descriptive statistics and consists of five values and all these values will help us to describe the data.

- The minimum value (the lowest value)
- 25th Percentile or Q1
- 50th Percentile or Q2 or Median
- 75th Percentile or Q3
- Maximum Value (the highest value)

How to calculate Five Number Summary

Step 1: Put your numbers in ascending order (from smallest to largest). For this particular data set, the order is:
Example: 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.

Step 2: Find the minimum and maximum for your data set. Now that your numbers are in order, this should be easy to spot.
In the example in step 1, the minimum (the smallest number) is 1 and the maximum (the largest number) is 27.

Step 3: Find the median. The median is the middle number. If you aren't sure how to find the median, see: How to find the mean mode and median.

Step 4: Place parentheses around the numbers above and below the median.
(This is not technically necessary, but it makes Q1 and Q3 easier to find).
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).

Step 5: Find Q1 and Q3. Q1 can be thought of as a median in the lower half of the data, and Q3 can be thought of as a median for the upper half of data.
(1, 2, 5, 6, 7), 9, (12, 15,18,19,27).

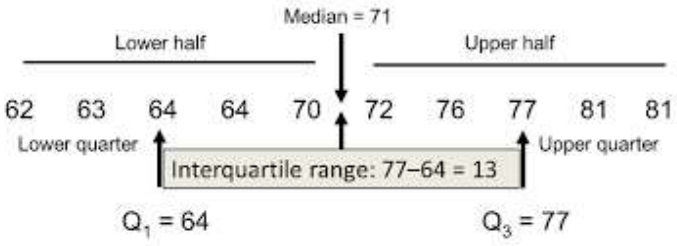
Step 6: Write down your summary found in the above steps.
minimum = 1, Q1 = 5, median = 9, Q3 = 18, and maximum = 27.

Interquartile Range (IQR)

- The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

- The distance between the first and third quartiles—the interquartile range (IQR)—is a measure of variability. It indicates the spread of the middle 50% of the data.

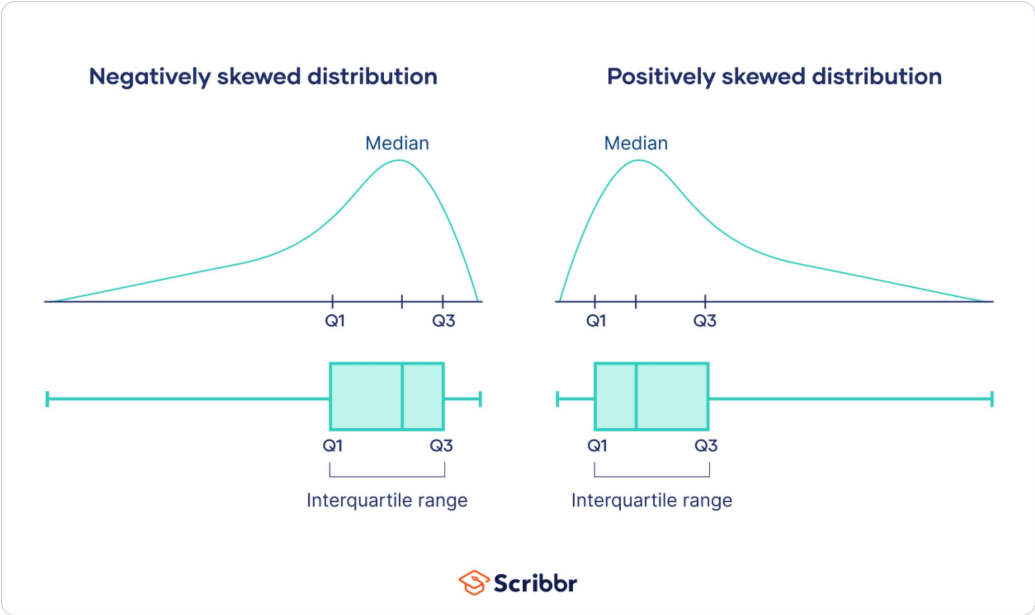
IQR = Q3 – Q1



➢ The IQR is an especially good measure of variability for skewed distributions or distributions with outliers. IQR only includes the middle 50% of the data, so, unlike the range, the IQR isn't affected by extreme values.

Skewness

The distance between quartiles can give you a hint about whether a distribution is skewed or symmetrical. It's easiest to use a boxplot to look at the distances between quartiles:

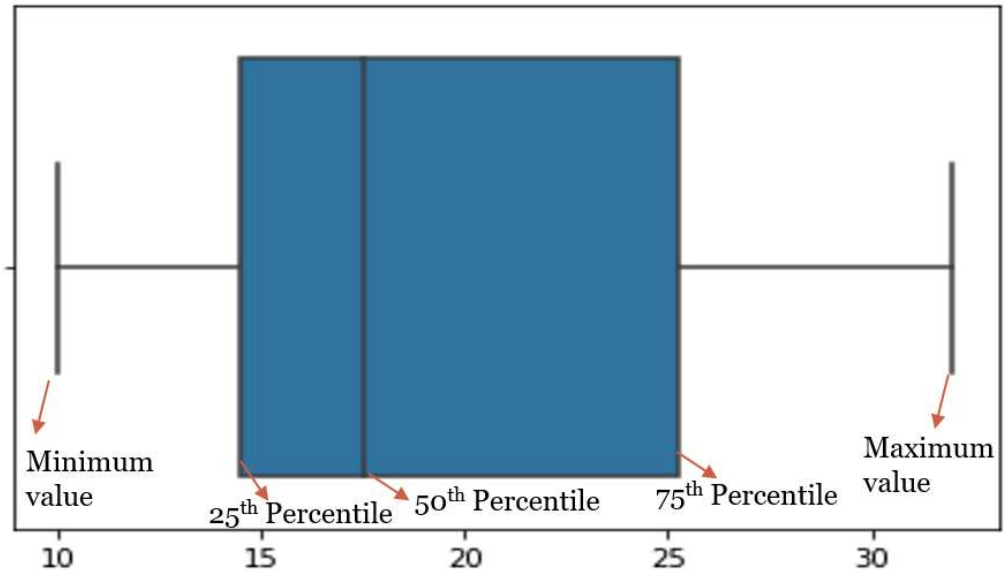


Click to enlarge

More about Skewness : [Skewness Article - Scribbr]

Box plots and how they are constructed?

Boxplots are the graphical representation of the distribution of the data using Five Number summary values. It is one of the most efficient ways to detect outliers in our dataset.



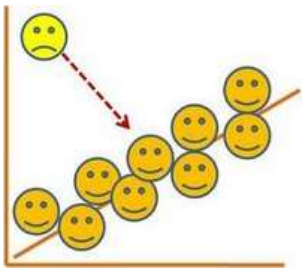
Plotting the boxplot of the data points taken for the above example (11,23,32,26,16,19,30,14,16,10) where the Five number summary was :

- Minimum value: 10
- 25th Percentile: 14
- 50th Percentile: 17.5
- 75th Percentile: 26
- Maximum value: 32

Effect Of Outliers And Its Removal

➢ In statistics, an outlier is a single data point that goes far outside the average value of a group of statistics. An outlier may be due to variability in the measurement or it may indicate experimental error / human error. An outlier can cause serious problems in statistical analyses.

➢ Outliers increase the variability in your data, which decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant.



What do they affect?

In statistics, we have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

- Mean is the accurate measure to describe the data when we do not have any outliers present.
- Median is used if there is an outlier in the dataset.
- Mode is used if there is an outlier AND about ½ or more of the data is the same.

'Mean' is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.

Example:

Consider a small dataset, sample= **[15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9].** By looking at it, one can quickly say '101' is an outlier that is much larger than the other values.

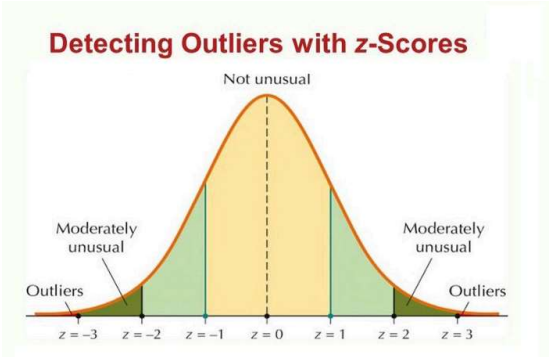
with outlier	without outlier
Mean: 20.08	Mean: 12.72
Median: 14.0	Median: 13.0
Mode: 15	Mode: 15
Variance: 614.74	Variance: 21.28
Std dev: 24.79	Std dev: 4.61

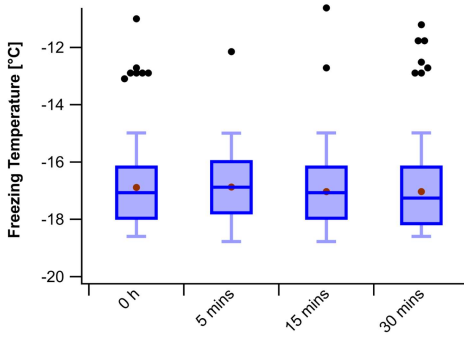
Detecting Outliers

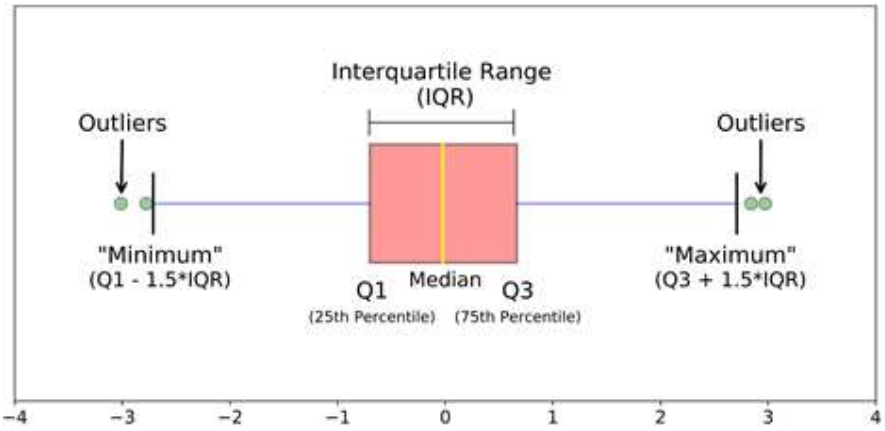
If our dataset is small, we can detect the outlier by just looking at the dataset. But what if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques.

Below are some of the techniques of detecting outliers

- ➢ Boxplots
- ➢ Z-score
- ➢ Inter Quantile Range(IQR)



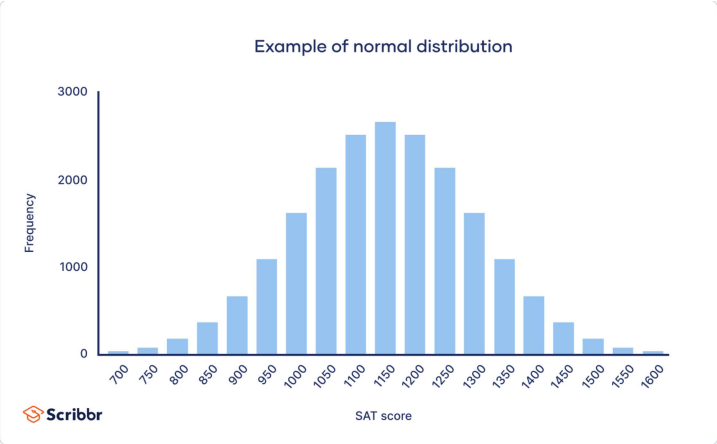




More about Outliers : [Outliers detection - Analytics Vidya]

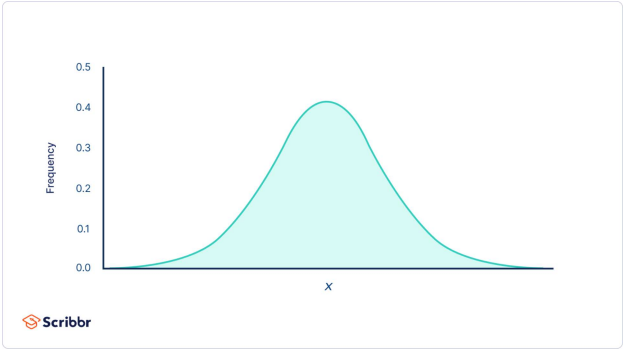
Normal or Gaussian distribution or Symmetrical distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

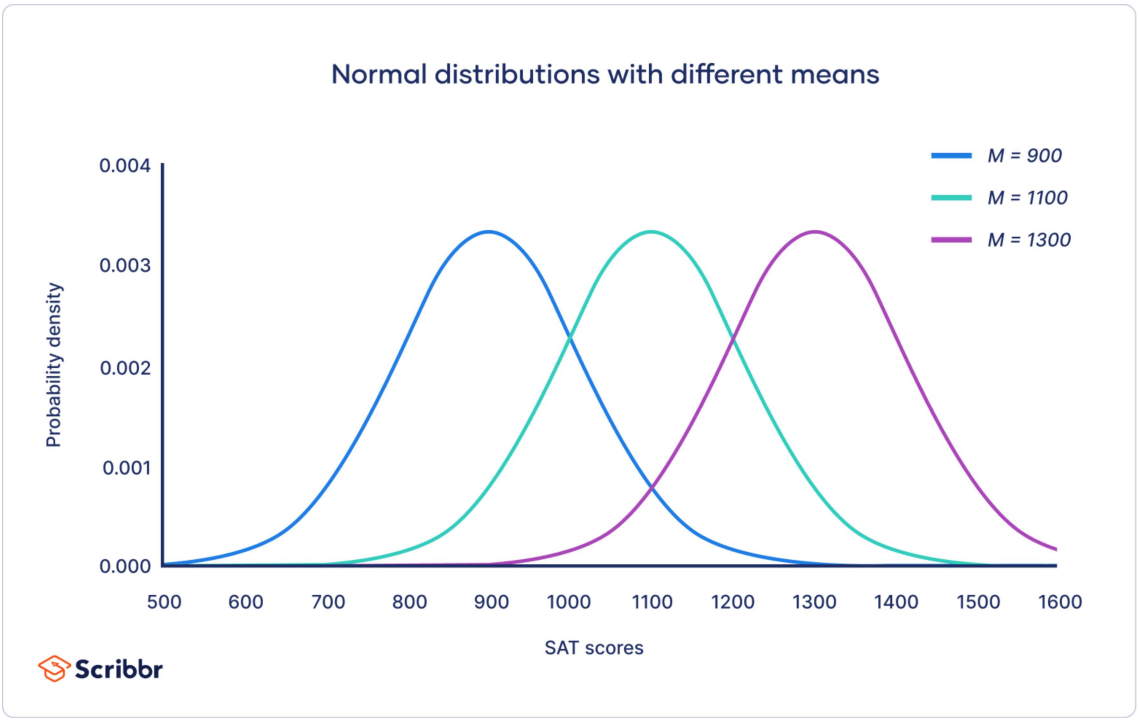


Properties of Normal distributions

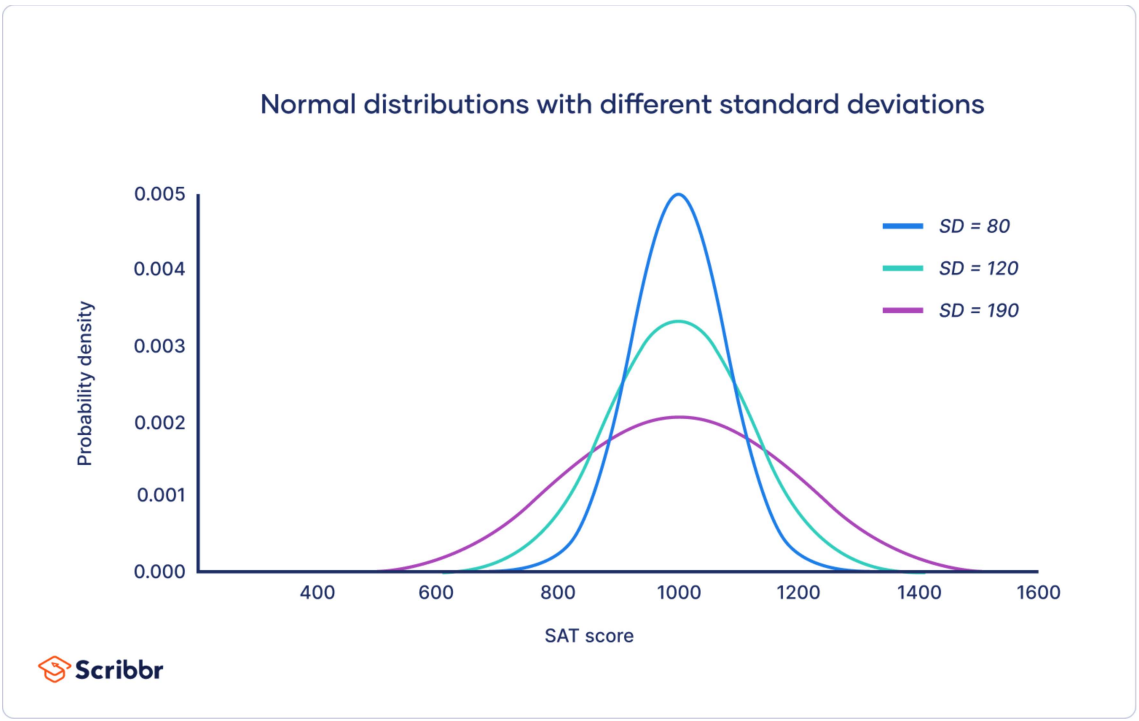
- In a normal distribution, data are symmetrically distributed with no skew. Most values cluster around a central region.
- The measures of central tendency (mean, mode, and median) are exactly the same in a normal distribution.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.



The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.



The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.



Empirical rule in Normal or Gaussian Distribution

The **empirical rule**, or the **68-95-99.7** rule, tells you where most of your values lie in a normal distribution:

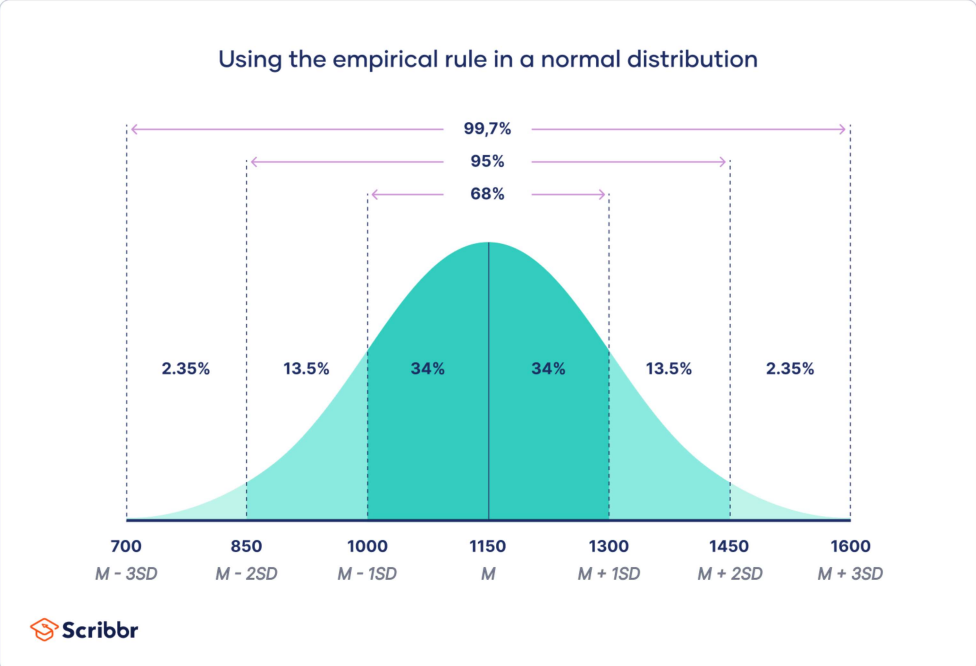
- Around **68%** of values are within 1 standard deviation from the mean.
- Around **95%** of values are within 2 standard deviations from the mean.
- Around **99.7%** of values are within 3 standard deviations from the mean.

Example: Using the empirical rule in a normal distribution -

You collect SAT(Scholastic Assessment Test) scores from students in a new test preparation course. The data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150.

Following the empirical rule:

- Around **68%** of scores are between **1,000 and 1,300**, 1 standard deviation above and below the mean.
- Around **95%** of scores are between **850 and 1,450**, 2 standard deviations above and below the mean.
- Around **99.7%** of scores are between **700 and 1,600**, 3 standard deviations above and below the mean.



The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don't follow this pattern.

Covariance

- ➢ Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

- ➢ In other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.

- ➢ The covariance value can range from **-∞ to +∞,** with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

Types of Covariance -

Covariance can have both positive and negative values. Based on this, it has two types:

- Positive Covariance
- Negative Covariance

Positive Covariance

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

Negative Covariance

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

Covariance Formula

It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by **Cov(X,Y).** The formula is given below for both population covariance and sample covariance.

Population Covariance Formula

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

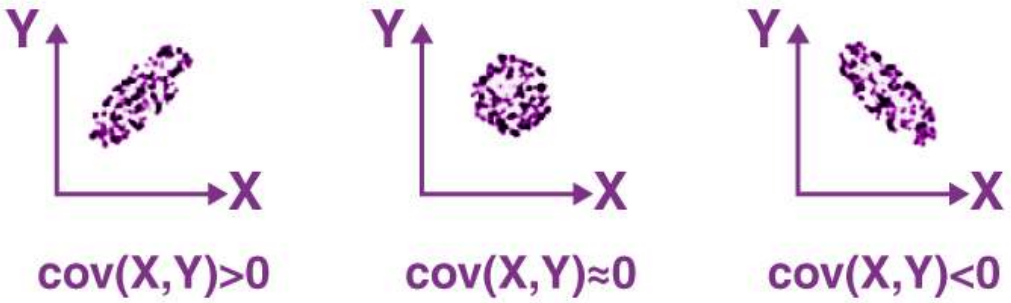
$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - y)}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

Covariance of X and Y

Below figure shows the covariance of X and Y.



➢ If cov(X, Y) is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

➢ If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

➢ If cov(X, Y) is zero, then we can say that there is no relation between two variables.

Pearson's Correlation Coefficient

- The correlation coefficient is a statistical measure that measures the strength and direction of the linear relationship between two variables. It is denoted by the **symbol "r"** and takes values between **-1 and 1.**

- When **r = 1,** it indicates a perfect positive linear relationship, which means that as one variable increases, the other variable also increases.

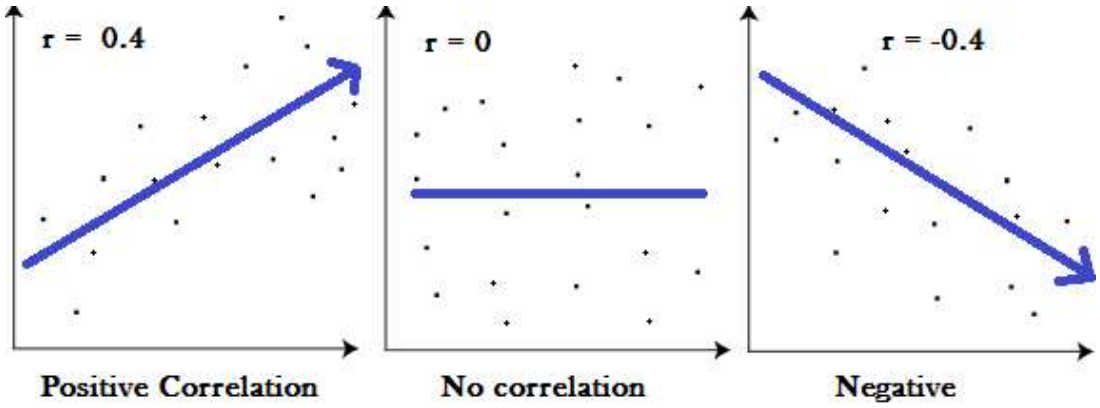
- When **r = -1,** it indicates a perfect negative linear relationship, which means that as one variable increases, the other variable decreases.

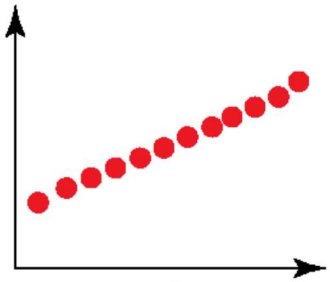
- When **r = 0,** it indicates no linear relationship between the two variables.

- A correlation coefficient between **-1 and 1** indicates the strength of the linear relationship between the two variables. The closer r is to -1 or 1, the stronger the relationship, while the closer r is to 0, the weaker the relationship.

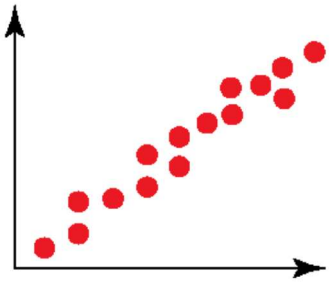
- One of the most commonly used formulas is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

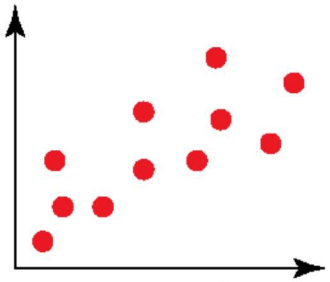




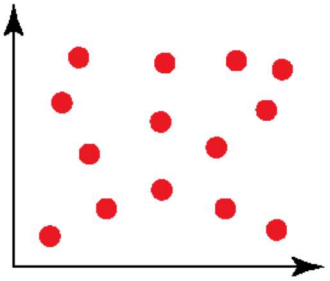
Perfect Positive Correlation



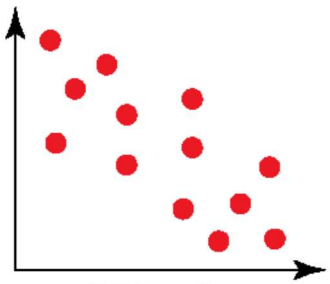
Strong Positive Correlation



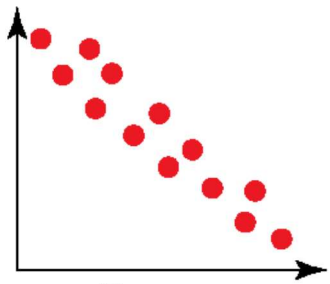
Weak Positive Correlation



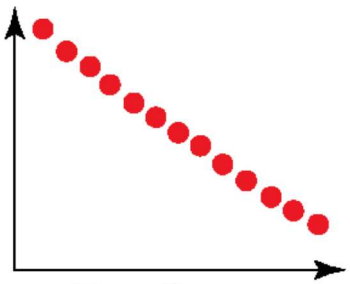
No Correlation



Weak Negative Correlation



Strong Negative Correlation



Perfect Negative Correlation