MEASURES OF CENTRAL TENDENCY

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean.

The mode is the *most commonly occurring value* **in a distribution.**

Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This table shows a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data

In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

The median is the *middle value* **in distribution when the values are arranged in ascending or descending order.**

The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 years:

52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

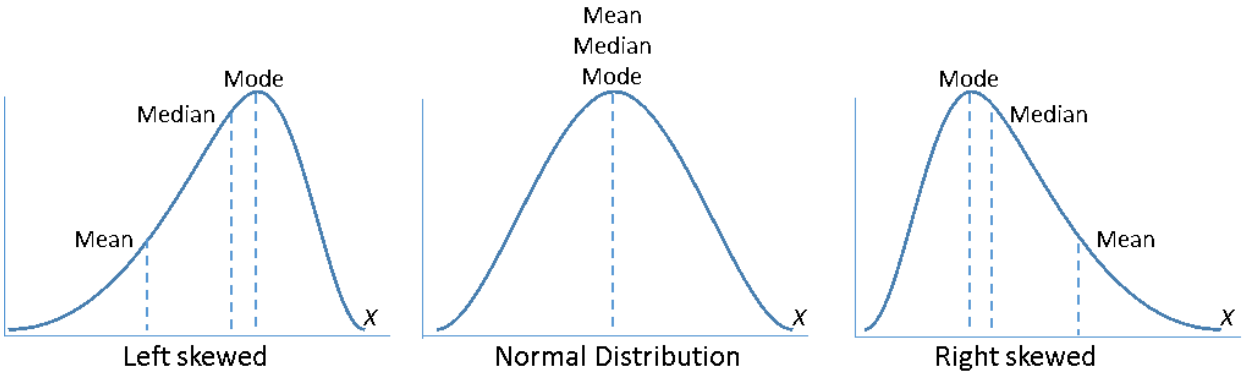
Looking at the retirement age distribution again:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

The mean can be used for both continuous and discrete numeric data.

The mean cannot be calculated for categorical data, as the values cannot be summed.



Harmonic Mean

A simple way to define a harmonic mean is to call it the reciprocal of the arithmetic mean of the reciprocals of the observations. The most important criteria for it is that none of the observations should be zero.

A harmonic mean is used in averaging of ratios. The most common examples of ratios are that of speed and time, cost and unit of material, work and time etc. The harmonic mean (H.M.) of n observations is

$$\mathrm{HM}(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

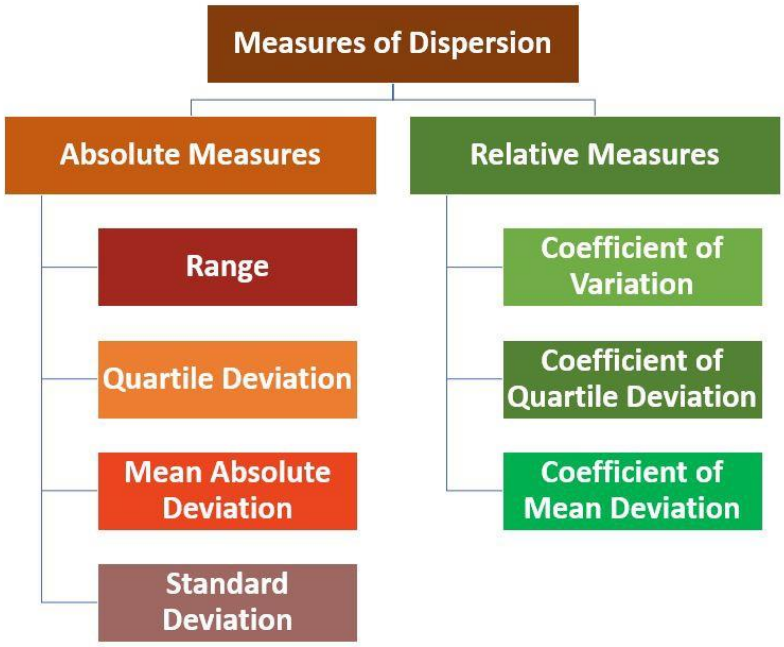
Geometric Mean

A geometric mean is a mean or average which shows the central tendency of a set of numbers by using the product of their values. For a set of n observations, a geometric mean is the nth root of their product. The geometric mean G.M., for a set of numbers $x_1, x_2, \ldots, x_n$ is given as

$$\mathrm{GM}(x_1, \ldots, x_n) = \sqrt[n]{|x_1 \times \cdots \times x_n|}$$

MEASURES OF DISPERSION

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.



Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measure of Dispersion

Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
2. **Variance:** Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance $(\sigma^2)=\sum(X-\mu)^2/N$
3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.
4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

1. Co-efficient of Range
2. Co-efficient of Variation
3. Co-efficient of Standard Deviation
4. Co-efficient of Quartile Deviation
5. Co-efficient of Mean Deviation

RANDOM VARIABLES

A random variable is a rule that assigns a numerical value to each outcome in a sample space. Random variables may be either discrete or continuous. A random variable is said to be discrete

if it assumes only specified values in an interval. Otherwise, it is continuous. We generally denote the random variables with capital letters such as X and Y.

As a function, a random variable is needed to be measured, which allows probabilities to be assigned to a set of potential values.

Types of Random Variable

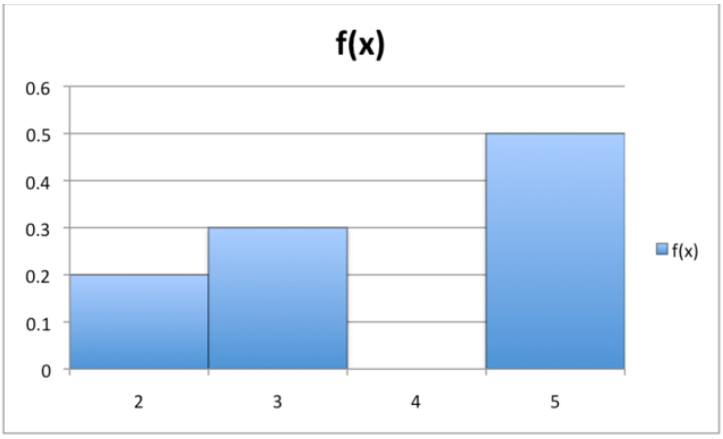
- Discrete Random Variable
- Continuous Random Variable

Discrete Random Variable

A discrete random variable can take only a finite number of distinct values such as 0, 1, 2, 3, 4, …

Examples of discrete random variables include:

- The number of eggs that a hen lays in a given day (it can't be 2.3)
- The number of people going to a given soccer match
- The number of students that come to class on a given day



Continuous Random Variables

Continuous random variables, on the other hand, take on values that vary continuously within one or more real intervals, and have a cumulative distribution function (CDF) that is absolutely continuous. As a result, the random variable has an uncountable infinite number of possible values

DISCRETE PROBABILITY DISTRIBUTIONS

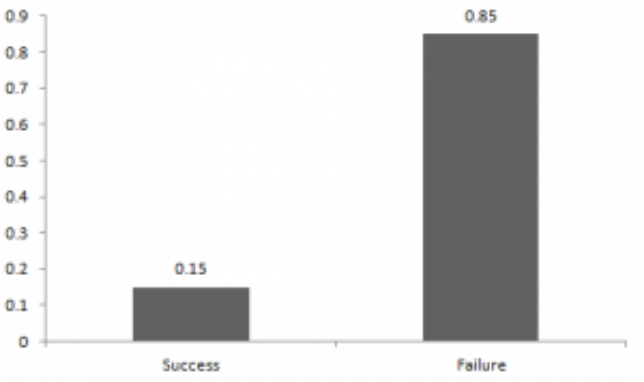
A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

The probability mass function is given by

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

The probabilities of success and failure need not be equally likely, like the result of a fight between me and hulk. He is pretty much certain to win. So in this case probability of my success is 0.15 while my failure is 0.85

Here, the probability of success(p) is not same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.



The expected value of a random variable X from a Bernoulli distribution is found as follows:

$$E(X) = 1*p + 0*(1-p) = p$$

The variance of a random variable from a bernoulli distribution is:

$$V(X) = E(X^2) – [E(X)]^2 = p – p^2 = p(1-p)$$

Binomial Distribution

Suppose that you won the toss today and this indicates a successful event. You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Assign a random variable, say X, to the number of times you won the toss. What can

be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

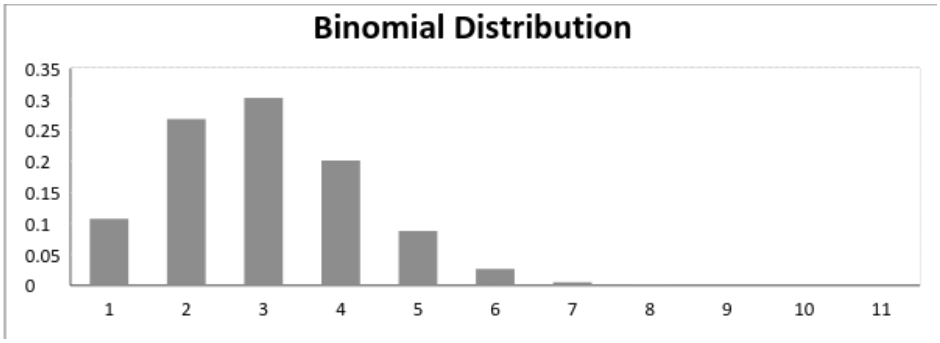
On the basis of the above explanation, the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Poisson Distribution

Suppose you work at a call center, approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

1. The number of emergency calls recorded at a hospital in a day.
2. The number of thefts reported in an area on a day.
3. The number of customers arriving at a salon in an hour.
4. The number of suicides reported in a particular city.
5. The number of printing errors at each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

- λ is the rate at which an event occurs,
- t is the length of a time interval,
- And X is the number of events in that time interval.

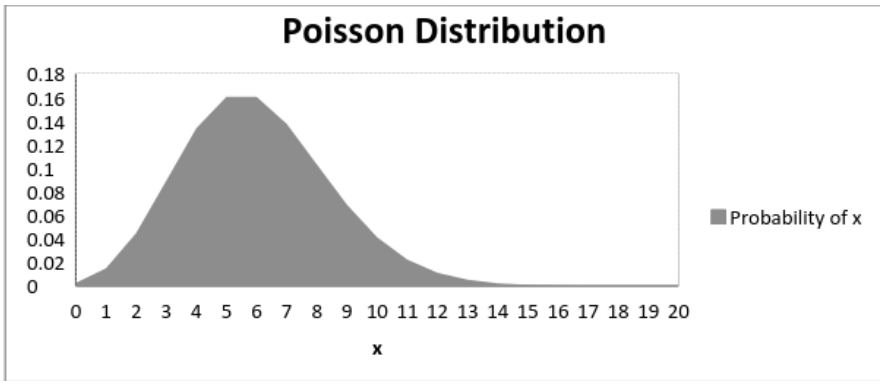
Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t.

The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x = 0, 1, 2, \ldots..$$

The mean μ is the parameter of this distribution. μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



CONTINUOUS PROBABILITY DISTRIBUTIONS

Exponential Distribution

Consider the call center example. What about the interval of time between the calls ? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

Other examples are:

1. Length of time beteeen metro arrivals,
2. Length of time between arrivals at a gas station
3. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an **exponential distribution** with PDF:

$$f(x) = \{ \lambda e^{-\lambda x}, \ x \geq 0$$

and parameter **$\lambda > 0$** which is also called the rate.

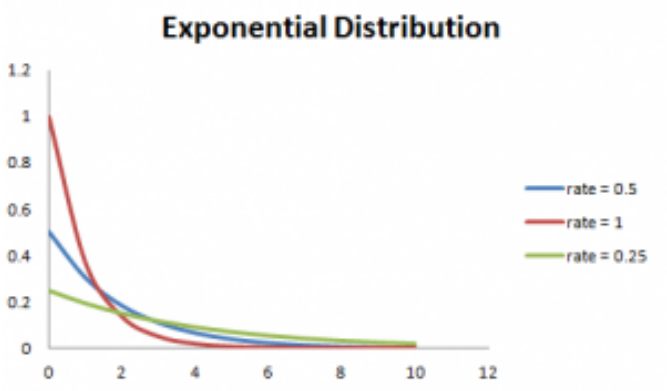
For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

$$\text{Mean} \rightarrow E(X) = 1/\lambda$$

$$\text{Variance} \rightarrow Var(X) = (1/\lambda)^2$$

Also, the greater the rate, the faster the curve drops and the lower the rate, flatter the curve. This is explained better with the graph shown below.



CHI-SQUARE DISTRIBUTION

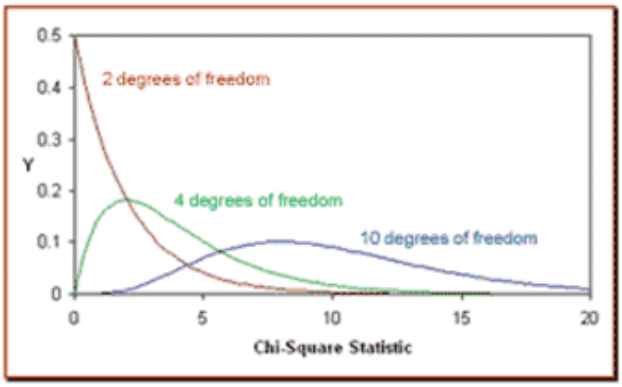
A random variable χ follows chi-square distribution ,it can be written as a sum of squared

standard normal variables.

$$\chi^2 = \sum Z_i^2$$

Degrees of freedom:

Degrees of freedom refers to the maximum number of logically independent values, which have the freedom to vary. In simple words, it can be defined as the total number of observations minus the number of independent constraints imposed on the observations.



In the above figure, we could see Chi-Square distribution for different degrees of freedom. We can also observe that as the degrees of freedom increase Chi-Square distribution approximates to normal distribution.

Chi-Square Test for Feature Selection

A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c = \text{degrees of freedom}$

$O = \text{observed value(s)}$

$E = \text{expected value(s)}$

Let's consider a scenario where we need to determine the relationship between the independent category feature (predictor) and dependent category feature(response). In feature selection, we aim to select the features which are highly dependent on the response.

When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training.

Steps to perform the Chi-Square Test:

1. Define Hypothesis.
2. Find the expected values.
3. Calculate the Chi-Square statistic.
4. Accept or Reject the Null Hypothesis.

Normal Distribution

Normal distribution represents the behavior of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is given by:

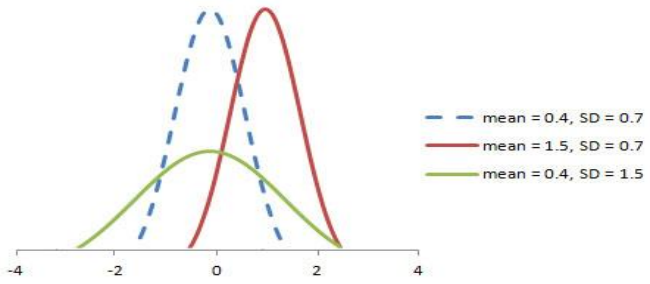
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-u}{\sigma})^2\}} \qquad for -\infty < x < \infty.$$

The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean -> E(X) = μ

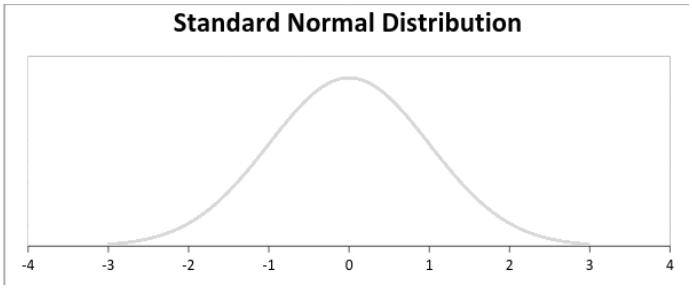
Variance -> Var(X) = σ^2

Here, μ (mean) and σ (standard deviation) are the parameters.
The graph of a random variable X ~ N (μ, σ) is shown below.



A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1. For such a case, the PDF becomes:

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad for -\infty < x < \infty$$



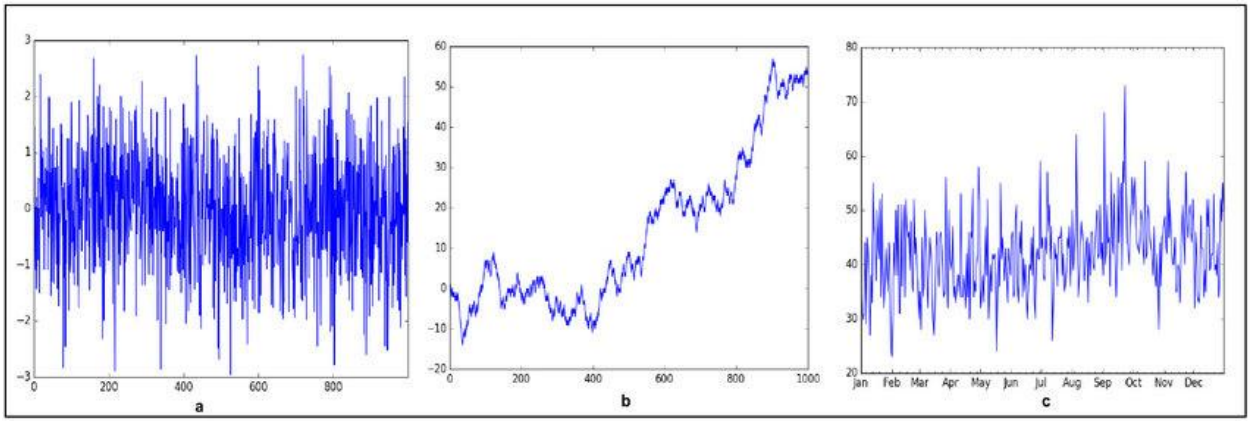
WHITE-NOISE PROCESS

White noise is an important concept in time series forecasting.

If a time series is white noise, it is a sequence of random numbers and cannot be predicted. If the series of forecast errors are not white noise, it suggests improvements could be made to the predictive model.

It is important for two main reasons:

1. **Predictability**: If your time series is white noise, then, by definition, it is random. You cannot reasonably model it and make predictions.
2. **Model Diagnostics**: The series of errors from a time series forecast model should ideally be white noise.



VARIANCE

In statistics, variance refers to the spread of a data set. It's a measurement used to identify how far each number in the data set is from the mean.

While performing market research, variance is particularly useful when calculating probabilities of future events. Variance is a great way to find all of the possible values and likelihoods that a random variable can take within a given range.

A variance value of zero represents that all of the values within a data set are identical, while all variances that are not equal to zero will come in the form of positive numbers.

The larger the variance, the more spread in the data set.

A large variance means that the numbers in a set are far from the mean and each other. A small variance means that the numbers are closer together in value.

How to Calculate Variance

Variance is calculated by taking the differences between each number in a data set and the mean, squaring those differences to give them positive value, and dividing the sum of the resulting squares by the number of values in the set.

The formula for variance is as follows:

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

In this formula, X represents an individual data point, u represents the mean of the data points, and N represents the total number of data points.

Note that while calculating a sample variance in order to estimate a population variance, the denominator of the variance equation becomes N – 1. This removes bias from the estimation, as it prohibits the researcher from underestimating the population variance.

An Advantage of Variance

One of the primary advantages of variance is that it treats all deviations from the mean of the data set in the same way, regardless of direction.

This ensures that the squared deviations cannot sum to zero, which would result in giving the appearance that there was no variability in the data set at all.

CORRELATION COEFFICIENT

The correlation coefficient is the term used to refer to the resulting correlation measurement. It will always maintain a value between one and negative one.

When the correlation coefficient is one, the variables under examination have a perfect positive correlation. In other words, when one moves, so does the other in the same direction, proportionally.

If the correlation coefficient is less than one, but still greater than zero, it indicates a less than perfect positive correlation. The closer the correlation coefficient gets to one, the stronger the correlation between the two variables.

When the correlation coefficient is zero, it means that there is no identifiable relationship between the variables. If one variable moves, it's impossible to make predictions about the movement of the other variable.

If the correlation coefficient is negative one, this means that the variables are perfectly negatively or inversely correlated. If one variable increases, the other will decrease at the same proportion. The variables will move in opposite directions from each other.

If the correlation coefficient is greater than negative one, it indicates that there is an imperfect negative correlation. As the correlation approaches negative one, the correlation grows.

COVARIANCE

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each

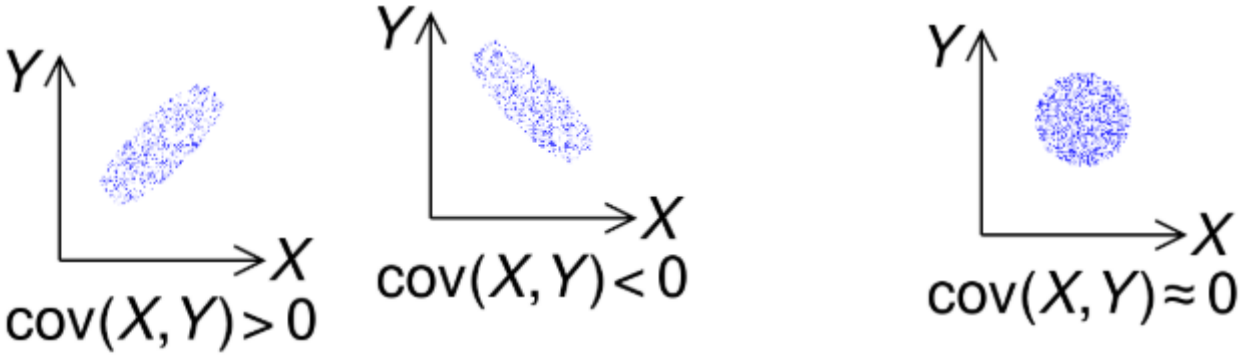
other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The values of covariance can be any number between the two opposite infinities. Also, it's important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:

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

The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables. Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude. Below are the plots which help us understand how the covariance between two variables would look in different directions.



CORRELATION

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables.

To do so we have to normalize the covariance by dividing it with the product of the standard deviations of the two variables, thus providing a correlation between the two variables.

The main result of a correlation is called the correlation coefficient.

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The correlation coefficient is a dimensionless metric and its value ranges from -1 to +1.

The closer it is to +1 or -1, the more closely the two variables are related.

If there is no relationship at all between two variables, then the correlation coefficient will certainly be 0. However, if it is 0 then we can only say that there is no linear relationship. There could exist other functional relationships between the variables.

When the correlation coefficient is positive, an increase in one variable also increases the other. When the correlation coefficient is negative, the changes in the two variables are in opposite directions.

HYPOTHESIS AND INFERENCE

A statistical hypothesis is an assumption about a population which may or may not be true. Hypothesis testing is a set of formal procedures used by statisticians to either accept or reject statistical hypotheses. Statistical hypotheses are of two types:

- **Null hypothesis,** H0- represents a hypothesis of chance basis.
- **Alternative hypothesis,** Ha - represents a hypothesis of observations which are influenced by some non-random cause.

Example

suppose we wanted to check whether a coin was fair and balanced. A null hypothesis might say, that half flips will be of head and half will of tails whereas alternative hypothesis might say that flips of head and tail may be very different.

H0: P=0.5

Ha: P≠0.5

For example if we flipped the coin 50 times, in which 40 Heads and 10 Tails results. Using result, we need to reject the null hypothesis and would conclude, based on the evidence, that the coin was probably not fair and balanced.

Hypothesis Tests

Following formal process is used by statistican to determine whether to reject a null hypothesis, based on sample data. This process is called hypothesis testing and is consists of following four steps:

1. **State the hypotheses** - This step involves stating both null and alternative hypotheses. The hypotheses should be stated in such a way that they are mutually exclusive. If one is true then other must be false.
2. **Formulate an analysis plan** - The analysis plan is to describe how to use the sample data to evaluate the null hypothesis. The evaluation process focuses around a single test statistic.
3. **Analyze sample data** - Find the value of the test statistic (using properties like mean score, proportion, t statistic, z-score, etc.) stated in the analysis plan.
4. **Interpret results** - Apply the decisions stated in the analysis plan. If the value of the test statistic is very unlikely based on the null hypothesis, then reject the null hypothesis.