## Statistics: Definition, Importance, Limitation

Statistics is a form of mathematical analysis that uses quantified models, representations and synopses for a given set of experimental data or real-life studies. Statistics studies methodologies to gather, review, analyze and draw conclusions from data. Some statistical measures include mean, regression analysis, skewness, kurtosis, variance and analysis of variance.

Statistics is a term used to summarize a process that an analyst uses to characterize a data set. If the data set depends on a sample of a larger population, then the analyst can develop interpretations about the population primarily based on the statistical outcomes from the sample. Statistical analysis involves the process of gathering and evaluating data and then summarizing the data into a mathematical form.

Statistical methods analyze large volumes of data and their properties. Statistics is used in various disciplines such as psychology, business, physical and social sciences, humanities, government and manufacturing. Statistical data is gathered using a sample procedure or other methods. Two types of statistical methods are used in analyzing data: descriptive statistics and inferential statistics. Descriptive statistics are used to synopsize data from a sample exercising the mean or standard deviation. Inferential statistics are used when data is viewed as a subclass of a specific population.

## Importance and Scope of Statistics

### (i) Statistics in Planning

Statistics is indispensable in planning—may it be at business, economics or government level. The modern age is termed as the 'age of planning' and almost all organizations in the government or business or management are resorting to planning for efficient working and for formulating policy decisions.

To achieve this end, the statistical data relating to production, consumption, birth, death, investment, and income are of paramount importance. Today efficient planning is a must for almost all countries, particularly developing economies for their economic development.

### (ii) Statistics in Mathematics

Statistics is intimately related to and essentially dependent upon mathematics. The modern theory of Statistics has its foundations in the theory of probability which in turn is a particular branch of more advanced mathematical theory of Measures and Integration. Ever increasing role of mathematics in statistics has led to the development of a new branch of statistics called Mathematical Statistics.

Thus, Statistics may be considered to be an important member of the mathematics family. In the words of Connor, "Statistics is a branch of applied mathematics which specializes in data."

### (iii) Statistics in Economics

Statistics and Economics are so intermixed with each other that it looks foolish to separate them. DeveThe development modern statistical methods has led to an extensive use of statistics in Economics.

All the important branches of Economics—consumption, production, exchange, distribution, public finance—use statistics for the purpose of comparison, presentation, interpretation, etc. Problem of spending of income on and by different sections of the people, production of national wealth, adjustment of demand and supply, effect of economic policies on the economy etc. simply indicate the importance of statistics in the field of economics and in its different branches.

Statistics of Public Finance enables us to impose tax, to provide subsidy, to spend on various heads, amount of money to be borrowed or lent etc. So we cannot think of Statistics without Economics or Economics without Statistics.

### (iv) Statistics in Social Sciences

Every social phenomenon is affected to a marked extent by a multiplicity of factors which bring out the variation in observations from time to time, place to place and object to object. Statistical tools of Regression and Correlation Analysis can be used to study and isolate the effect of each of these factors on the given observation.

Sampling Techniques and Estimation Theory are very powerful and indispensable tools for conducting any social survey, pertaining to any strata of society and then analyzing the results and drawing valid inferences. The most important application of statistics in sociology is in the field of Demography for studying mortality (death rates), fertility (birth rates), marriages, population growth and so on.

**(v) Statistics in Trade**

As already mentioned, statistics is a body of methods to make wise decisions in the face of uncertainties. Business is full of uncertainties and risks. We have to forecast at every step. Speculation is just gaining or losing by way of forecasting. Can we forecast without taking into view the past? Perhaps, no. The future trend of the market can only be expected if we make use of statistics. Failure in anticipation will mean failure of business.

Changes in demand, supply, habits, fashion etc. can be anticipated with the help of statistics. Statistics is of utmost significance in determining prices of the various products, determining the phases of boom and depression etc. Use of statistics helps in smooth running of the business, in reducing the uncertainties and thus contributes towards the success of business.

**(vi) Statistics in Research Work**

The job of a research worker is to present the result of his research before the community. The effect of a variable on a particular problem, under differing conditions, can be known by the research worker only if he makes use of statistical methods. Statistics are everywhere basic to research activities. To keep alive his research interests and research activities, the researcher is required to lean upon his knowledge and skills in statistical methods.

**Limitations of Statistics**

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. Statisticians improve the quality of data with the design of experiments and survey sampling.

(i) Statistics does not deal with isolated measurement

(ii) Statistics deals with only quantitative characteristics

(iii) Statistics laws are true on average. Statistics are aggregates of facts. So single observation is not a statistics, it deals with groups and aggregates only.

(iv) Statistical methods are best applicable on quantitative data.

(v) Statistical cannot be applied to heterogeneous data.

(vi) It sufficient care is not exercised in collecting, analyzing and interpreting the data, statistical results might be misleading.

(vii) Only a person who has expert knowledge of statistics can handle statistical data efficiently.

(viii) Some errors are possible in statistical decisions. Particularly inferential statistics involves certain errors. We do not know whether an error has been committed or not.

**Measures of Central Tendency: Mean, Median, and Mode**

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. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

**MEAN (ARITHMETIC)**

The mean (or average) is the most popular and well-known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values $X_1$, $X_2$, $X_3$… $X_n$ , the sample mean, usually denoted by (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capital letter, , pronounced "sigma", which means "sum of...".

$$\bar{x} = \frac{\sum x}{n}$$

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lowercase letter “mu", denoted as µ:

$$\mu = \frac{\sum x}{n}$$

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

**MEDIAN**

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 | 92 |

We first need to rearrange that data into order of magnitude (smallest first):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 35 | 45 | 55 | 55 | **56** | 56 | 65 | 87 | 89 | 92 |

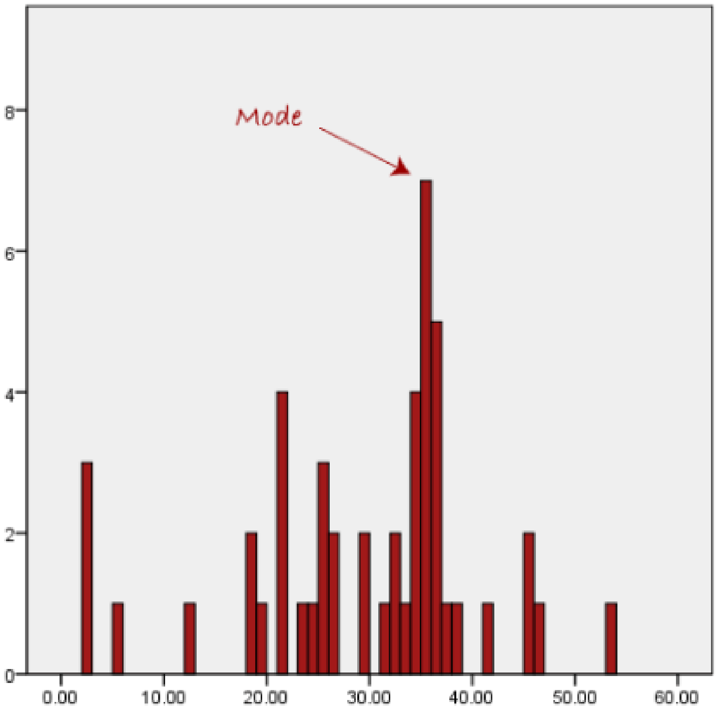
Our median mark is the middle mark — in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 35 | 45 | 55 | **55** | **56** | 56 | 65 | 87 | 89 |

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

**MODE**

The mode is the most frequent score in our data set. On a histogram, it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:



**Partition Values: Quartile, Deciles, Percentiles**

Partition values or fractiles such a quartile, a decile, etc. are the different sides of the same story. In other words, these are values that divide the same set of observations in different ways. So, we can fragment these observations into several equal parts.

**QUARTILE**

Whenever we have an observation and we wish to divide it, there is a chance to do it in different ways. So, we use the median when a given observation is divided into two parts that are equal. Likewise, quartiles are values that divide a complete given set of observations into four equal parts.

Basically, there are three types of quartiles, first quartile, second quartile, and third quartile. The other name for the first quartile is lower quartile. The representation of the first quartile is 'QI.' The other name for the second quartile is median.

The representation of the second quartile is by 'Q2 .c The other name for the third quartile is the upper quartile. The representation of the third quartile is by 'Q3.'

First Quartile is generally the one-fourth of any sort of observation. However, the point to note here is, this one-fourth value is always less than or equal to 'QI.' Similarly, it goes for the values of 'Q2' and 'Q3.'

**DECILES**

Deciles are those values that divide any set of a given observation into a total of ten equal parts. Therefore, there are a total of nine deciles. These representations of these deciles are as follows — D1, D2, D3, D4, D9.

D1 is the typical peak value for which one-tenth (1/10) of any given observation is either less or equal to DI. However, the remaining nine-tenths(9/10) of the same observation is either greater than or equal to the value of DI.

**PERCENTILES**

Last but not the least, comes the percentiles. The other name for percentiles is centiles. A centile or a percentile basically divides any given observation into a total of 100 equal parts. The representation of these percentiles or centiles is given as —Pl, P2, P3, P4, P99.

P1 is the typical peak value for which one-hundredth (1/100) of any given observation is either less or equal to Pl. However, the remaining ninety-nine-hundredth (99/100) of the same observation is either greater than or equal to the value of Pl.

This takes place once all the given observations are arranged in a specific manner i.e., ascending order. So, in case the data we have doesn't have a proper classification, then the representation of the $p^{th}$ quartile is $(n + 1 )p^{th}$

Here,

n = total number of observations.

p = 1/4, 2/4, 3/4 for different values of QI, Q2, and Q3 respectively.

p = 1/10, 2/10, .... 9/10 for different values of DI, D2, D9 respectively.

p = 1/100, 2/100, .... 99/100 for different values of Pl, P2, P99 respectively.

Formula

At times, the grouping of frequency distribution takes place. For which, we use the following formula during the

computation:

$$Q = l1 + [(Np - Ni)/(Nu - Ni)] * C$$

Here,

l1 = lower class boundary of the specific class that contains the median.

Ni = less than the cumulative frequency in correspondence to 11 (Post Median Class)

Nu = less than the cumulative frequency in correspondence to 12 (Pre Median Class)

C = Length of the median class

The symbol 'p' has its usual value. The value of 'p' varies completely depending on the type of quartile. There are different ways to find values or quartiles. We use this way in a grouped frequency distribution. The best way to do it is by drawing an ogive for the present frequency distribution.

Hence, all that we need to do to find one specific quartile is, find the point and draw a horizontal axis through the same. This horizontal line must pass through Np. The next step is to draw a perpendicular. The perpendicular comes up from the same point of intersection of the ogive and the horizontal line. Hence, the value of the quartile comes from the value of 'x' of the given perpendicular line.

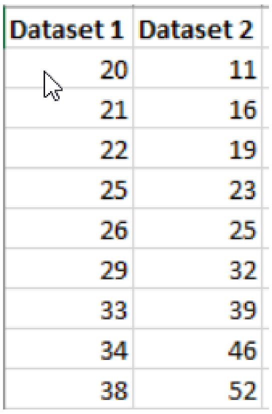
**Measures of Variation: Range, IQR**

A measure of variation is a summary statistic that represents the amount of dispersion in a dataset. While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the centre. We talk about variability in the context of a distribution of values. A low dispersion indicates that the data points tend to be clustered tightly around the centre. High dispersion signifies that they tend to fall further away.

In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution. Just as there are multiple measures of central tendency, there are several measures of variability.

**RANGE**

Let's start with the range because it is the most straightforward measure of variability to calculate and the simplest to understand. The range of a dataset is the difference between the largest and smallest values in that dataset. For example, in the two datasets below, dataset 1 has a range of 20 — 38 = 18 while dataset 2 has a range of 11 — 52 = 41. Dataset 2 has a wider range and, hence, more variability than dataset 1.



| Dataset 1 | Dataset 2 |
|---|---|
| 20 | 11 |
| 21 | 16 |
| 22 | 19 |
| 25 | 23 |
| 26 | 25 |
| 29 | 32 |
| 33 | 39 |
| 34 | 46 |
| 38 | 52 |

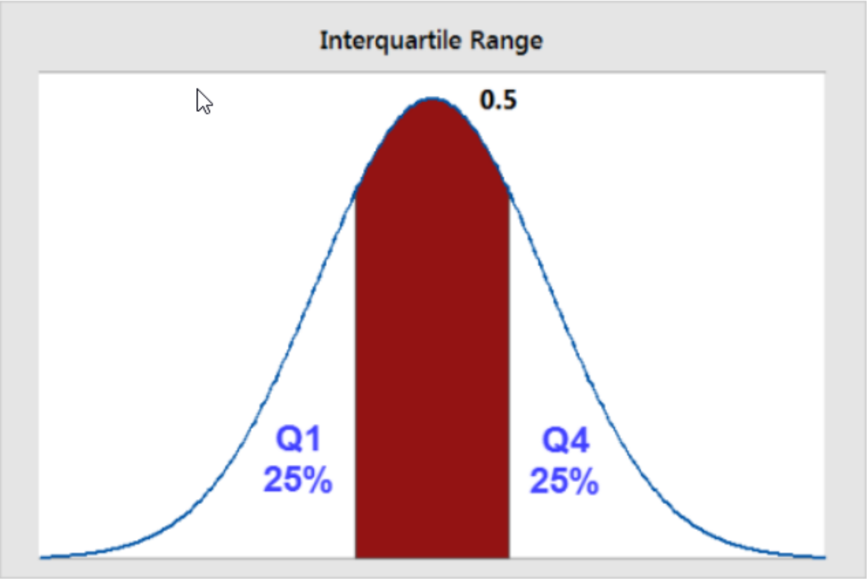
While the range is easy to understand, it is based on only the two most extreme values in the dataset, which makes it very susceptible to outliers. If one of those numbers is unusually high or low, it affects the entire range even if it is atypical.

Additionally, the size of the dataset affects the range. In general, you are less likely to observe extreme values. However, as you increase the sample size, you have more opportunities to obtain these extreme values. Consequently, when you draw random samples from the same population, the range tends to increase as the sample size increases. Consequently, use the range to compare variability only when the sample sizes are similar.

THE INTERQUARTILE RANGE (IQR)

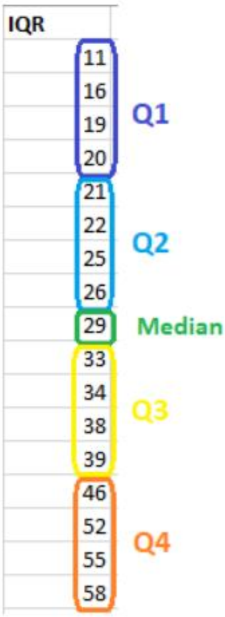
The interquartile range is the middle half of the data. To visualize it, think about the median value that splits the dataset in half. Similarly, you can divide the data into quarters. Statisticians refer to these quarters as quartiles and denote them from low to high as Q1, Q2, Q3, and Q4. The lowest quartile (Q1) contains the quarter of the dataset with the smallest values. The upper quartile (Q4) contains the quarter of the dataset with the highest values. The interquartile range is the middle half of the data that is in between the upper and lower quartiles. In other words, the interquartile range includes the 50% of data points that fall in Q2 and

The IQR is the red area in the graph below.



The interquartile range is a robust measure of variability in a similar manner that the median is a robust measure of centra tendency. Neither measure is influenced dramatically by outliers because they don't depend on every value. Additionally, the interquartile range is excellent for skewed distributions, just like the median. As you'll learn, when you have a normal distribution, the standard deviation tells you the percentage of observations that fall specific distances from the mean. However, this doesn't work for skewed distributions, and the IQR is a great alternative.

I've divided the dataset below into quartiles. The interquartile range (IQR) extends from the low end of Q2 to the upper limit of Q3. For this dataset, the range is 21-39.



| IQR | |
|---|---|
| 11 | Q1 |
| 16 | |
| 19 | |
| 20 | |
| 21 | Q2 |
| 22 | |
| 25 | |
| 26 | |
| 29 | Median |
| 33 | Q3 |
| 34 | |
| 38 | |
| 39 | |
| 46 | Q4 |
| 52 | |
| 55 | |
| 58 | |

# Mean Deviation and Standard Deviation

## Mean Deviation

To understand the dispersion of data from a measure of central tendency, we can use mean deviation. It comes as an improvement over the range. It basically measures the deviations from a value. This value is generally mean or median. Hence although mean deviation about mode can be calculated, mean deviation about mean and median are frequently used.

Note that the deviation of an observation from a value a is **d= x-a.** To find out mean deviation we need to take the mean of these deviations. However, when this value of a is taken as mean, the deviations are both negative and positive since it is the central value.

This further means that when we sum up these deviations to find out their average, the sum essentially vanishes. Thus to resolve this problem we use absolute values or the magnitude of deviation. The basic formula for finding out mean deviation is :

**Mean deviation= Sum of absolute values of deviations from 'a' ÷ The number of observations**

## Standard Deviation

As the name suggests, this quantity is a standard measure of the deviation of the entire data in any distribution. Usually represented by *s* or *σ*. It uses the arithmetic mean of the distribution as the reference point and normalizes the deviation of all the data values from this mean.

Therefore, we define the formula for the standard deviation of the distribution of a variable X with n data points as:

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

This formula can sometimes be useful in this alternate form as well –

$$s = \sqrt{\frac{\Sigma x_i^2}{n} - \bar{x}^2}$$

Also, sometimes if appropriate for a frequency distribution f(X), we can compute the standard deviation as –

$$s = \sqrt{\frac{\Sigma f_i(x_i - \bar{x})^2}{n}}$$

Alternatively –

$$s = \sqrt{\frac{\Sigma f_i x_i^2}{n} - \bar{x}^2}$$

# Variance, Coefficient of Variance

## Variance

Another statistical term that is related to the distribution is the variance, which is the standard deviation squared (variance = $SD^2$ ). The SD may be either positive or negative in value because it is calculated as a square root, which can be either positive or negative. By squaring the SD, the problem of signs is eliminated. One common application of the variance is its use in the F-test to compare the variance of two methods and determine whether there is a statistically significant difference in the imprecision between the methods.

In many applications, however, the SD is often preferred because it is expressed in the same concentration units as the data. Using the SD, it is possible to predict the range of control values that should be observed if the method remains stable. As discussed in an earlier lesson, laboratorians often use the SD to impose "gates" on the expected normal distribution of control values.

## Coefficient of Variation

Another way to describe the variation of a test is calculate the coefficient of variation, or CV. The CV expresses the variation as a percentage of the mean, and is calculated as follows:

**CV% = (SD/Xbar)100**

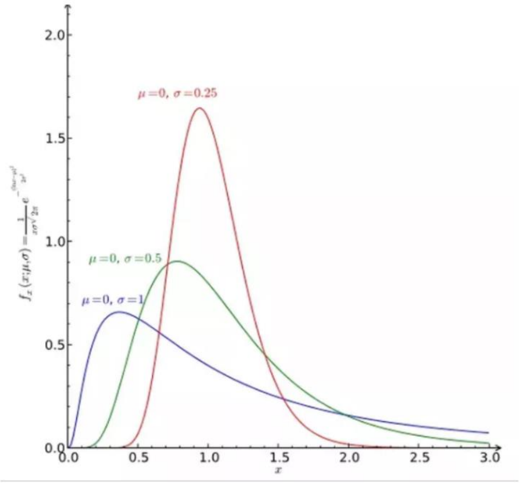
In the laboratory, the CV is preferred when the SD increases in proportion to concentration. For example, the data from a replication experiment may show an SD of 4 units at a concentration of 100 units and an SD of 8 units at a concentration of 200 units. The CVs are 4.0% at both levels and the CV is more useful than the SD for describing method performance at concentrations in between. However, not all tests will demonstrate imprecision that is constant in terms of CV. For some tests, the SD may be constant over the analytical range.

The CV also provides a general "feeling" about the performance of a method. CVs of 5% or less generally give us a feeling of good method performance, whereas CVs of 10% and higher sound bad. However, you should look carefully at the mean value before judging a CV. At very low concentrations, the CV may be high and at high concentrations the CV may be low. For example, a bilirubin test with an SD of 0.1 mg/dL at a mean value of 0.5 mg/dL has a CV of 20%, whereas an SD of 1.0 mg/dL at a concentration of 20 mg/dL corresponds to a CV of 5.0%.

# Skewness ant Types

Skewness, in statistics, is the degree of distortion from the symmetrical bell curve, or normal distribution, in a set of data. Skewness can be negative, positive, zero or undefined. A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew.

The three probability distributions depicted below depict increasing levels of right (or positive) skewness. Distributions can also be left (negative) skewed. Skewness is used along with kurtosis to better judge the likelihood of events falling in the tails of a probability distribution.



Right skewness

**Key Takeaways**

- Skewness, in statistics, is the degree of distortion from the symmetrical bell curve in a probability

distribution.

- Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degree.

- Investors note skewness when judging a return distribution because it, like kurtosis, considers the extremes of the data set rather than focusing solely on the average.

Broadly speaking, there are two types of skewness: They are

(1) Positive skewness and

(2) Negative skewness.

**Positive skewness**

A series is said to have positive skewness when the following characteristics are noticed:

- Mean > Median > Mode.

- The right tail of the curve is longer than its left tail when the data are plotted through a histogram or a frequency polygon.

- The formula of Skewness and its coefficient give positive figures.

**Negative skewness**

A series is said to have negative skewness when the following characteristics are noticed:

- Mode> Median > Mode.

- The left tail of the curve is longer than the right tail, when the data are plotted through a histogram, or a frequency polygon.

- The formula of skewness and its coefficient give negative figures.

**Thus, a statistical distribution may be three types viz.**

Symmetric

Positively skewed

Negatively skewed

**Kurtosis**

Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values. Along with skewness, kurtosis is an important descriptive statistic of data distribution. However, the two concepts must not be confused with each other. Skewness essentially measures the symmetry of the distribution while kurtosis determines the heaviness of the distribution tails.

In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk of an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

**Excess Kurtosis**

An excess kurtosis is a metric that compares the kurtosis of a distribution against the kurtosis of a normal distribution. The kurtosis of a normal distribution equals 3. Therefore, the excess kurtosis is found using the formula below:

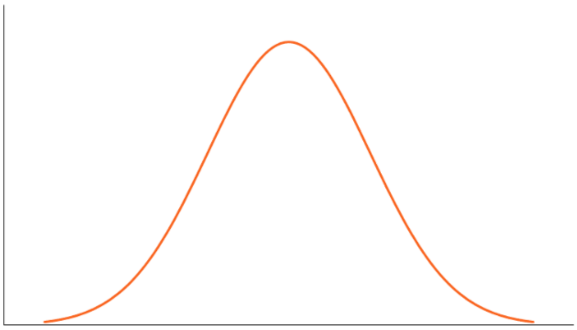
Excess Kurtosis = Kurtosis — 3

**Types of Kurtosis**

The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values as well, as values close to zero.

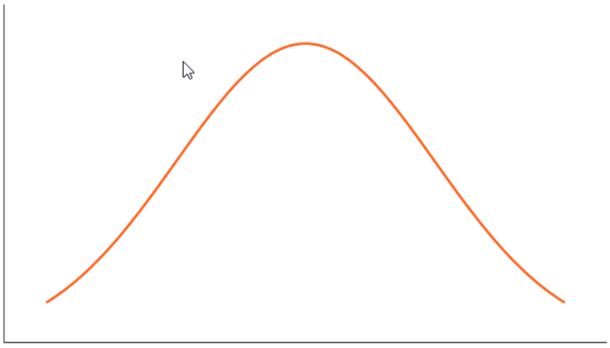
**1. Mesokurtic**

Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. It means that if the data follows a normal distribution, it follows a mesokurtic distribution.



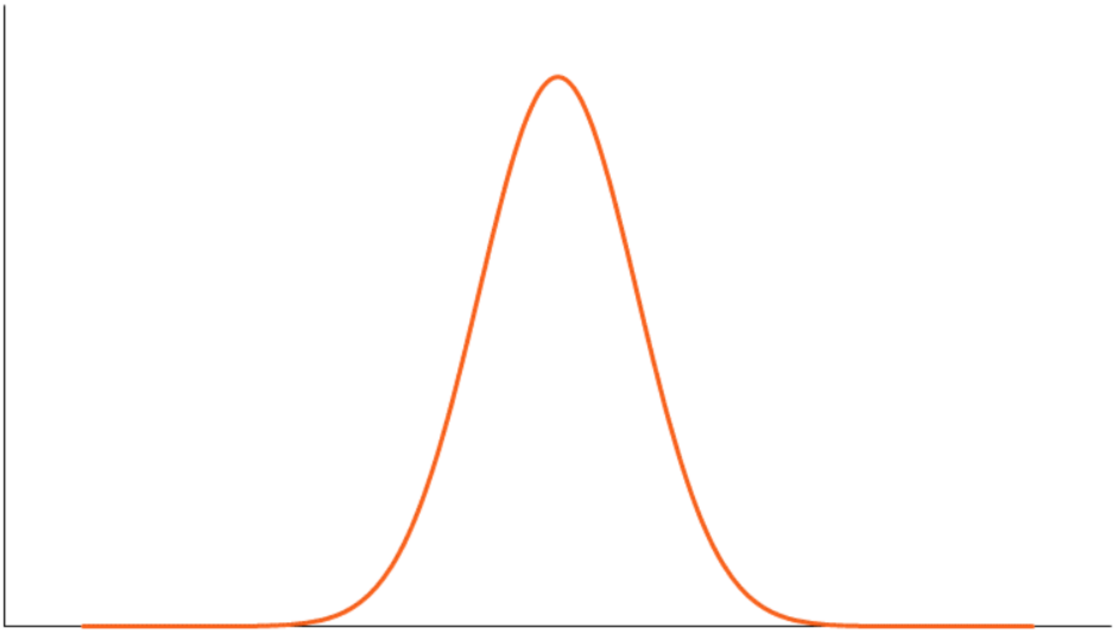
**2. Leptokurtic**

Leptokurtic indicates a positive excess kurtosis distribution. The leptokurtic distribution shows heavy tails on either side, indicating the large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.



### 3. Platykurtic

A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.



# Unit 2: Correlation & Regression Analysis

### Correlation analysis

**Correlation analysis** is a statistical method used to measure the relationship between two or more variables. The goal of correlation analysis is to determine whether there is a relationship between two variables and if so, to what extent.

Correlation analysis is typically used when we have two or more *continuous variables*, such as height and weight, or temperature and humidity. The strength and direction of the relationship between two variables can be measured by calculating the *correlation coefficient*.

The most commonly used correlation coefficient is Pearson's correlation coefficient, which ranges from -1 to +1. A coefficient of +1 indicates a perfect positive correlation, while a coefficient of -1 indicates a perfect negative correlation. A coefficient of 0 indicates no correlation between the two variables.

Correlation analysis can be used in a variety of fields, such as psychology, economics, and biology, to investigate the relationship between variables and to help make predictions or inform decision-making. However, *correlation does not necessarily imply causation,* and other factors may be responsible for the observed correlation. Therefore, it is important to use caution when interpreting the results of correlation analysis.

### The properties of correlation are as follows:

**Correlation coefficient ranges from -1 to +1:** The correlation coefficient is a standardized measure that ranges from -1 to +1. A correlation of -1 indicates a perfect negative correlation, while a correlation of +1 indicates a perfect positive correlation. A correlation coefficient of 0 indicates no correlation between the two variables.

**Correlation is symmetric:** The correlation between two variables is always symmetric, which means that the correlation between variable A and variable B is the same as the correlation between variable B and variable A.

**Correlation is affected by outliers:** Correlation can be affected by outliers, which are data points that are far away from the rest of the data. Outliers can artificially inflate or deflate the correlation coefficient and can lead to erroneous conclusions.

**Correlation does not imply causation:** Correlation does not imply causation, which means that just because two variables are correlated, it does not necessarily mean that one causes the other. Other factors may be responsible for the observed correlation.

**Correlation is affected by scale**: Correlation is affected by the scale of measurement of the variables. For example, if one variable is measured in inches and the other variable is measured in centimeters, the correlation coefficient will be affected by the difference in scale.

**Correlation can be influenced by sample size:** The correlation coefficient can be influenced by sample size, with larger sample sizes leading to more accurate estimates of the true correlation between variables.

**Correlation is not affected by the units of measurement:** Correlation is a unitless measure, which means that it is not affected by the units of measurement used for the variables. For example, the correlation between height and weight will be the same whether height is measured in inches or centimeters.

### Karl Pearson's coefficient of correlation:

Karl Pearson's coefficient of correlation, also known as Pearson's correlation coefficient or simply Pearson's r, is a measure of the linear relationship between two continuous variables. Pearson's r ranges from -1 to +1, with a value of +1 indicating a perfect positive correlation, a value of 0 indicating no correlation, and a value of -1 indicating a perfect negative correlation.

Pearson's correlation coefficient is calculated as the covariance between the two variables divided by the product of their standard deviations. The formula for Pearson's r is:

$$r = ( \Sigma (xi - x) (yi - y) ) / ( (n - 1) * s_x * s_y )$$

**where:**

Σ represents the sum of the values

xi is the value of variable x for observation i

x is the mean of variable x

yi is the value of variable y for observation i

y is the mean of variable y

n is the number of observations

s_x is the standard deviation of variable x

s_y is the standard deviation of variable y

Pearson's correlation coefficient is widely used in many fields of research, including psychology, biology, economics, and engineering, to measure the strength and direction of the relationship between two continuous variables. However, Pearson's r only measures linear relationships and may not capture non-linear relationships or relationships between variables that are not normally distributed.

**There are several assumptions associated with Pearson's correlation coefficient:**

**Linearity:** The relationship between the two variables being correlated should be linear, which means that the relationship should be roughly equal across the range of both variables.

**Normality:** Both variables should be normally distributed, or at least approximately normally distributed. This is important because Pearson's correlation coefficient assumes that the variables are normally distributed.

**Homoscedasticity:** Homoscedasticity refers to the assumption that the variance of the two variables should be roughly equal across the range of both variables.

**Independence:** The observations of the two variables being correlated should be independent of each other. In other words, the value of one variable should not depend on the value of the other variable.

**Outliers:** Pearson's correlation coefficient is sensitive to outliers, which can lead to inaccurate estimates of the correlation coefficient. Therefore, it is important to check for outliers before using Pearson's correlation coefficient.

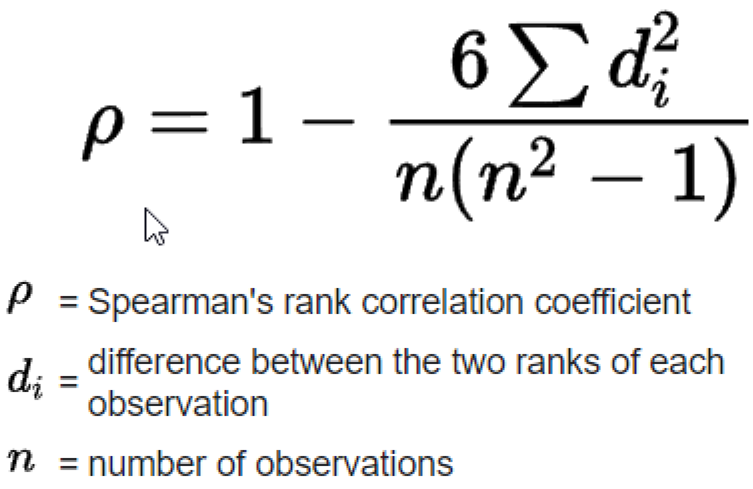
**Range of scores:** Pearson's correlation coefficient is also sensitive to the range of scores of the two variables being correlated. If the range of scores is too narrow, it may be difficult to detect a relationship between the two variables.

Violations of these assumptions can affect the accuracy and interpretation of Pearson's correlation coefficient. Therefore, it is important to check for these assumptions before using Pearson's correlation coefficient and to consider alternative methods of analysis if any of these assumptions are violated.

**Spearman's rank correlation coefficient**

**Spearman's rank correlation coefficient** is a non-parametric measure of the monotonic relationship between two variables. It is often used as an alternative to Pearson's correlation coefficient when the assumption of normality is not met, or when the relationship between the variables is not linear.

*Spearman's rank correlation coefficient is based on the ranks of the data, rather than the actual values.* The formula for Spearman's rank correlation coefficient is:



$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$\rho$ = Spearman's rank correlation coefficient

$d_i$ = difference between the two ranks of each observation

$n$ = number of observations

Spearman's rank correlation coefficient ranges from -1 to +1, with a value of +1 indicating a perfect positive monotonic relationship, a value of 0 indicating no monotonic relationship, and a value of -1 indicating a perfect negative monotonic relationship.

The assumptions for using Spearman's rank correlation coefficient are less stringent than those for Pearson's correlation coefficient. Spearman's rank correlation coefficient does not assume that the data are normally distributed, and it is less sensitive to outliers. However, Spearman's rank correlation coefficient assumes that the relationship between the variables is monotonic, rather than linear.

Spearman's rank correlation coefficient is commonly used in fields such as psychology, social sciences, and biology to measure the relationship between two variables when the assumptions for using Pearson's correlation coefficient are not met.

**Assumptions of Spearman's Rank Correlation:**

Spearman's rank correlation coefficient is a *non-parametric measure* of correlation and therefore, does not make any assumptions about the distribution of the data. However, there are some assumptions related to the use of Spearman's rank correlation coefficient, which are:

**Independence:** The observations of the two variables being correlated should be independent of each other. In other words, the value of one variable should not depend on the value of the other variable.

**Monotonicity:** The relationship between the two variables should be monotonic, which means that the relationship should be either strictly increasing or decreasing. The strength of the monotonic relationship is measured by the Spearman's rank correlation coefficient.

**Rankability:** The two variables should be capable of being ranked. This means that there should be no ties in the data or that the ties should be resolved using appropriate methods.

**Outliers:** Although Spearman's rank correlation coefficient is less sensitive to outliers compared to Pearson's correlation coefficient, outliers can still influence the results. Therefore, it is important to check for outliers before using Spearman's rank correlation coefficient.

It is important to note that Spearman's rank correlation coefficient does not measure causality, and a significant correlation does not necessarily mean that there is a causal relationship between the two variables. Therefore, it is important to use caution when interpreting the results of Spearman's rank correlation coefficient.

## Regression analysis

Regression analysis is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to develop a model that can accurately predict the values of the dependent variable based on the values of the independent variables.

**There are two main types of regression analysis: simple linear regression and multiple linear regression:**

Simple linear regression involves one independent variable and one dependent variable. The goal is to determine the linear relationship between the two variables and develop a linear equation that can be used to predict the value of the dependent variable based on the value of the independent variable.

Multiple linear regression involves two or more independent variables and one dependent variable. The goal is to determine the linear relationship between the dependent variable and all the independent variables together, and develop a linear equation that can be used to predict the value of the dependent variable based on the values of the independent variables.

Regression analysis uses statistical measures such as the coefficient of determination (R-squared), which indicates the proportion of variation in the dependent variable that can be explained by the independent variables, and the coefficients of the independent variables, which indicate the magnitude and direction of the effect of each independent variable on the dependent variable.

Regression analysis can be used in many different fields, including economics, finance, social sciences, engineering, and biology, to name a few. It is a powerful tool for making predictions and understanding the relationships between variables.

Fitting a regression line involves finding the equation of a line that best fits the data points in a scatter plot. The regression line represents the linear relationship between the independent variable and the dependent variable. The line can then be used to make predictions about the value of the dependent variable for a given value of the independent variable.

The equation of a regression line can be written as:

$Y = a + bX$

Where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope of the line. The slope of the line (b) represents the change in Y for every unit change in X, while the intercept (a) represents the value of Y when X is equal to zero.

To fit a regression line, various methods can be used, including the method of least squares. The method of least squares involves minimizing the sum of the squared differences between the actual values of the dependent variable and the predicted values of the dependent variable from the regression line.

Interpreting the results of a regression analysis involves several steps. First, the coefficient of determination (R-squared) should be calculated, which is a measure of how well the regression line fits the data. R-squared ranges from 0 to 1, with values closer to 1 indicating a better fit.

Second, the coefficients of the regression line should be examined. The slope coefficient (b) indicates the direction and strength of the relationship between the independent variable and the dependent variable. A positive slope coefficient indicates a positive relationship, while a negative slope coefficient indicates a negative relationship. The intercept coefficient (a) represents the value of the dependent variable when the independent variable is equal to zero.

Third, the significance of the coefficients should be evaluated using hypothesis testing. The null hypothesis is that the coefficient is equal to zero, indicating no relationship between the independent variable and the dependent variable. If the p-value associated with the coefficient is less than the level of significance (usually 0.05), the null hypothesis can be rejected, and it can be concluded that the coefficient is significantly different from zero.

Finally, the regression line can be used to make predictions about the value of the dependent variable for a given value of the independent variable. The predicted value of the dependent variable can be calculated by plugging the value of the independent variable into the regression equation.

**Simple linear regression:**

Simple linear regression is a type of regression analysis that involves only one independent variable and one dependent variable. The goal of simple linear regression is to establish a linear relationship between the two variables and develop a linear equation that can be used to predict the value of the dependent variable based on the value of the independent variable.

The equation of a simple linear regression line can be written as:

**Y = a + bX**

Where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope of the line. The slope of the line (b) represents the change in Y for every unit change in X, while the intercept (a) represents the value of Y when X is equal to zero.

To fit a simple linear regression line, the method of **least squares** is commonly used. The method of least squares involves minimizing the sum of the squared differences between the actual values of the dependent variable and the predicted values of the dependent variable from the regression line.

The coefficient of determination (R-squared) is a measure of how well the regression line fits the data. R-squared ranges from 0 to 1, with values closer to 1 indicating a better fit. R-squared represents the proportion of the total variation in the dependent variable that is explained by the independent variable.

The slope coefficient (b) indicates the direction and strength of the relationship between the independent variable and the dependent variable. A positive slope coefficient indicates a positive relationship, while a negative slope coefficient indicates a negative relationship. The intercept coefficient (a) represents the value of the dependent variable when the independent variable is equal to zero.

The significance of the coefficients can be evaluated using hypothesis testing. The null hypothesis is that the coefficient is equal to zero, indicating no relationship between the independent variable and the dependent variable. If the p-value associated with the coefficient is less than the level of significance (usually 0.05), the null hypothesis can be rejected, and it can be concluded that the coefficient is significantly different from zero.

The regression line can be used to make predictions about the value of the dependent variable for a given value of the independent variable. The predicted value of the dependent variable can be calculated by plugging the value of the independent variable into the regression equation. However, it is important to note that predictions made outside the range of the independent variable used to fit the regression line may not be accurate.

**Multiple linear regression**

Multiple linear regression is a type of regression analysis that involves more than one independent variable and one dependent variable. The goal of multiple linear regression is to establish a linear relationship between the dependent variable and multiple independent variables and develop a linear equation that can be used to predict the value of the dependent variable based on the values of the independent variables.

The equation of a multiple linear regression line can be written as:

$$\mathbf{Y = a + b1X1 + b2X2 + ... + bnxn}$$

Where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, a is the intercept, and b1, b2, ..., bn are the coefficients of the independent variables. The coefficients represent the change in Y for every unit change in the respective independent variable, holding all other independent variables constant.

To fit a multiple linear regression line, the method of least squares is commonly used. The method of least squares involves minimizing the sum of the squared differences between the actual values of the dependent variable and the predicted values of the dependent variable from the regression line.

The coefficient of determination (R-squared) is a measure of how well the regression line fits the data. R-squared ranges from 0 to 1, with values closer to 1 indicating a better fit. R-squared represents the proportion of the total variation in the dependent variable that is explained by the independent variables.

The significance of the coefficients can be evaluated using hypothesis testing. The null hypothesis is that the coefficient is equal to zero, indicating no relationship between the independent variable and the dependent variable, holding all other independent variables constant. If the p-value associated with the coefficient is less than the level of significance (usually 0.05), the null hypothesis can be rejected, and it can be concluded that the coefficient is significantly different from zero.

***Note:** The regression line can be used to make predictions about the value of the dependent variable for a given set of values of the independent variables. The predicted value of the dependent variable can be calculated by plugging the values of the independent variables into the regression equation. However, it is important to note that predictions made outside the range of the independent variables used to fit the regression line may not be accurate. Additionally, it is important to ensure that the independent variables are not highly correlated with each other, as this can lead to issues with multicollinearity.*

### Properties of Regression Coefficients

Regression coefficients have several properties that are important to understand when interpreting the results of a regression analysis. Some of the key properties are:

**Sign:** The sign of the coefficient indicates the direction of the relationship between the independent variable and the dependent variable. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.

**Magnitude:** The magnitude of the coefficient indicates the strength of the relationship between the independent variable and the dependent variable. A larger coefficient indicates a stronger relationship, while a smaller coefficient indicates a weaker relationship.

**Standard error:** The standard error of the coefficient measures the precision of the estimate. A smaller standard error indicates a more precise estimate, while a larger standard error indicates a less precise estimate.

**Confidence interval:** The confidence interval provides a range of values within which the true value of the coefficient is likely to fall with a certain level of confidence (usually 95%). A narrower confidence interval indicates a more precise estimate, while a wider confidence interval indicates a less precise estimate.

**T-statistic:** The t-statistic is a measure of the significance of the coefficient. A larger absolute value of the t-statistic indicates a more significant coefficient, while a smaller absolute value of the t-statistic indicates a less significant coefficient.

**P-value:** The p-value indicates the probability of observing a coefficient as extreme as the one calculated, assuming the null hypothesis that the true value of the coefficient is zero. A smaller p-value indicates a more significant coefficient, while a larger p-value indicates a less significant coefficient.

**Multicollinearity:** The presence of multicollinearity (high correlation between independent variables) can lead to unstable and unreliable coefficient estimates. In such cases, it may be necessary to use alternative methods such as regularization or dimensionality reduction.

It is important to consider these properties when interpreting regression coefficients, as they can provide valuable insights into the strength, precision, and significance of the relationship between the independent variable and the dependent variable.

### Relationship between Regression and Correlation:

- Regression analysis is used to predict the value of the dependent variable based on the values of one or more independent variables.
- Correlation analysis is used to determine the degree of association or relationship between two or more variables.
- The coefficient of correlation (r) is often used to assess the strength and direction of the linear relationship between the independent and dependent variables in a regression analysis.
- The coefficient of correlation (r) ranges from -1 to 1, with values closer to -1 indicating a strong negative linear relationship, values closer to 1 indicating a strong positive linear relationship, and values close to 0 indicating little or no linear relationship.
- Regression and correlation are complementary statistical techniques that can be used together to examine the relationship between variables.

**Difference between Regression and Correlation:**

Regression and correlation are two statistical techniques that are often used together to examine the relationship between variables. Here are the key differences between regression and correlation:

**Purpose:** The purpose of regression analysis is to develop a mathematical equation that can be used to predict the value of the dependent variable based on the values of one or more independent variables. In contrast, the purpose of correlation analysis is to determine the degree of association or relationship between two or more variables.

**Directionality:** Regression analysis involves the dependent variable and one or more independent variables, and it attempts to identify a functional relationship between them. Correlation analysis examines the association between two or more variables without specifying the direction of the relationship.

**Causality:** Regression analysis can be used to determine the cause-and-effect relationship between the dependent and independent variables. Correlation analysis, however, cannot determine causality; it only indicates whether a relationship exists between two variables.

**Values:** The output of regression analysis is a set of regression coefficients that represent the relationship between the dependent and independent variables. The output of correlation analysis is the correlation coefficient (r), which represents the strength and direction of the association between two variables.

**Linearity:** Regression analysis assumes that the relationship between the dependent and independent variables is linear. Correlation analysis does not make this assumption and can detect both linear and non-linear relationships.

Overall, regression analysis is used to predict the value of the dependent variable based on the values of one or more independent variables, while correlation analysis is used to determine the degree of association or relationship between two or more variables.

| Feature | Regression Analysis | Correlation Analysis |
|---|---|---|
| **Purpose** | Predict the value of dependent variable(s) | Determine the degree of association |
| **Directionality** | Dependent and independent variables | Two or more variables |
| **Causality** | Can determine cause-and-effect | Cannot determine causality |
| **Values** | Regression coefficients | Correlation coefficient (r) |
| **Linearity Assumption** | Assumes linear relationship | Can detect both linear and non-linear relationships |

**Time Series Analysis: Concept, Additive and Multiplicative models**

Time series analysis is a statistical method used to analyze and forecast data that is collected over time. It is a useful tool for understanding the behaviour of a variable over time and for making predictions about the future values of the variable.

In time series analysis, two commonly used models are additive and multiplicative models. These models are used to describe the relationship between the observed data and time.

**Component of Time Series**

The components of a time series are the different sources of variation or patterns that are present in the data over time. These components can help to explain the underlying structure of the time series and can be used to develop models for forecasting future values.

There are generally four main components of a time series:

**Trend:** The trend component represents the long-term behaviour or direction of the time series. It reflects the underlying pattern or tendency of the series to increase or decrease over time. The trend can be linear, non-linear, or even irregular, depending on the nature of the data.

**Seasonality:** The seasonality component refers to the regular and repeating patterns of variation that occur at fixed intervals of time within a year or over multiple years. These patterns can be daily, weekly, monthly, quarterly, or annually. Examples of seasonality include increased sales during holiday seasons, higher temperatures in summer, and lower temperatures in winter.

**Cyclical Variation:** The cyclical component of the time series refers to the patterns of variation that occur over a period of several years or even decades. These patterns are typically driven by economic, political, or social factors that affect the overall level of the series. Cyclical variation is often confused with seasonal variation but is fundamentally different since it does not occur at regular intervals of time.

**Irregular Variation**: The irregular component, also known as the residual component, refers to the random or unpredictable fluctuations that are left after the trend, seasonality, and cyclical components have been accounted for. These fluctuations can arise due to factors such as measurement error, sampling error, or other random events that affect the series.

By understanding and modelling the different components of a time series, analysts can develop effective forecasting models that take into account the underlying structure of the data.

**Additive and Multiplicative Models:**

An ***additive model*** assumes that the observed data is the sum of several components, each of which varies independently over time. These components include a trend component, a seasonal component, and a random or noise component. The trend component represents the long-term behaviour of the series, while the seasonal component captures the periodic fluctuations in the series. The random or noise component represents the random fluctuations in the data that are not accounted for by the trend or seasonal components. Mathematically, an additive model can be expressed as:

$Y = T + S + C + I$

where Y is the observed data at time t, T is the trend component at a time, S is the seasonal component at a time, and I is the random or noise component at time t.

A ***multiplicative model,*** on the other hand, assumes that the observed data is the product of several components, it also assumes that the effect of one component is proportional to the others. These components include a trend component, a seasonal component, and a random or noise component. Mathematically, a multiplicative model can be expressed as:

$Y = T * S * C * I$

where Y is the observed data at time t, T is the trend component at a time, S is the seasonal component at a time, and I is the random or noise component at time t.

The choice of whether to use an additive or multiplicative model depends on the nature of the time series data. If the variation in the data is proportional to the level of the data, then a multiplicative model is appropriate. If the variation in the data is independent of the level of the data, then an additive model is appropriate.

Both additive and multiplicative models can be used for forecasting future values of the time series. The choice of model depends on the specific characteristics of the time series data and the goals of the analysis.

**Trend Analysis**

Trend analysis is a method used to analyze and forecast trends in time series data. One common approach to trend analysis is the least squares method, which involves fitting a line (or curve) to the data that minimizes the sum of the squared differences between the predicted values and the observed values.

The **Least Squares method** is a statistical technique used to find the best-fit line or curve that represents the relationship between two variables. It is commonly used in linear regression analysis, but can also be used for non-linear regression analysis.

In simple **linear regression analysis**, we are interested in finding the line that best represents the relationship between a dependent variable (Y) and an independent variable (X). The line is determined by finding the values of the slope (b) and the intercept (a) that minimize the sum of the squared differences between the predicted values of Y and the actual values of Y.

The equation for the line is given by:

$\mathbf{Y = a + bX}$

where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope.

To find the values of a and b that minimize the sum of the squared differences, we can use the following formulas:

$b = (n\sum XY - \sum X\sum Y) / (n\sum X^2 - (\sum X)^2)$

$a = (\sum Y - b\sum X) / n$

where n is the number of observations, $\sum XY$ is the sum of the products of X and Y, $\sum X$ and $\sum Y$ are the sums of X and Y, respectively, and $\sum X^2$ is the sum of the squares of X.

Once we have determined the values of a and b, we can use the equation of the line to predict future values of Y.

In **non-linear regression analysis**, the least squares method can be used to find the best-fit curve that represents the relationship between the variables. The curve can take many different forms, depending on the specific pattern observed in the data. The values of the parameters in the equation of the curve are determined by minimizing the sum of the squared differences between the predicted values of Y and the actual values of Y.

Overall, the least squares method is a powerful tool for analyzing relationships between variables and making predictions based on those relationships. However, it is important to be aware of the assumptions underlying the method and to interpret the results carefully.

### Assumptions of Least Squares Method

The least squares model relies on several assumptions in order to produce accurate results. These assumptions include:

**Linearity:** The relationship between the dependent variable and the independent variable(s) should be linear. This means that the change in the dependent variable should be proportional to the change in the independent variable(s).

**Independence:** The observations should be independent of each other. This means that the value of one observation should not be influenced by the value of any other observation.

**Homoscedasticity**: The variance of the errors (i.e., the difference between the predicted values and the actual values) should be constant across all levels of the independent variable(s). This means that the spread of the errors should be the same for all values of the independent variable(s).

**Normality:** The errors should be normally distributed. This means that the distribution of the errors should be symmetrical around zero.

**Outliers:** The data should not contain any outliers that have a significant impact on the results. Outliers are observations that are significantly different from the other observations and can have a large influence on the estimated coefficients.

It is important to check these assumptions before using the least squares model to make predictions. Violations of these assumptions can lead to inaccurate results and biased predictions. If any of these assumptions are violated, it may be necessary to use a different modelling approach or to transform the data in order to meet the assumptions.

### Applications in business decision-making:

The least squares method is widely used in business decision-making because it can be applied to a variety of scenarios where there is a need to analyze the relationship between variables and make predictions based on that relationship. Some of the applications of the least squares method in business decision-making include:

**Forecasting:** The least squares method can be used to forecast future values of a variable based on historical data. This is useful in scenarios where there is a need to predict sales, demand, or other key performance indicators.

**Pricing:** The least squares method can be used to analyze the relationship between pricing and sales. This can help businesses to determine the optimal price point that maximizes revenue and profitability.

**Quality control:** The least squares method can be used to analyze the relationship between quality control measures and defects. This can help businesses to identify the key factors that contribute to defects and to develop strategies to improve quality control.

**Marketing:** The least squares method can be used to analyze the relationship between marketing activities (e.g., advertising, promotions) and sales. This can help businesses to optimize their marketing spend and to develop more effective marketing campaigns.

**Financial analysis:** The least squares method can be used to analyze the relationship between financial variables (e.g., revenue, expenses, profits) and to make predictions about future financial performance.

Overall, the least squares method is a powerful tool for business decision-making because it allows businesses to analyze the relationships between variables and to make predictions based on those relationships. This can help businesses to make more informed decisions and to optimize their operations for maximum efficiency and profitability.

### Index Number and meaning

Index numbers are a statistical measure used to track changes in the value of a variable over time or across different groups. They are typically used to measure changes in the price level, but can also be used to measure changes in other variables such as output, employment, and population.

An index number is a measure of the relative change in the value of a variable compared to a base period or base year. The base period is typically set to 100 and the index number for subsequent periods is expressed as a percentage of the base period.

For example, suppose the price of a basket of goods in the base year was $100 and in the current year it has risen to $120. The index number for the current year would be 120, indicating a 20% increase in prices compared to the base year.

Index numbers are useful because they allow us to compare changes in the value of a variable over time or across different groups, even if the absolute values of the variable are not comparable. For example, the price of a basket of goods in one country may be higher than the price of the same basket of goods in another country, but we can use index numbers to compare changes in prices over time within each country.

Index numbers can also be used to calculate inflation rates, which is the rate at which the general price level is increasing. Inflation is calculated by taking the percentage change in the index number for a particular period and subtracting the percentage change in the index number for the previous period.

Overall, index numbers are a valuable tool for measuring changes in the value of a variable over time or across different groups. They allow us to compare changes in the value of a variable, even if the absolute values are not comparable, and can provide useful insights into economic trends and patterns.

**The key features of index numbers include:**

**Relative measure:** Index numbers are a relative measure of change, meaning that they measure the change in the value of a variable relative to a base period or base year.

**Base period or base year:** Index numbers are calculated relative to a base period or base year. The base period or base year is usually set to 100, and the index number for subsequent periods is expressed as a percentage of the base period.

**Aggregation:** Index numbers can be aggregated to provide an overall measure of change. For example, the Consumer Price Index (CPI) aggregates price changes for a basket of goods and services to provide an overall measure of inflation.

**Weighting:** Index numbers can be weighted to reflect the importance of different components. For example, the CPI assigns weights to different categories of goods and services based on their relative importance in the average consumer's budget.

**Time series analysis:** Index numbers can be used to analyze changes in a variable over time. This can help to identify trends and patterns, and to make forecasts about future values of the variable.

Cross-sectional analysis: Index numbers can be used to compare changes in a variable across different groups. For example, the GDP deflator can be used to compare changes in the general price level across different countries.

Overall, index numbers are a useful tool for measuring changes in the value of a variable over time or across different groups. They provide a relative measure of change, can be aggregated and weighted to reflect the importance of different components, and can be used for time series analysis and cross-sectional analysis.

**There are various types of index numbers, each designed to measure changes in different variables. Here are some of the most common types:**

**Price Index Numbers:** These measure changes in the price of goods and services over time. Examples include the Consumer Price Index (CPI), Producer Price Index (PPI), and Wholesale Price Index (WPI).

**Quantity Index Numbers:** These measure changes in the quantity of goods and services produced over time. Examples include the Industrial Production Index (IPI) and the Retail Sales Index (RSI).

**Value Index Numbers:** These measure changes in the total value of goods and services produced over time. Examples include the Gross Domestic Product (GDP) deflator, which measures changes in the price level of all goods and services produced in a country.

**Cost of Living Index Numbers:** These measure changes in the cost of living over time. Examples include the Consumer Price Index for All Urban Consumers (CPI-U) and the Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI-W).

**Stock Market Index Numbers:** These measure changes in the value of stocks traded on a stock exchange. Examples include the Dow Jones Industrial Average (DJIA) and the Standard & Poor's 500 Index (S&P 500).

**Quality Index Numbers:** These measure changes in the quality of goods and services over time. Examples include the Hedonic Price Index (HPI) used to measure changes in the quality of consumer goods.

Overall, index numbers are a valuable tool for measuring changes in various variables over time. The choice of index number depends on the variable being measured and the purpose of the analysis.

**Index numbers have various uses in different fields. Some common uses of index numbers include:**

**Economic analysis:** Index numbers are used to measure changes in economic variables over time, such as inflation, production, and employment. They help to identify trends, patterns, and fluctuations in the economy.

**Business planning:** Index numbers are used in business planning to forecast future values of variables, such as sales, costs, and profits. They can help businesses to make informed decisions about production, pricing, and investments.

**Investment analysis:** Index numbers are used to evaluate the performance of investments, such as stocks, bonds, and mutual funds. They help to compare the returns on different investments over time and across different markets.

**Marketing research:** Index numbers are used in marketing research to measure consumer behavior and preferences. For example, the Consumer Confidence Index (CCI) is used to measure consumer sentiment and spending habits.

**Government policy-making:** Index numbers are used by governments to monitor and evaluate the effectiveness of policies, such as fiscal and monetary policies. They help to identify the impact of policies on the economy and society.

Overall, index numbers are a valuable tool for measuring changes in various variables over time. They provide a relative measure of change, can be aggregated and weighted to reflect the importance of different components, and can be used for time series analysis and cross-sectional analysis.

**Price, quantity, and volume indices are constructed differently depending on the variable being measured. Here are the basic steps for constructing each type of index:**

**Price Index:** Price indices measure changes in the price of goods and services over time. To construct a price index, follow these steps:

- Select a base year and a basket of goods and services that represent the typical consumption patterns of a population.
- Collect prices for each item in the basket in the base year.

- Repeat the process for subsequent years.
- Calculate the price index for each year by dividing the total cost of the basket in that year by the total cost of the basket in the base year, then multiplying by 100.

**Quantity Index:** Quantity indices measure changes in the quantity of goods and services produced over time. To construct a quantity index, follow these steps:

- Select a base year and a set of goods and services that represent the typical production patterns of an industry or economy.
- Collect data on the quantity of each item produced in the base year.
- Repeat the process for subsequent years.
- Calculate the quantity index for each year by dividing the total quantity of each item produced in that year by the total quantity of each item produced in the base year, then multiplying by 100.

**Volume Index:** Volume indices measure changes in the value of goods and services produced over time, adjusting for changes in price. To construct a volume index, follow these steps:

- Select a base year and a set of goods and services that represent the typical production patterns of an industry or economy.
- Collect data on the quantity and price of each item produced in the base year.
- Repeat the process for subsequent years.
- Calculate the value of each item produced in each year by multiplying the quantity by the price.
- Calculate the volume index for each year by dividing the total value of all items produced in that year by the total value of all items produced in the base year, then multiplying by 100.

Overall, constructing price, quantity, and volume indices involves collecting data on the relevant variable and comparing it over time relative to a base year or period. The resulting index provides a useful measure of change in the variable over time, adjusting for changes in price or quantity.

**Fixed base and chain base methods are two different approaches for constructing index numbers. Here's a brief explanation of each method:**

**Fixed Base Method:** In the fixed base method, the index is calculated relative to a fixed base year. The prices or quantities of the items in the basket are collected for the base year, and then the same basket of items is priced or quantified for each year under consideration. The index is calculated by dividing the cost or quantity of the basket in each year by the cost or quantity of the basket in the base year and multiplying by 100. For example, if the base year is 2010 and the cost of the basket of goods in 2015 is $500 and in 2020 is $600, the index for 2015 would be calculated as (500/1000) x 100 = 50, and the index for 2020 would be calculated as (600/1000) x 100 = 60.

**Chain Base Method:** In the chain base method, the index is calculated relative to the preceding year instead of a fixed base year. This method is used to link together different base years, creating a continuous series of index numbers that account for changes in the composition of the basket of goods and services over time. The prices or quantities of the items in the basket are collected for the first year under consideration, and the index is calculated relative to that year. In the second year, a new basket of goods and services is chosen, and the index is calculated relative to the previous year's basket. This process is repeated for each subsequent year, with each year's basket being linked to the previous year's basket. For example, if the base year is 2010 and the index for 2015 is 120 and the index for 2020 is 150, the index for 2015 is calculated as (120/100) x 100 = 120 and the index for 2020 is calculated as (150/120) x 100 = 125.

Overall, the choice of fixed base or chain base method depends on the nature of the data and the purpose of the index. The fixed base method is useful when comparing changes in prices or quantities over time relative to a fixed point, while the chain base method is useful for creating a continuous series of index numbers that reflect changes in the composition of the basket over time.

Probability theory is a branch of mathematics that deals with the study of random events and phenomena. It provides a framework for understanding and analysing uncertain situations, enabling us to make informed decisions in the face of uncertainty.

The basic concepts of probability theory include events, outcomes, sample space, probability, and random variables. An event is a set of outcomes of an experiment, while an outcome is a possible result of the experiment. The sample space is the set of all possible outcomes of the experiment, and the probability is a measure of the likelihood of an event occurring.

Random variables are used to model uncertain quantities in probability theory. They are variables whose values depend on the outcome of a random experiment. For example, the number of heads obtained when flipping a coin is a random variable.

The three main types of probability are classical probability, empirical probability, and subjective probability. Classical probability is used when all outcomes of an experiment are equally likely, and empirical probability is based on observations and experiments. Subjective probability is based on personal judgments and opinions.

Probability theory has many applications, including in statistics, finance, engineering, and computer science. It is used to model and analyze various real-world phenomena, such as the behavior of stock prices, the reliability of machines, and the spread of infectious diseases.

**The theory of probability is based on a set of fundamental concepts and principles, including:**

- **Sample space:** The set of all possible outcomes of a random experiment.
- **Event:** A subset of the sample space representing a particular outcome or set of outcomes.
- **Probability measure:** A function that assigns a numerical value between 0 and 1 to each event, representing the likelihood of that event occurring.
- **Probability distribution:** A function that describes the probabilities of all possible outcomes in a random experiment.
- **Independence:** Two events are independent if the occurrence of one event does not affect the likelihood of the other event occurring.
- **Conditional probability:** The probability of an event given that another event has occurred.
- **Bayes' theorem:** A formula for calculating the probability of an event based on prior knowledge or information.

The theory of probability has numerous applications in various fields, including statistics, finance, engineering, physics, and computer science. It is used to model and analyse a wide range of phenomena, such as the behaviour of stock prices, the likelihood of natural disasters, and the spread of infectious diseases.

**Hypothesis testing** is a statistical method that allows researchers to make conclusions about population parameters based on data collected from a sample. The purpose of hypothesis testing is to determine whether an assumption about a population parameter is likely to be true or false based on the sample data. In this section, we will discuss the different components of hypothesis testing in detail.

Hypothesis testing is a statistical method used to determine whether a particular hypothesis about a population parameter is likely to be true or false. It involves making an assumption about a population parameter, collecting data, and then using statistical tests to determine whether the data support or refute the hypothesis.

**Steps in Hypothesis Testing:**

**Null Hypothesis and Alternative Hypothesis**

The first step in hypothesis testing is to state the null hypothesis (H0) and the alternative hypothesis (Ha). The null hypothesis is the assumption that there is no significant difference between a population parameter and a specific value, whereas the alternative hypothesis is the assumption that there is a significant difference between a population parameter and a specific value.

For example, let's say we are interested in determining whether the mean height of a population of people is different from a specific value, such as 170 cm. The null hypothesis in this case would be that the mean height of the population is equal to 170 cm (H0: $\mu = 170$), and the alternative hypothesis would be that the mean height of the population is not equal to 170 cm (Ha: $\mu \neq 170$).

**Significance Level**

The next step is to set the significance level (alpha), which is the probability of rejecting the null hypothesis when it is true. The significance level is typically set at 0.05 or 0.01, which means that there is a 5% or 1% chance of rejecting the null hypothesis when it is true.

**Type I and Type II Errors**

Before we move on to the next steps, it's essential to understand the concept of Type I and Type II errors. Type I error occurs when we reject the null hypothesis when it is true, whereas Type II error occurs when we accept the null hypothesis when it is false. The probability of Type I error is denoted by alpha, whereas the probability of Type II error is denoted by beta.

**Test Statistics**

The next step in hypothesis testing is to calculate the test statistic, which is a measure of how far the sample estimate is from the hypothesized value. The test statistic depends on the type of hypothesis test being conducted. For example, if we are conducting a t-test, the test statistic would be calculated using the t-distribution.

**P-Value**

The p-value is the probability of obtaining a sample result as extreme or more extreme than the one observed if the null hypothesis is true. It measures the strength of evidence against the null hypothesis. The p-value is compared to the significance level (alpha) to determine whether to reject or fail to reject the null hypothesis.

If the p-value is less than alpha, we reject the null hypothesis and conclude that the alternative hypothesis is true. If the p-value is greater than alpha, we fail to reject the null hypothesis.

**Confidence Intervals**

Another way of testing hypotheses is to use confidence intervals. A confidence interval is a range of values that is likely to contain the true population parameter with a certain degree of confidence. The degree of confidence is typically set at 95% or 99%. If the hypothesized value falls outside the confidence interval, we reject the null hypothesis.

**Conclusion**

In conclusion, hypothesis testing is a statistical method used to test whether a hypothesis about a population parameter is likely to be true or false based on the sample data. The process involves stating the null hypothesis and the alternative hypothesis, setting the significance level, calculating the test statistic, determining the p-value, and drawing a conclusion. The choice of the hypothesis test and the method used to interpret the results depend on the nature of the data and the research question being investigated.

**Null Hypothesis**

In statistical hypothesis testing, the null hypothesis (H0) is a hypothesis that represents the status quo or the current state of knowledge. It is often a statement that there is no significant difference between a population parameter and a specific value or that there is no relationship between two variables.

For example, if we are interested in testing whether a new medication is effective in reducing blood pressure, the null hypothesis would be that the new medication is not effective, and there is no significant difference in blood pressure between the group that received the medication and the group that did not receive it.

The null hypothesis is tested against an alternative hypothesis (Ha), which is a hypothesis that represents the opposite of the null hypothesis. In the example above, the alternative hypothesis would be that the new medication is effective, and there is a significant difference in blood pressure between the two groups.

The purpose of testing the null hypothesis is to determine whether there is enough evidence to reject it in favor of the alternative hypothesis. If the data collected provides enough evidence to reject the null hypothesis, then the alternative hypothesis is accepted. If the data does not provide enough evidence to reject the null hypothesis, then it is accepted.

It is important to note that failing to reject the null hypothesis does not mean that the null hypothesis is true. It simply means that there is not enough evidence to reject it. Therefore, the null hypothesis is always considered the default assumption until evidence to the contrary is presented.

**Alternative hypothesis**

In statistical hypothesis testing, the alternative hypothesis (Ha) is a hypothesis that represents the opposite of the null hypothesis (H0). It is often a statement that there is a significant difference between a population parameter and a specific value or that there is a relationship between two variables.

For example, if we are interested in testing whether a new medication is effective in reducing blood pressure, the alternative hypothesis would be that the new medication is effective, and there is a significant difference in blood pressure between the group that received the medication and the group that did not receive it. This is the opposite of the null hypothesis, which states that the new medication is not effective and there is no significant difference in blood pressure between the two groups.

The purpose of testing the alternative hypothesis is to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. If the data collected provides enough evidence to reject the null hypothesis, then the alternative hypothesis is accepted. If the data does not provide enough evidence to reject the null hypothesis, then it is accepted.

It is important to note that the alternative hypothesis is often more specific than the null hypothesis. For example, the alternative hypothesis may state that the population parameter is greater than or less than a specific value, whereas the null hypothesis may only state that the population parameter is equal to a specific value. This specificity helps to guide the statistical test and provides more precise conclusions.

In summary, the alternative hypothesis represents the hypothesis that is being tested against the null hypothesis in statistical hypothesis testing. It is often a statement that there is a significant difference or relationship between two variables, and the purpose of testing it is to determine whether there is enough evidence to reject the null hypothesis.

**Types of Error in Hypothesis:**

In statistical hypothesis testing, there are two types of errors that can occur: Type I error and Type II error.

**Type I error** occurs when the null hypothesis (H0) is rejected when it is actually true. This means that the statistical test concludes that there is a significant difference or relationship between two variables when there is not. Type I error is also known as a **false positive.**

For example, if a researcher concludes that a new medication is effective in reducing blood pressure (rejects the null hypothesis) when in fact it is not effective (null hypothesis is true), then this would be a Type I error.

The probability of making a Type I error is denoted by the Greek letter alpha (α) and is usually set at 0.05 or 0.01, which represents the level of significance of the statistical test.

**Type II error** occurs when the null hypothesis (H0) is not rejected when it is actually false. This means that the statistical test concludes that there is no significant difference or relationship between two variables when there is. Type II error is also known as a **false negative**.

For example, if a researcher concludes that a new medication is not effective in reducing blood pressure (fails to reject the null hypothesis) when in fact it is effective (alternative hypothesis is true), then this would be a Type II error.

The probability of making a Type II error is denoted by the Greek letter beta (β) and depends on various factors, including the sample size, the effect size, and the level of significance of the test.

In statistical hypothesis testing, it is important to minimize both Type I and Type II errors, but it is not always possible to do so simultaneously. The balance between these two types of errors is often a trade-off, and researchers must carefully consider the costs and consequences of each type of error when interpreting the results of a statistical test.

**Testing of Hypothesis:**

1. **Small-Scale Test**
2. **Large-Scale Test**

**Small sample tests** are statistical tests that are used when the sample size is small (typically less than 30) and the population standard deviation is unknown. These tests rely on the t-distribution, which is a distribution that takes into account the added uncertainty due to the small sample size.

**Some of the common small sample tests include:**

**t-test:** A t-test is a statistical test used to test a hypothesis about a population mean when the population standard deviation is unknown. It is used when the sample size is small (typically less than 30) and the population standard deviation is unknown.

**Paired t-test:** A paired t-test is a statistical test used to test the difference between two means when the two samples are paired. It is used when the sample size is small (typically less than 30) and the population standard deviation is unknown.

Small sample tests are typically less powerful than large sample tests because of the added uncertainty due to the small sample size. However, they are appropriate when the sample size is small, and the population standard deviation is unknown. It is important to choose the appropriate test based on the research question and the type of data being analysed to ensure accurate and reliable results.

**Large sample tests** are statistical tests that are used when the sample size is sufficiently large (typically more than 30) and the population standard deviation is known or estimated. These tests rely on the central limit theorem, which states that the distribution of the sample mean approaches a normal distribution as the sample size increases.

**Some of the common large sample tests include:**

**Z-test:** A Z-test is a statistical test used to test a hypothesis about a population mean when the population standard deviation is known. It is used when the sample size is large (typically more than 30) and the population standard deviation is known or can be estimated from a previous study.

**Chi-square test:** A chi-square test is a statistical test used to test the **independence** between two categorical variables. It is used when the sample size is large, and the expected cell frequency is at least 5.

**F-test:** An F-test is a statistical test used to test the equality of variances between two or more groups. It is used when the sample size is large, and the population standard deviation is known or can be estimated from a previous study.

**One-sample z-test:** A one-sample z-test is a statistical test used to test a hypothesis about a population mean when the population standard deviation is known. It is used when the sample size is large (typically more than 30) and the population standard deviation is known or can be estimated from a previous study.

Large sample tests are typically more powerful than small sample tests because they are less affected by sampling variability. However, they also require larger sample sizes to be reliable, and they may not be appropriate when the population standard deviation is unknown or when the distribution of the data is highly skewed.

**t-test, F-test, Z-test, and Chi-square test are commonly used statistical tests in hypothesis testing. Here's an overview of each test:**

**t-test:** A t-test is a statistical test used to ***compare the means of two groups***. It is used when the sample size is small (less than 30) or the population standard deviation is unknown. The t-test calculates a t-value, which is used to determine whether the difference between the means is statistically significant.

**F-test:** An F-test is a statistical test used to ***compare the variances of two or more groups***. It is used when the sample size is large (greater than 30) and the population standard deviation is known. The F-test calculates an F-value, which is used to determine whether the variances are statistically significant.

**Z-test:** A Z-test is a statistical test used to ***compare a sample mean to a known population mean.*** It is used when the sample size is large (greater than 30) and the population standard deviation is known. The Z-test calculates a Z-score, which is used to determine whether the difference between the sample mean and population mean is statistically significant.

**Chi-square test:** A chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. It is used when the data are categorical, and the sample size is large. The chi-square test calculates a chi-square statistic, which is used to determine whether the association between the two variables is statistically significant.

Each of these tests has different assumptions and requirements, and the choice of test depends on the research question and the type of data being analyzed. It is important to choose the appropriate test to ensure accurate and reliable results.

**Business Analytics:**

Business analytics is the process of using data analysis techniques and statistical methods to extract insights from business data and use them to make informed decisions. It involves gathering and analyzing data from a variety of sources, including financial data, customer data, market trends, and operational data.

The concept of business analytics is to help organizations make better decisions by providing them with a deeper understanding of their business operations and performance. It helps organizations identify trends, patterns, and relationships in their data, which can be used to develop strategies, improve operations, and optimize decision-making.

Business analytics can be applied to a variety of business functions, including marketing, sales, operations, finance, and human resources. It involves using various analytical tools such as predictive modelling, data visualization, data mining, and machine learning to analyze data and generate insights.

The process of business analytics involves several stages, including data collection, data preparation, data analysis, and data visualization. In the data collection stage, relevant data is collected from various sources and stored in a centralized database. In the data preparation stage, the data is cleaned, transformed, and organized in a way that is suitable for analysis. In the data analysis stage, various statistical methods and analytical tools are used to analyze the data and identify patterns and trends. Finally, in the data visualization stage, the insights generated from the data analysis are presented in a visually appealing format, such as charts, graphs, and dashboards.

Overall, the concept of business analytics is to help organizations make **data-driven decisions** that lead to better business outcomes. By analyzing data and generating insights, businesses can identify areas for improvement, make better decisions, and gain a competitive advantage in the market.

**Application of Business Analytics:**

Business analytics has many applications across different industries and business functions. Some of the common applications of business analytics include:

**Marketing:** Business analytics can be used to analyze customer behavior, identify buying patterns, and develop targeted marketing campaigns. By understanding customer preferences and behavior, businesses can create personalized marketing strategies and improve customer engagement.

**Sales:** Business analytics can be used to analyze sales data, identify trends, and forecast future sales. By understanding customer behavior and market trends, businesses can optimize sales strategies and improve revenue.

**Operations:** Business analytics can be used to optimize supply chain management, production planning, and inventory management. By analyzing operational data, businesses can identify inefficiencies and improve processes to reduce costs and improve productivity.

**Finance:** Business analytics can be used to analyze financial data, identify trends, and forecast future performance. By understanding financial trends and patterns, businesses can make informed decisions about investments, budgeting, and risk management.

**Human resources:** Business analytics can be used to analyze workforce data, identify talent gaps, and optimize hiring and retention strategies. By understanding employee behavior and preferences, businesses can create a more engaged and productive workforce.

**Risk management:** Business analytics can be used to analyze data related to risks, such as fraud, cybersecurity, and compliance. By identifying potential risks and developing mitigation strategies, businesses can protect themselves from potential losses and reputational damage.

Overall, the application of business analytics can help businesses improve decision-making, optimize processes, and gain a competitive advantage in the market. By analyzing data and generating insights, businesses can create more efficient and effective operations, improve customer satisfaction, and increase profitability.

## Descriptive analytics

**Descriptive analytics** is a branch of business analytics that involves analyzing historical data to understand patterns, trends, and insights. It helps to summarize and describe the characteristics of a particular dataset or population, providing information about what has happened in the past.

Descriptive analytics includes a variety of statistical methods and techniques, such as mean, median, mode, standard deviation, and correlation analysis. These methods are used to describe and summarize the characteristics of a dataset, such as the average value, the spread of the data, and the relationship between different variables.

**Some common applications of descriptive analytics include:**

**Business performance analysis:** Descriptive analytics can be used to analyze historical financial data and identify trends in revenue, costs, and profits. It can help businesses understand their performance over time and make informed decisions about future investments.

**Customer behavior analysis:** Descriptive analytics can be used to analyze customer data and identify patterns in their behavior, such as purchase history, product preferences, and demographic information. It can help businesses understand their customers and develop targeted marketing strategies.

**Operations analysis:** Descriptive analytics can be used to analyze operational data, such as production output, inventory levels, and quality control data. It can help businesses identify inefficiencies and areas for improvement in their operations.

**Risk analysis:** Descriptive analytics can be used to analyze historical data related to risks, such as fraud, cyber attacks, and natural disasters. It can help businesses understand the likelihood and impact of these risks and develop mitigation strategies.

Overall, descriptive analytics provides a valuable foundation for other branches of business analytics, such as predictive and prescriptive analytics. By understanding patterns and trends in historical data, businesses can make more informed decisions and improve their performance over time.

## Predictive analytics

**Predictive analytics** is a branch of business analytics that uses statistical modelling and machine learning techniques to analyze data and make predictions about future outcomes. It involves using historical data to identify patterns and trends, and then using these insights to make predictions about what is likely to happen in the future.

Predictive analytics involves a range of techniques, such as regression analysis, decision trees, and neural networks. These techniques are used to build predictive models that can be used to forecast future events, identify trends, and make recommendations.

**Some common applications of predictive analytics include:**

**Sales forecasting:** Predictive analytics can be used to analyse sales data and identify trends in customer behavior, such as changes in buying patterns or preferences. This information can be used to forecast future sales and optimize sales strategies.

**Customer churn prediction:** Predictive analytics can be used to analyse customer data and identify factors that are likely to lead to customer churn, such as low engagement or dissatisfaction. This information can be used to develop retention strategies and improve customer loyalty.

**Fraud detection:** Predictive analytics can be used to analyse transactional data and identify patterns of fraudulent activity. This information can be used to detect and prevent fraud in real-time.

**Supply chain optimization:** Predictive analytics can be used to analyse supply chain data and identify potential bottlenecks or inefficiencies. This information can be used to optimize production and logistics processes.

**Risk management:** Predictive analytics can be used to analyse data related to risks, such as credit risk or cybersecurity risk. This information can be used to identify potential risks and develop mitigation strategies.

Overall, predictive analytics provides businesses with the ability to make more informed decisions based on insights about future events. By identifying patterns and trends in historical data, businesses can forecast future outcomes, make better predictions, and optimize their operations to achieve better results.

# Statistics Cheat Sheet

## Population

The entire group one desires information about

## Sample

A subset of the population taken because the entire population is usually too large to analyze

Its characteristics are taken to be representative of the population

## Mean

Also called the arithmetic mean or average

The sum of all the values in the sample divided by the number of values in the sample/population

$\mu$ is the mean of the population; $\bar{x}$ is the mean of the sample

## Median

The value separating the higher half of a sample/population from the lower half

Found by arranging all the values from lowest to highest and taking the middle one (or the mean of the middle two if there are an even number of values)

## Variance

Measures dispersion around the mean

Determined by averaging the squared differences of all the values from the mean

Variance of a population is $\sigma^2$

$$\sigma^2 = \frac{\sum (x-\mu)^2}{n}$$

Can be calculated by subtracting the square of the mean from the average of the squared scores:

$$\sigma^2 = \frac{\sum x^2}{n} - \mu^2$$

Variance of a sample is $s^2$; note the *n-1*

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

Can be calculated by:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

## Standard Deviation

Square root of the variance

Also measures dispersion around the mean but in the same units as the values (instead of square units with variance)

$\sigma$ is the standard deviation of the population and $s$ is the standard deviation of the sample

## Standard Error

An estimate of the standard deviation of the sampling distribution—the set of all samples of size *n* that can be taken from a population

Reflects the extent to which a statistic changes from sample to sample

For a mean, $\frac{s}{\sqrt{n}}$

For the difference between two means,

Assuming equal variances $\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$; unequal variances $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

## T-test

### *One-Sample*

Tests whether the mean of a normally distributed population is different from a specified value

Null Hypothesis ($H_0$): states that the population mean is equal to some value ($\mu_0$)
Alternative Hypothesis ($H_a$): states that the mean does not equal/is greater than/is less than $\mu_0$
t-statistic: standardizes the difference between $\bar{x}$ and $\mu_0$

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Degrees of freedom (df) = *n-1*

Read the table of t-distribution critical values for the p-value (probability that the sample mean was obtained by chance given $\mu_0$ is the population mean) using the calculated t-statistic and degrees of freedom.

$H_a$: $\mu>\mu_0$ → the t-statistic is likely positive; read table as given

$H_a$: $\mu<\mu_0$ → the t-statistic is likely negative; the t-distribution is symmetrical so read the probability as if the t-statistic were positive

Note: if the t-statistic is of the 'wrong' sign, the p-value is 1 minus the *p* given in the chart

$H_a$: $\mu\neq\mu_0$ → read the p-value as if the t-statistic were positive and double it (to consider both less than and greater than)

If the p-value is less than the predetermined value for significance (called $\alpha$ and is usually 0.05), reject the null hypothesis and accept the alternative hypothesis.

***Example***:
You are experiencing hair loss and skin discoloration and think it might be because of selenium toxicity. You decide to measure the selenium levels in your tap water once a day for one week. Your results are given below. The EPA maximum contaminant level for safe drinking water is 0.05 mg/L. Does the selenium level in your tap water exceed the legal limit (assume $\alpha$=0.05)?

| Day | Selenium mg/L |
|---|---|
| 1 | 0.051 |
| 2 | 0.0505 |
| 3 | 0.049 |
| 4 | 0.0516 |
| 5 | 0.052 |
| 6 | 0.0508 |
| 7 | 0.0506 |

$H_0$: $\mu$=0.05; $H_a$: $\mu$>0.05
Calculate the mean and standard deviation of your sample:

$$\bar{x} = 0.0508$$

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{(0.051-0.0508)^2 + (0.0505-0.0508)^2 + etc...}{6} = 9.15\times10^{-7}$$

$$s = \sqrt{s^2} = 9.56\times10^{-4}$$

The t-statistic is: $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.0508-0.05}{\frac{9.56\times10^{-4}}{\sqrt{7}}} = 2.17$ and the degrees of freedom are *n-1* = 7-1 = 6

Looking at the t-distribution of critical values table, 2.17 with 6 degrees of freedom is between *p*=0.05 and *p*=0.025. This means that the p-value is less than 0.05, so you can reject $H_0$ and conclude that the selenium level in your tap water exceeds the legal limit.

## T-test

### *Two-Sample*

Tests whether the means of two populations are significantly different from one another

***Paired***

Each value of one group corresponds directly to a value in the other group; ie: before and after values after drug treatment for each individual patient

Subtract the two values for each individual to get one set of values (the differences) and use $\mu_0 = 0$ to perform a one-sample t-test

***Unpaired***

The two populations are independent

$H_0$: states that the means of the two populations are equal ($\mu_1=\mu_2$)

$H_a$: states that the means of the two populations are unequal or one is greater than the other ($\mu_1\neq\mu_2$, $\mu_1>\mu_2$, $\mu_1<\mu_2$)

t-statistic:

assuming equal variances: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ assuming unequal variances: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}}$

degrees of freedom = ($n_1$-1)+($n_2$-1)

Read the table of t-distribution critical values for the p-value using the calculated t-statistic and degrees of freedom. Remember to keep the sign of the t-statistic clear (order of subtracting the sample means) and to double the p-value for an $H_a$ of $\mu_1 \neq \mu_2$.

***Example:***

Consider the lifespan of 18 rats. 12 were fed a restricted calorie diet and lived an average of 700 days (standard deviation=21 days). The other 6 had unrestricted access to food and lived an average of 668 days (standard deviation=30 days). Does a restricted calorie diet increase the lifespan of rats (assume $\alpha$=0.05)?

$\mu_1$=700, $s_1$=21, $n_1$=12; $\mu_2$=668, $s_2$=30, $n_2$=6

$H_0$: $\mu_1=\mu_2$

$H_a$: $\mu_1>\mu_2$ (because we are only asking if a restricted calorie diet increases lifespan)

We cannot assume that the variances of the two populations are equal because the different diets could also affect the variability in lifespan.

The t-statistic is: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}} = \dfrac{700-668}{\sqrt{\dfrac{21^2}{12} + \dfrac{30^2}{6}}} = 2.342$

Degrees of freedom = ($n_1$-1)+($n_2$-1) = (12-1)+(6-1)=16

From the t-distribution table, the p-value falls between 0.01 and 0.02, so we do reject $H_0$. The restricted calorie diet does increase the lifespan of rats.

## Chi-Square Test

### *For Goodness of Fit*

Checks whether or not an observed pattern of data fits some given distribution

$H_0$: the observed pattern fits the given distribution

$H_a$: the observed pattern does not fit the given distribution

The chi-square statistic is: $\chi^2 = \sum \dfrac{(O-E)^2}{E}$ (*O* is the observed value and *E* is the expected value)

Degrees of freedom = number of categories in the distribution – 1

Get the p-value from the table of $\chi^2$ critical values using the calculated $\chi^2$ and df values. If the p-value is less than $\alpha$, the observed data does not fit the expected distribution. If $p>\alpha$, the data likely fits the expected distribution

***Example 1*:**

You breed puffskeins and would like to determine the pattern of inheritance for coat color and purring ability. Puffskeins come in either pink or purple and can either purr or hiss. You breed a purebred, pink purring male with a purebred, purple hissing female. All individuals of the $F_1$ generation are pink and purring. The $F_2$ offspring are shown below. Do the alleles for coat color and purring ability assort independently (assume $\alpha$=0.05)?

| Pink and Purring | Pink and Hissing | Purple and Purring | Purple and Hissing |
|---|---|---|---|
| 143 | 60 | 55 | 18 |

Independent assortment means a phenotypic ratio of 9:3:3:1, so:

$H_0$: the observed distribution of $F_2$ offspring fits a 9:3:3:1 distribution

$H_a$: the observed distribution of $F_2$ offspring does not fit a 9:3:3:1 distribution

The expected values are:

| Pink and Purring | Pink and Hissing | Purple and Purring | Purple and Hissing |
|---|---|---|---|
| 155.25 | 51.75 | 51.75 | 17.25 |

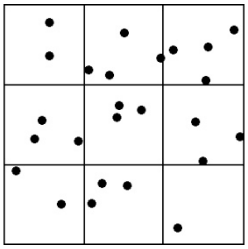
$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(143-155.25)^2}{155.25} + \frac{(60-51.75)^2}{51.75} + \frac{(55-51.75)^2}{51.75} + \frac{(18-17.25)^2}{17.25} = 2.519$$

df=4-1=3

From the table of $\chi^2$ critical values, the p-value is greater than 0.25, so the alleles for coat color and purring ability do assort independently in puffskeins.

***Example 2:***

You are studying the pattern of dispersion of king penguins and the diagram on the right represents an area you sampled. Each dot is a penguin. Do the penguins display a uniform distribution (assume $\alpha$=0.05)?



$H_0$: there is a uniform distribution of penguins

$H_a$: there is not a uniform distribution of penguins

There are a total of 25 penguins, so if there is a uniform distribution, there should be 2.778 penguins per square. There actual observed values are 2, 4, 4, 3, 3, 3, 2, 3, 1, so the $\chi^2$ statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(1-2.778)^2}{2.778} + 2\left(\frac{(2-2.778)^2}{2.778}\right) + 4\left(\frac{(3-2.778)^2}{2.778}\right) + 2\left(\frac{(4-2.778)^2}{2.778}\right) = 2.72$$

df=9-1=8

From the table of $\chi^2$ critical values, the p-value is greater than 0.25, so we do not reject $H_0$. The penguins do display a uniform distribution.

## Chi-Square Test

### *For Independence*

Checks whether two categorical variables are related or not (independence)

$H_0$: the two variables are independent

$H_a$: the two variables are not independent

Does not make any assumptions about an expected distribution

The observed values ($\#_1$, $\#_2$, $\#_3$, and $\#_4$) are usually presented as a table. Each row is a category of variable 1 and each column is a category of variable 2.

| | | Variable 1 | | Totals |
|---|---|---|---|---|
| | | Category *x* | Category *y* | |
| Variable 2 | Category *a* | $\#_1$ | $\#_2$ | $\#_1$+$\#_2$ |
| | Category *b* | $\#_3$ | $\#_4$ | $\#_3$+$\#_4$ |
| Totals | | $\#_1$+$\#_3$ | $\#_2$+$\#_4$ | $\#_1$+$\#_2$+$\#_3$+$\#_4$ |

The proportion of category *x* of variable 1 is the number of individuals in category *x* divided by the total number of individuals $\left(\frac{\#_1+\#_3}{\#_1+\#_2+\#_3+\#_4}\right)$. Assuming independence, the expected number of individuals that fall within category *a* of variable 2 is the proportion of category *x* multiplied by the number of individuals in category *a* $\left(\frac{\#_1+\#_3}{\#_1+\#_2+\#_3+\#_4}\right)(\#_1+\#_2)$. Thus, the expected value is:

$$E = \frac{(\#_1+\#_3)(\#_1+\#_2)}{\#_1+\#_2+\#_3+\#_4} = \frac{(row\ \ total)(column\ \ total)}{grand\ \ total}$$

Degrees of freedom = (*r*-1)(*c*-1)    where *r* is the number of rows and *c* is the number of columns

The chi-square statistic is still $\chi^2 = \sum \frac{(O-E)^2}{E}$

Read the p-values from the table of $\chi^2$ critical values.

***Example:***

Given the data below, is there a relationship between fitness level and smoking habits (assume $\alpha$=0.05)?

| | Fitness Level | | | | |
|---|---|---|---|---|---|
| | Low | Medium-Low | Medium-High | High | |
| Never smoked | 113 | 113 | 110 | 159 | 495 |
| Former smokers | 119 | 135 | 172 | 190 | 616 |
| 1 to 9 cigarettes daily | 77 | 91 | 86 | 65 | 319 |
| $\geq$ 10 cigarettes daily | 181 | 152 | 124 | 73 | 530 |
| | 490 | 491 | 492 | 487 | 1960 |

$H_0$: fitness level and smoking habits are independent
$H_a$: fitness level and smoking habits are not independent
First, we calculate the expected counts. For the first cell, the expected count is:

$$E = \frac{(row\ total)(column\ total)}{grand\ total} = \frac{(495)(490)}{1960} = 123.75$$

| | Fitness Level | | | |
|---|---|---|---|---|
| | Low | Medium-Low | Medium-High | High |
| Never smoked | 123.75 | 124 | 124.26 | 122.99 |
| Former smokers | 154 | 154.31 | 154.63 | 153.06 |
| 1 to 9 cigarettes daily | 79.75 | 79.91 | 80.08 | 79.26 |
| ≥ 10 cigarettes daily | 132.5 | 132.77 | 133.04 | 131.69 |

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(113-123.75)^2}{123.75} + \frac{(113-124)^2}{124} + \frac{(110-124.26)^2}{124.26} + etc\ldots = 91.73$$

df=$(r$-1$)(c$-1$)$=(4-1)(4-1)=9

From the table of $\chi^2$ critical values, the p-value is less than 0.001, so we reject $H_0$ and conclude that there is a relationship between fitness level and smoking habits.

## Type I error

The probability of rejecting a true null hypothesis
Equals $\alpha$

## Type II error

The probability of failing to reject a false null hypothesis

## Probability

### *Joint Probability*

The probability of events A and B occurring
$P(A \text{ and } B) = P(A) \times P(B)$ when events A and B are independent

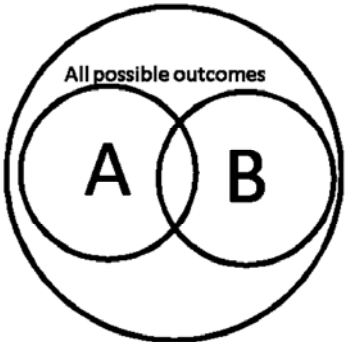
### *Union of Events*

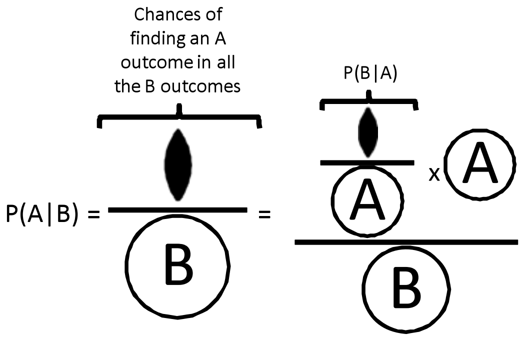
The probability of either event A or event B occurring
$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

### *Conditional Probability*

The probability of event A occurring given that event B has occurred

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \qquad \text{or} \qquad P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$





***Example 1***:
Assume that eye color is an autosomally inherited trait controlled by one gene with two alleles. Brown is dominant to blue. A brown-eyed man with genotype Bb and a blue-eyed woman have three children. The first has blue eyes. What is the probability that all three children have blue eyes?

Without considering the first child, the probability that the couple has three children with blue eyes is $0.5 \times 0.5 \times 0.5 = 0.125 = P(A \text{ and } B) = P(2 \text{ children} = \text{bb and 1st child bb})$

With his parents, the probability that the 1st child is bb is: $P(B) = P(\text{1st child} = \text{bb}) = 0.5$

Therefore, $P(2 \text{ children} = \text{bb} \mid \text{ 1st child bb}) = P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.125}{0.5} 0.25$

***Example 2:***

Based on an analysis of her pedigree, it is determined that a woman has a 70% chance of being Zz and a 30% chance of being ZZ for a sex-linked trait, where Z is dominant to z. If she now has a son with the Z phenotype, what is the probability of her being Zz?

We're looking for: P(W=Zz|S=Z)

But it's hard to find P(W=Zz and S=Z) because the two events are not independent. Instead, let us use:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

$P(S = Z \mid W = Zz) = 0.5$ (50% chance of passing on the Z allele)

$P(W = Zz) = 0.7$ (given)

$P(S = Z) = (0.7 \times 0.5) + (0.3 \times 1) = 0.65$ (son can be Z from the woman being either Zz or ZZ)

$$P(W = Zz \mid S = Z) = \frac{0.5 \times 0.7}{0.65} = 0.538$$

## *Multiple Experiments*

### ***Binomial distribution***

For when you are not concerned about the order of the events, only that they occur

$$P(X = m) = \frac{n! \times p^m \times (1-p)^{(n-m)}}{m! \times (n-m)!}$$

for *m* outcomes of event X in *n* total trials with *p*=probability of X occurring once

***Example:***

What is the probability that a couple has one boy out of five children?

$$P(\text{1 boy of 5 children}) = \frac{5! \times 0.5^1 \times 0.5^4}{1! \times (4)!} = 0.15625$$

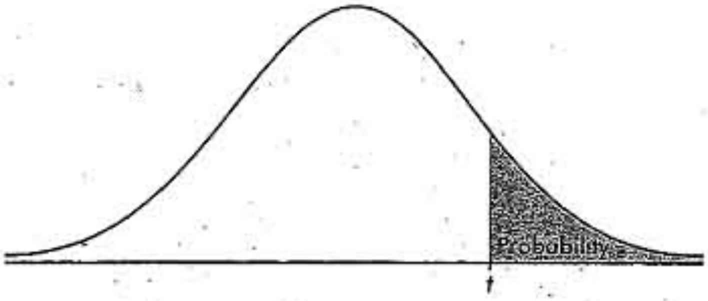
### ***Poisson distribution***

The binomial distribution works for a small number of trials but as *n* gets too large, the factorials become unwieldy.

The Poisson distribution is an estimate of the binomial distribution for large *n*.

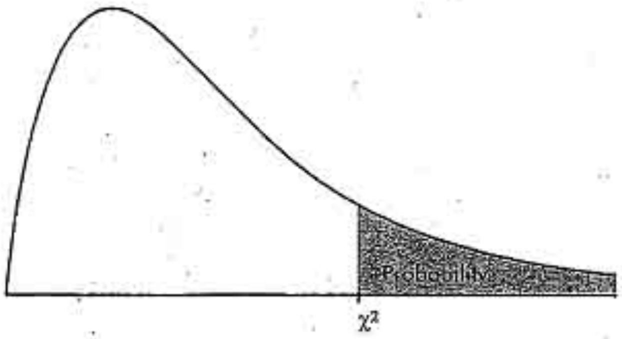
$$P(X = m) = \frac{e^{-np} \times (n \times p)^m}{m!}$$

Note: *np* is also known as the number of expected outcomes for event X



## TABLE B: t-DISTRIBUTION CRITICAL VALUES

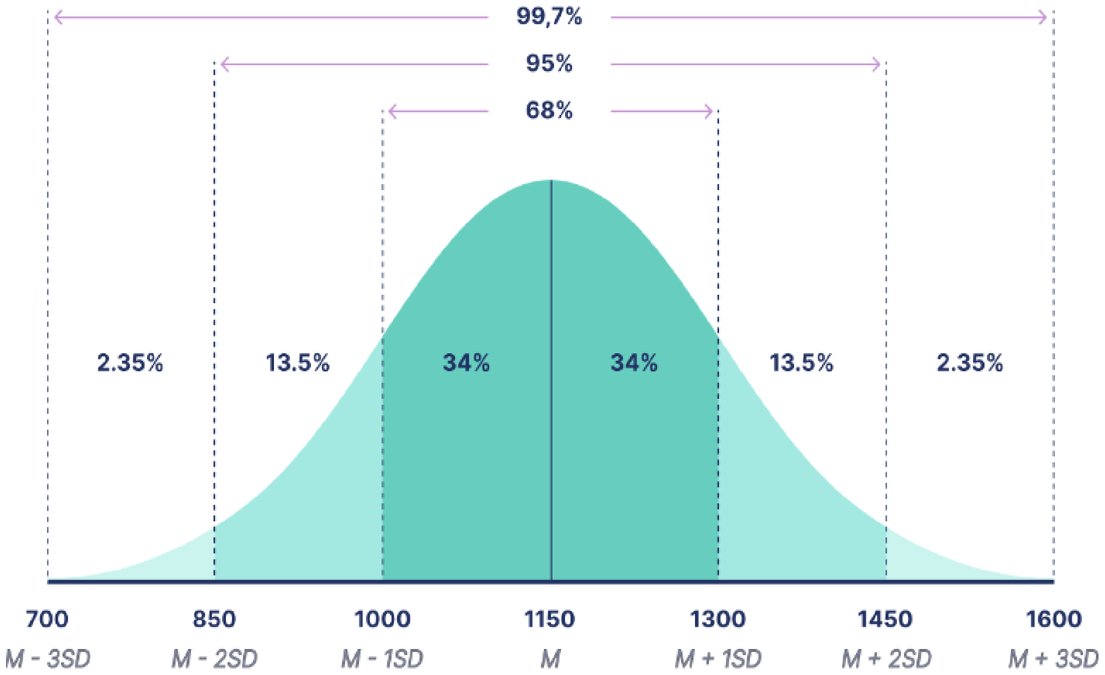
| | Tail probability $p$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| df | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 | .0005 |
| 1 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 15.89 | 31.82 | 63.66 | 127.3 | 318.3 | 636.6 |
| 2 | .816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 4.849 | 6.965 | 9.925 | 14.09 | 22.33 | 31.60 |
| 3 | .765 | .978 | 1.250 | 1.638 | 2.353 | 3.182 | 3.482 | 4.541 | 5.841 | 7.453 | 10.21 | 12.92 |
| 4 | .741 | .941 | 1.190 | 1.533 | 2.132 | 2.776 | 2.999 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | .727 | .920 | 1.156 | 1.476 | 2.015 | 2.571 | 2.757 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | .718 | .906 | 1.134 | 1.440 | 1.943 | 2.447 | 2.612 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | .711 | .896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.517 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | .706 | .889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.449 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | .703 | .883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.398 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | .700 | .879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.359 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | .697 | .876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.328 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | .695 | .873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.303 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | .694 | .870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.282 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | .692 | .868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.264 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | .691 | .866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.249 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | .690 | .865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.235 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | .689 | .863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.224 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | .688 | .862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.214 | 2.552 | 2.878 | 3.197 | 3.611 | 3.922 |
| 19 | .688 | .861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.205 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | .687 | .860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.197 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| 21 | .686 | .859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.189 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| 22 | .686 | .858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.183 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |
| 23 | .685 | .858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.177 | 2.500 | 2.807 | 3.104 | 3.485 | 3.768 |
| 24 | .685 | .857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.172 | 2.492 | 2.797 | 3.091 | 3.467 | 3.745 |
| 25 | .684 | .856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.167 | 2.485 | 2.787 | 3.078 | 3.450 | 3.725 |
| 26 | .684 | .856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.162 | 2.479 | 2.779 | 3.067 | 3.435 | 3.707 |
| 27 | .684 | .855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.158 | 2.473 | 2.771 | 3.057 | 3.421 | 3.690 |
| 28 | .683 | .855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.154 | 2.467 | 2.763 | 3.047 | 3.408 | 3.674 |
| 29 | .683 | .854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.150 | 2.462 | 2.756 | 3.038 | 3.396 | 3.659 |
| 30 | .683 | .854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.147 | 2.457 | 2.750 | 3.030 | 3.385 | 3.646 |
| 40 | .681 | .851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.123 | 2.423 | 2.704 | 2.971 | 3.307 | 3.551 |
| 50 | .679 | .849 | 1.047 | 1.299 | 1.676 | 2.009 | 2.109 | 2.403 | 2.678 | 2.937 | 3.261 | 3.496 |
| 60 | .679 | .848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.099 | 2.390 | 2.660 | 2.915 | 3.232 | 3.460 |
| 80 | .678 | .846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.088 | 2.374 | 2.639 | 2.887 | 3.195 | 3.416 |
| 100 | .677 | .845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.081 | 2.364 | 2.626 | 2.871 | 3.174 | 3.390 |
| 1000 | .675 | .842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.056 | 2.330 | 2.581 | 2.813 | 3.098 | 3.300 |
| ∞ | .674 | .841 | 1.036 | 1.282 | 1.645 | 1.960 | 2.054 | 2.326 | 2.576 | 2.807 | 3.091 | 3.291 |
| | 50% | 60% | 70% | 80% | 90% | 95% | 96% | 98% | 99% | 99.5% | 99.8% | 99.9% |
| | Confidence level $C$ | | | | | | | | | | | |



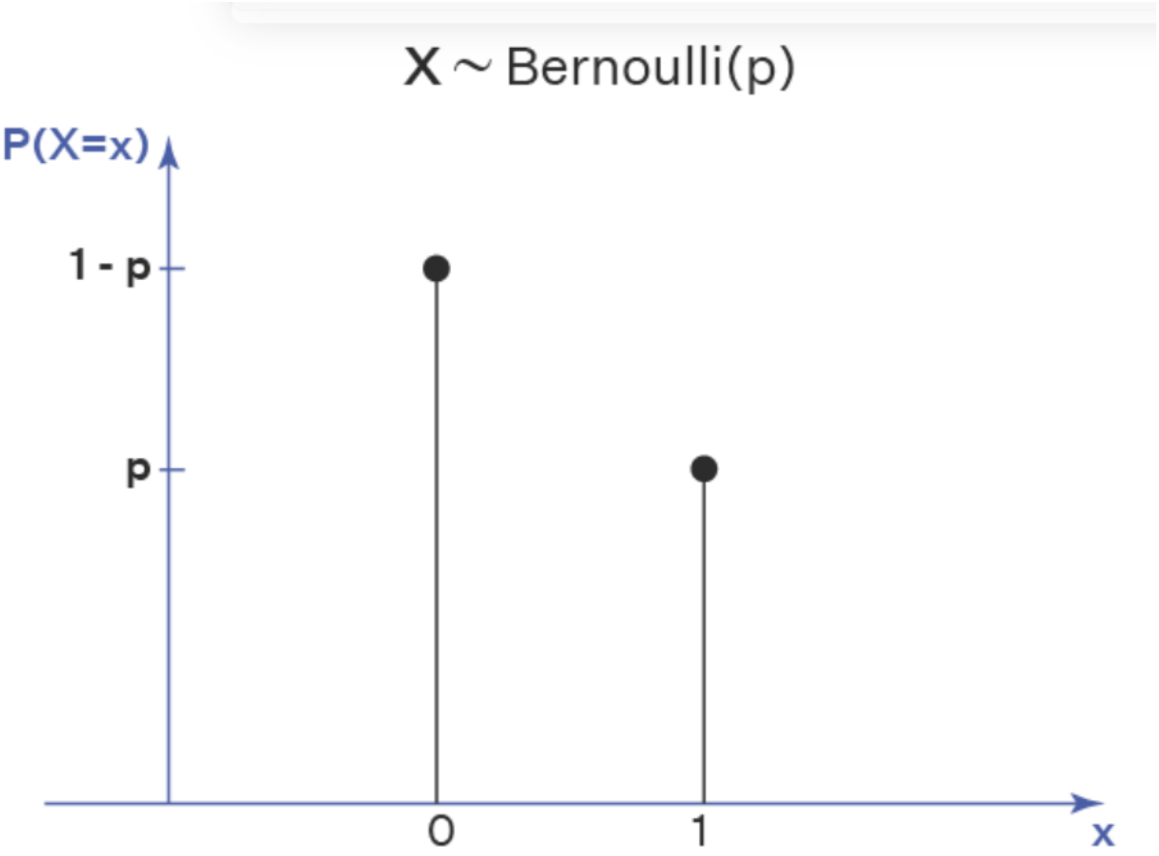
## TABLE C: $\chi^2$ CRITICAL VALUES

| | Tail probability $p$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| df | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 |
| 1 | 1.32 | 1.64 | 2.07 | 2.71 | 3.84 | 5.02 | 5.41 | 6.63 | 7.88 | 9.14 | 10.83 |
| 2 | 2.77 | 3.22 | 3.79 | 4.61 | 5.99 | 7.38 | 7.82 | 9.21 | 10.60 | 11.98 | 13.82 |
| 3 | 4.11 | 4.64 | 5.32 | 6.25 | 7.81 | 9.35 | 9.84 | 11.34 | 12.84 | 14.32 | 16.27 |
| 4 | 5.39 | 5.99 | 6.74 | 7.78 | 9.49 | 11.14 | 11.67 | 13.28 | 14.86 | 16.42 | 18.47 |
| 5 | 6.63 | 7.29 | 8.12 | 9.24 | 11.07 | 12.83 | 13.39 | 15.09 | 16.75 | 18.39 | 20.51 |
| 6 | 7.84 | 8.56 | 9.45 | 10.64 | 12.59 | 14.45 | 15.03 | 16.81 | 18.55 | 20.25 | 22.46 |
| 7 | 9.04 | 9.80 | 10.75 | 12.02 | 14.07 | 16.01 | 16.62 | 18.48 | 20.28 | 22.04 | 24.32 |
| 8 | 10.22 | 11.03 | 12.03 | 13.36 | 15.51 | 17.53 | 18.17 | 20.09 | 21.95 | 23.77 | 26.12 |
| 9 | 11.39 | 12.24 | 13.29 | 14.68 | 16.92 | 19.02 | 19.68 | 21.67 | 23.59 | 25.46 | 27.88 |
| 10 | 12.55 | 13.44 | 14.53 | 15.99 | 18.31 | 20.48 | 21.16 | 23.21 | 25.19 | 27.11 | 29.59 |
| 11 | 13.70 | 14.63 | 15.77 | 17.28 | 19.68 | 21.92 | 22.62 | 24.72 | 26.76 | 28.73 | 31.26 |
| 12 | 14.85 | 15.81 | 16.99 | 18.55 | 21.03 | 23.34 | 24.05 | 26.22 | 28.30 | 30.32 | 32.91 |
| 13 | 15.98 | 16.98 | 18.20 | 19.81 | 22.36 | 24.74 | 25.47 | 27.69 | 29.82 | 31.88 | 34.53 |
| 14 | 17.12 | 18.15 | 19.41 | 21.06 | 23.68 | 26.12 | 26.87 | 29.14 | 31.32 | 33.43 | 36.12 |
| 15 | 18.25 | 19.31 | 20.60 | 22.31 | 25.00 | 27.49 | 28.26 | 30.58 | 32.80 | 34.95 | 37.70 |
| 16 | 19.37 | 20.47 | 21.79 | 23.54 | 26.30 | 28.85 | 29.63 | 32.00 | 34.27 | 36.46 | 39.25 |
| 17 | 20.49 | 21.61 | 22.98 | 24.77 | 27.59 | 30.19 | 31.00 | 33.41 | 35.72 | 37.95 | 40.79 |
| 18 | 21.60 | 22.76 | 24.16 | 25.99 | 28.87 | 31.53 | 32.35 | 34.81 | 37.16 | 39.42 | 42.31 |
| 19 | 22.72 | 23.90 | 25.33 | 27.20 | 30.14 | 32.85 | 33.69 | 36.19 | 38.58 | 40.88 | 43.82 |
| 20 | 23.83 | 25.04 | 26.50 | 28.41 | 31.41 | 34.17 | 35.02 | 37.57 | 40.00 | 42.34 | 45.31 |
| 21 | 24.93 | 26.17 | 27.66 | 29.62 | 32.67 | 35.48 | 36.34 | 38.93 | 41.40 | 43.78 | 46.80 |
| 22 | 26.04 | 27.30 | 28.82 | 30.81 | 33.92 | 36.78 | 37.66 | 40.29 | 42.80 | 45.20 | 48.27 |
| 23 | 27.14 | 28.43 | 29.98 | 32.01 | 35.17 | 38.08 | 38.97 | 41.64 | 44.18 | 46.62 | 49.73 |
| 24 | 28.24 | 29.55 | 31.13 | 33.20 | 36.42 | 39.36 | 40.27 | 42.98 | 45.56 | 48.03 | 51.18 |
| 25 | 29.34 | 30.68 | 32.28 | 34.38 | 37.65 | 40.65 | 41.57 | 44.31 | 46.93 | 49.44 | 52.62 |
| 26 | 30.43 | 31.79 | 33.43 | 35.56 | 38.89 | 41.92 | 42.86 | 45.64 | 48.29 | 50.83 | 54.05 |
| 27 | 31.53 | 32.91 | 34.57 | 36.74 | 40.11 | 43.19 | 44.14 | 46.96 | 49.64 | 52.22 | 55.48 |
| 28 | 32.62 | 34.03 | 35.71 | 37.92 | 41.34 | 44.46 | 45.42 | 48.28 | 50.99 | 53.59 | 56.89 |
| 29 | 33.71 | 35.14 | 36.85 | 39.09 | 42.56 | 45.72 | 46.69 | 49.59 | 52.34 | 54.97 | 58.30 |
| 30 | 34.80 | 36.25 | 37.99 | 40.26 | 43.77 | 46.98 | 47.96 | 50.89 | 53.67 | 56.33 | 59.70 |
| 40 | 45.62 | 47.27 | 49.24 | 51.81 | 55.76 | 59.34 | 60.44 | 63.69 | 66.77 | 69.70 | 73.40 |
| 50 | 56.33 | 58.16 | 60.35 | 63.17 | 67.50 | 71.42 | 72.61 | 76.15 | 79.49 | 82.66 | 86.66 |
| 60 | 66.98 | 68.97 | 71.34 | 74.40 | 79.08 | 83.30 | 84.58 | 88.38 | 91.95 | 95.34 | 99.61 |
| 80 | 88.13 | 90.41 | 93.11 | 96.58 | 101.9 | 106.6 | 108.1 | 112.3 | 116.3 | 120.1 | 124.8 |
| 100 | 109.1 | 111.7 | 114.7 | 118.5 | 124.3 | 129.6 | 131.1 | 135.8 | 140.2 | 144.3 | 149.4 |

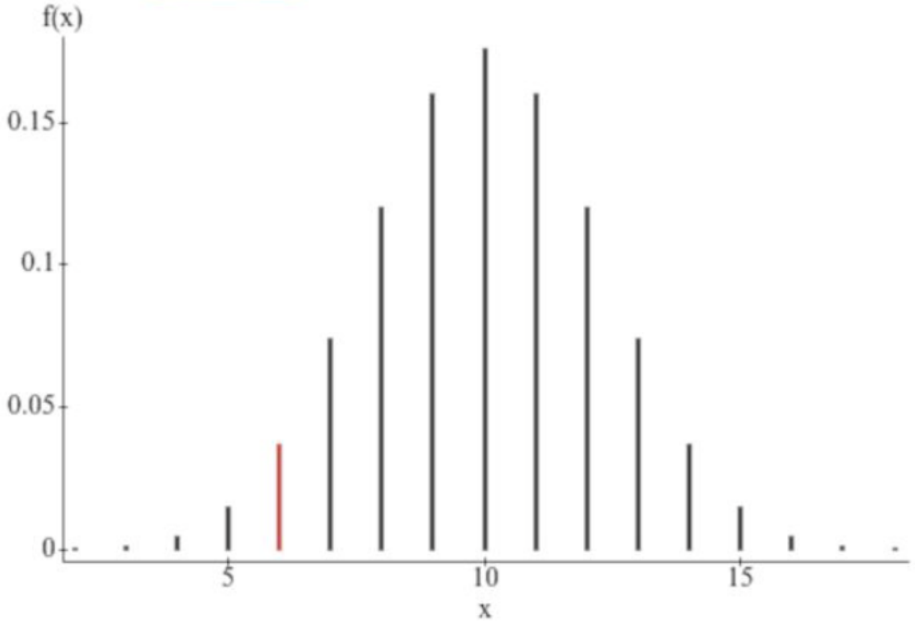
## Normal Distribution



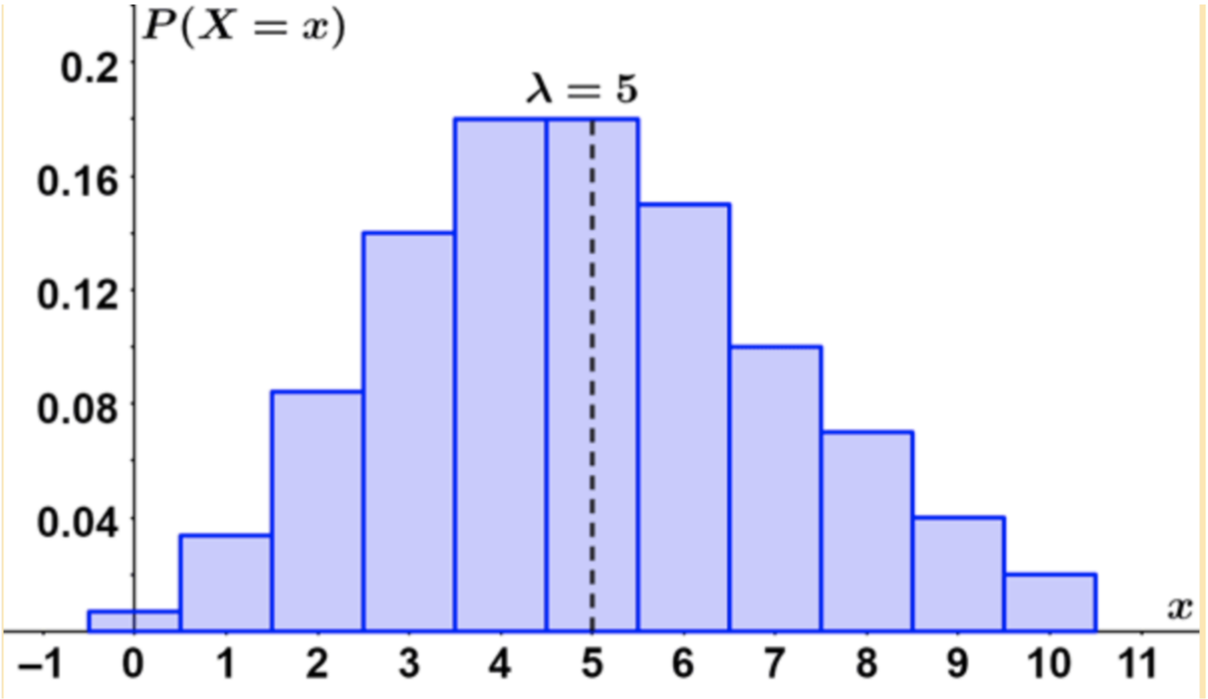
## Bernoulli Distribution



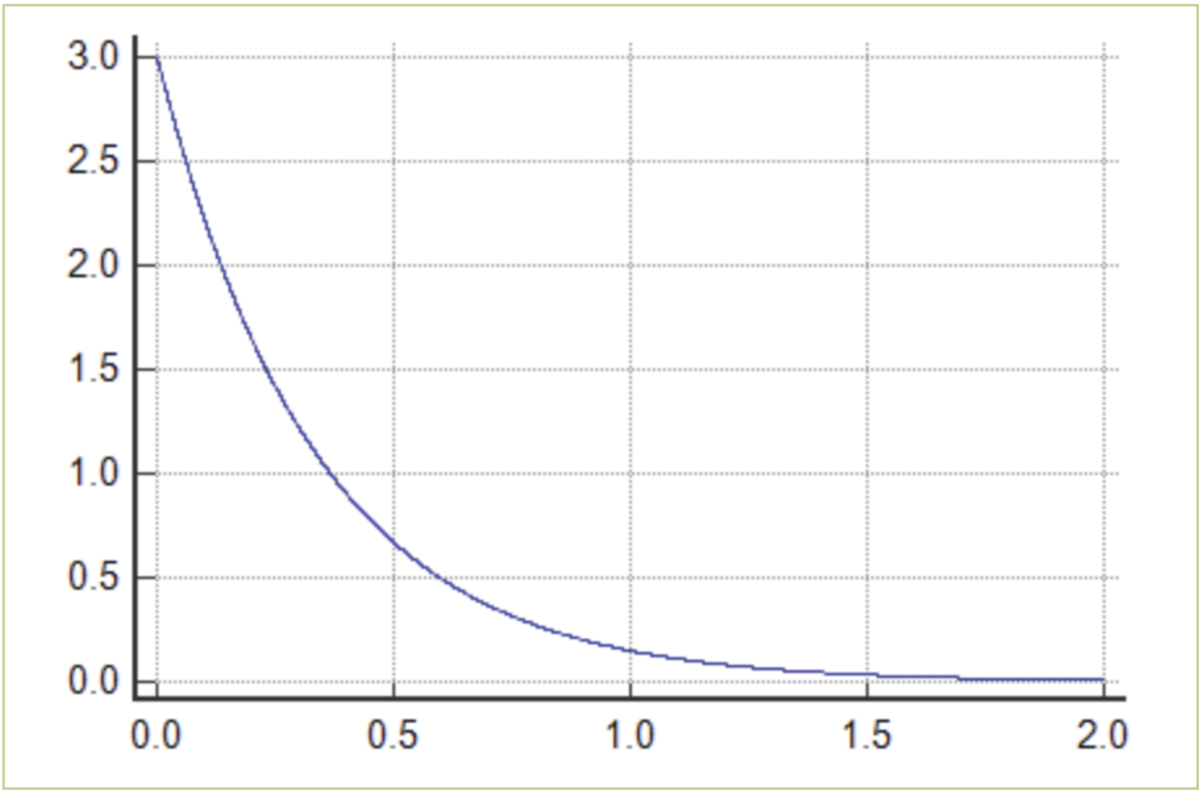
# Binomial Distribution



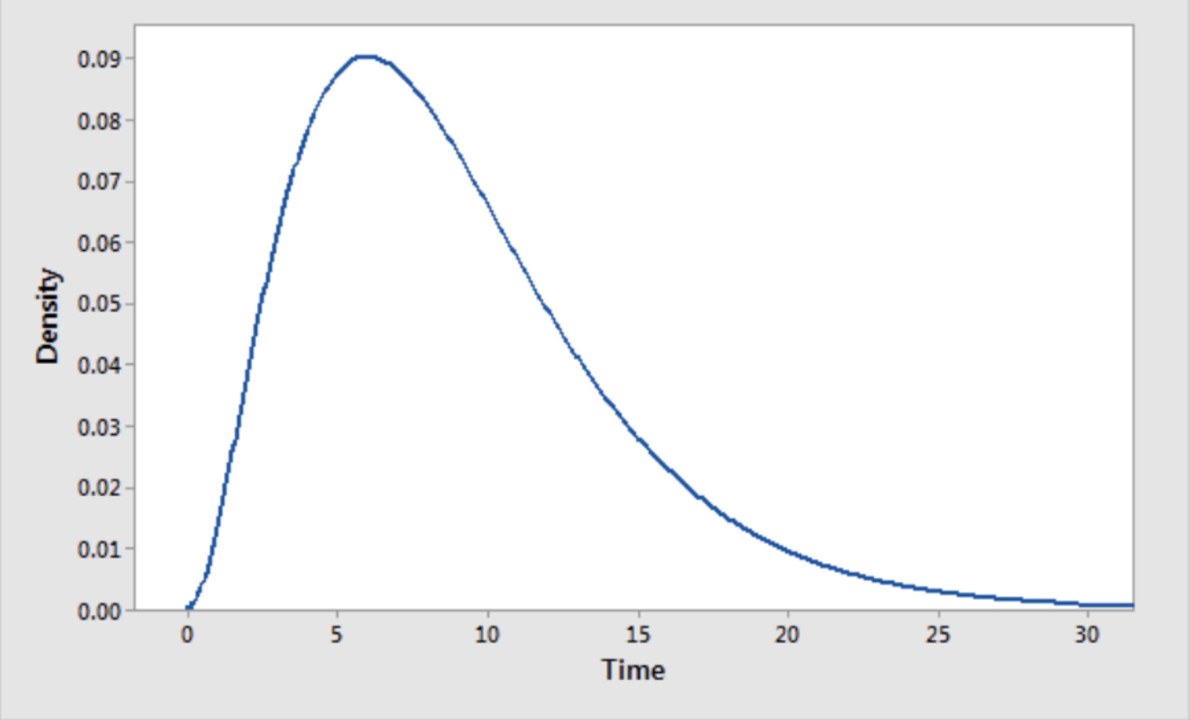
# Poisson Distribution



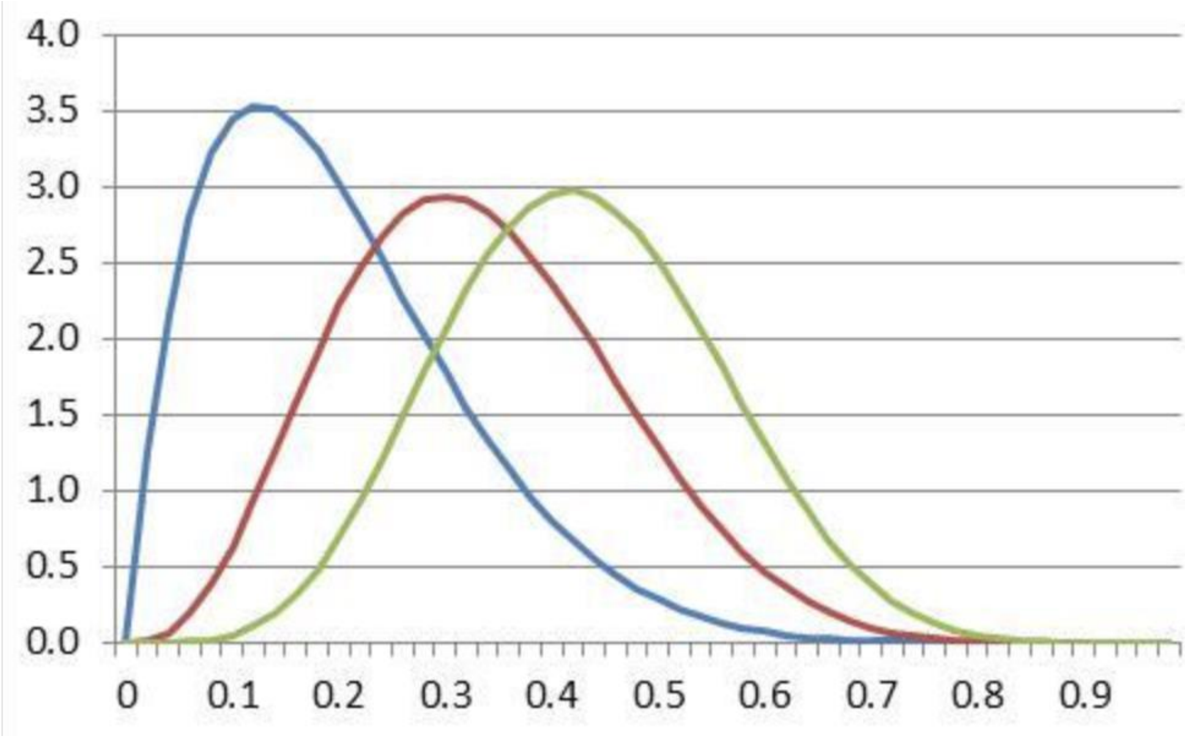
# Exponential Distribution



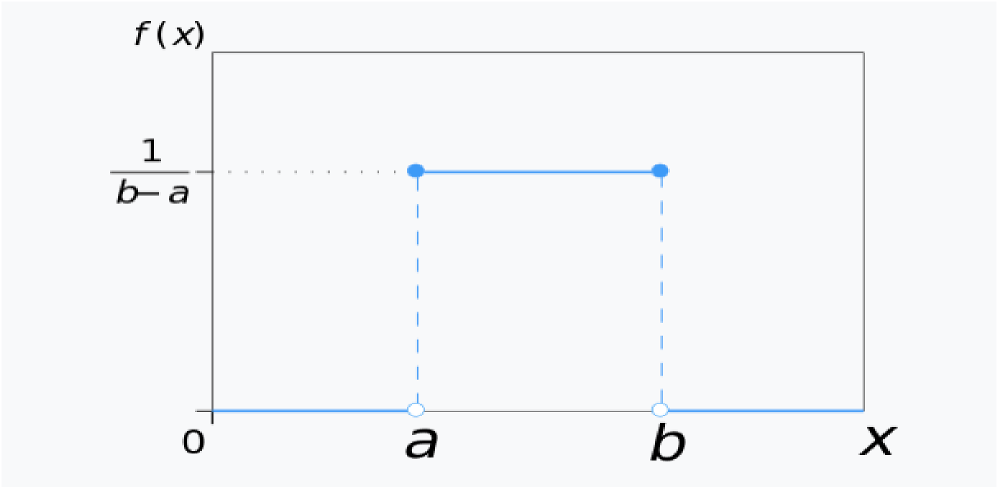
# Gamma Distribution



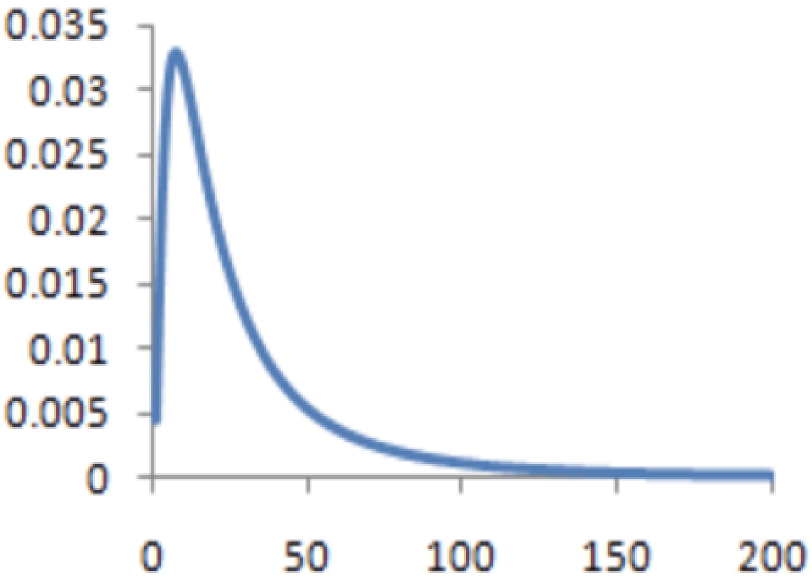
## Beta Distribution



## Uniform Distribution



# Log Normal Distribution


0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
50
100
150
200