Statistical Inference – I

Tushar B. Kute,
http://tusharkute.com

Contents

- Types of Statistical Inference, Descriptive Statistics, Inferential Statistics, Importance of Statistical Inference in Machine Learning.
- Descriptive Statistics, Measures of Central Tendency: Mean, Median, Mode, Mid-range,
- Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation.
- One sample hypothesis testing, Hypothesis, Testing of Hypothesis, Chi-Square Tests, t-test, ANOVA and ANOCOVA.
- Pearson Correlation, Bi-variate regression, Multi-variate regression, Chi-square statistics.

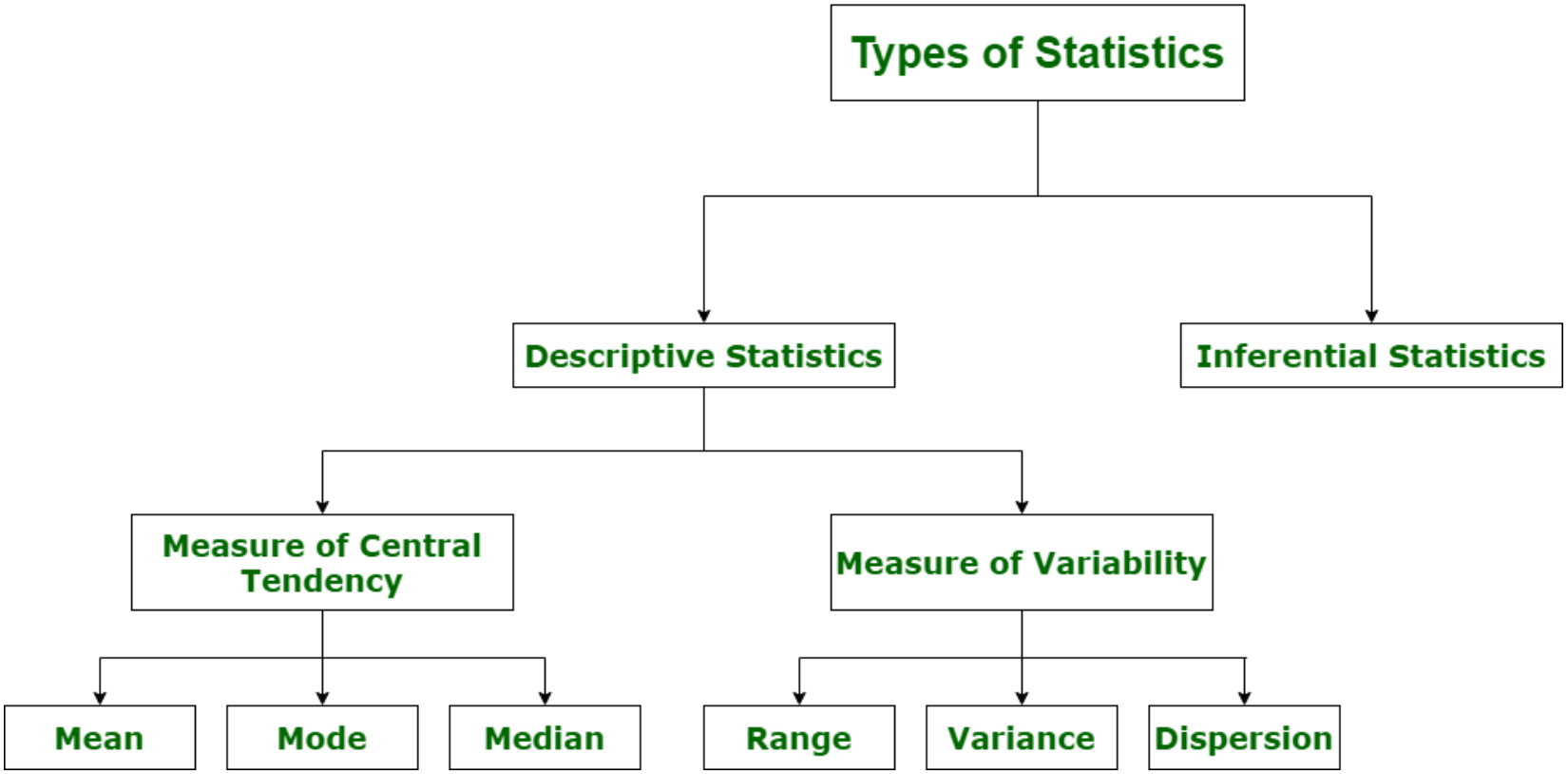
What is statistics?

- Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.
- Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments

What is statistics?

- When census data cannot be collected, statisticians collect data by developing specific experiment designs and survey samples.
- Representative sampling assures that inferences and conclusions can reasonably extend from the sample to the population as a whole.
- An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements.
- In contrast, an observational study does not involve experimental manipulation.

Types of statistics?



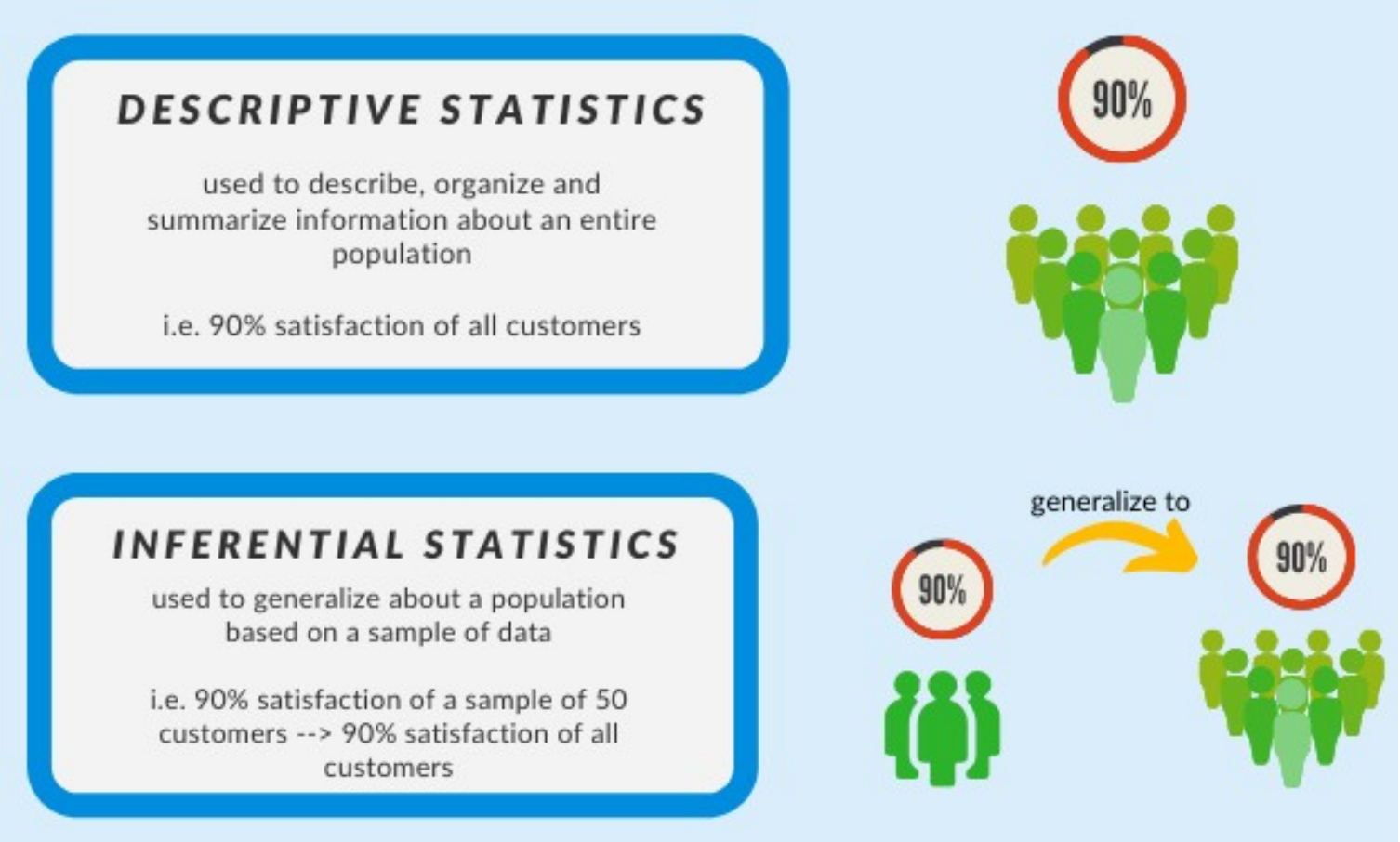
Descriptive Statistics

- A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics.
- Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Inferential Statistics

- Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.
- Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates.
- It is assumed that the observed data set is sampled from a larger population. Inferential statistics can be contrasted with descriptive statistics.
- Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Comparing


DESCRIPTIVE STATISTICS
used to describe, organize and summarize information about an entire population
i.e. 90% satisfaction of all customers
90%
INFERENTIAL STATISTICS
used to generalize about a population based on a sample of data
i.e. 90% satisfaction of a sample of 50 customers --> 90% satisfaction of all customers
90%
generalize to
90%

Why Statistics?

- What features are the most important?
- How should we design the experiment to develop our product strategy?
- What performance metrics should we measure?
- What is the most common and expected outcome?
- How do we differentiate between noise and valid data?

From data to knowledge

- In isolation, raw observations are just data. We use descriptive statistics to transform these observations into insights that make sense.
- Then we can use inferential statistics to study small samples of data and extrapolate our findings to the entire population.

Terminologies of statistics

- Population: It is an entire pool of data from where a statistical sample is extracted. It can be visualized as a complete data set of items that are similar in nature.
- Sample: It is a subset of the population, i.e. it is an integral part of the population that has been collected for analysis.
- Variable: A value whose characteristics such as quantity can be measured, it can also be addressed as a data point, or a data item.

Terminologies of statistics

- Distribution: The sample data that is spread over a specific range of values.
- Parameter: It is a value that is used to describe the attributes of a complete data set (also known as ‘population’). Example: Average, Percentage
- Quantitative analysis: It deals with specific characteristics of data- summarizing some part of data, such as its mean, variance, and so on.
- Qualitative analysis: This deals with generic information about the type of data, and how clean or structured it is.

Statistical Inference

"The use of a sample of data to draw inferences or conclusions about some aspect of the situation from which the data were taken."

Statistical Inference

- Statistical inference is the process of analysing the result and making conclusions from data subject to random variation.
- It is also called inferential statistics. Hypothesis testing and confidence intervals are the applications of the statistical inference.
- Statistical inference is a method of making decisions about the parameters of a population, based on random sampling.
- It helps to assess the relationship between the dependent and independent variables.
- The purpose of statistical inference to estimate the uncertainty or sample to sample variation.

Statistical Inference

- It allows us to provide a probable range of values for the true values of something in the population.
- The components used for making statistical inference are:
  - Sample Size
  - Variability in the sample
  - Size of the observed differences

Statistical Inference Procedure

- The procedure involved in inferential statistics are:
  - Begin with a theory
  - Create a research hypothesis
  - Operationalize the variables
  - Recognize the population to which the study results should apply
  - Formulate a null hypothesis for this population
  - Accumulate a sample from the population and continue the study
  - Conduct statistical tests to see if the collected sample properties are adequately different from what would be expected under the null hypothesis to be able to reject the null hypothesis

Statistical Inference Solution

- Statistical inference solutions produce efficient use of statistical data relating to groups of individuals or trials.
- It deals with all characters, including the collection, investigation and analysis of data and organizing the collected data.
- By statistical inference solution, people can acquire knowledge after starting their work in diverse fields. Some statistical inference solution facts are:
  - It is a common way to assume that the observed sample is of independent observations from a population type like Poisson or normal
  - Statistical inference solution is used to evaluate the parameter(s) of the expected model like normal mean or binomial proportion

Statistical Inference Solution

- An example of statistical inference is given below.

Suit	Spade	Clubs	Hearts	Diamonds
No.of times drawn	90	100	120	90

- Question: From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:
- While a card is tried at random, then what is the probability of getting a
  - Diamond cards
  - Black cards
  - Except for spade

Statistical Inference Solution

- By statistical inference solution,

  Total number of events = 400

  i.e.,90+100+120+90=400

- (1) The probability of getting diamond cards:
  - Number of trials in which diamond card is drawn = 90
  - Therefore, P(diamond card) = 90/400 = 0.225

Statistical Inference Solution

- (2) The probability of getting black cards:

  Number of trials in which black card showed up = 90+100 =190

  Therefore, P(black card) = 190/400 = 0.475

- (3) Except for spade

  Number of trials other than spade showed up = 90+100+120 =310

  Therefore, P(except spade) = 310/400 = 0.775

Statistical Machine Learning

- The methods used in statistics are important to train and test the data that is used as input to the machine learning model. Some of these include outlier/anomaly detection, sampling of data, data scaling, variable encoding, dealing with missing values, and so on.
- Statistics is also essential to evaluate the model that has been used, i.e. see how well the machine learning model performs on a test dataset, or on data that it has never seen before.
- Statistics is essential in selecting the final and appropriate model to deal with that specific data in a predictive modelling situation.
- It is also needed to show how well the model has performed, by taking various metrics and showing how the model has fared.

Discriptive Statistics

- Descriptive statistics summarize and organize characteristics of a data set.
- A data set is a collection of responses or observations from a sample or entire population.
- In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

Discriptive Statistics – Types

- There are 3 main types of descriptive statistics:
  - The distribution concerns the frequency of each value.
  - The central tendency concerns the averages of the values.
  - The variability or dispersion concerns how spread out the values are.
- You can apply these to assess only one variable at a time, in univariate analysis, or to compare two or more, in bivariate and multivariate analysis.

Discriptive Statistics – Example

- You want to study the popularity of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year:
  - Go to a library
  - Watch a movie at a theater
  - Visit a national park
- Your data set is the collection of responses to the survey.
- Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

Reference: https://www.scribbr.com

Measure of Central Tendency

- A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data.
- These one number summary is of three types.
  - Mean
  - Median
  - Mode

What is mean?

- Mean : Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations.
- This is also known as Average.
- Thus mean is a number around which the entire data set is spread.

Example:

- Consider the following data points.

  17, 16, 21, 18, 15, 17, 21, 19, 11, 23

  Mean — Mean is calculated as

$$Mean = \frac{17+16+21+18+15+17+21+19+11+23}{10} = \frac{178}{10} = 17.8$$

What is median?

- Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same.
- Median is calculated by first arranging the data in either ascending or descending order.
  - If the number of observations are odd, median is given by the middle observation in the sorted form.
  - If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.
- An important point to note that the order of the data (ascending or descending) does not effect the median.

What is median?

- To calculate Median, lets arrange the data in ascending order.

  11, 15, 16, 17, 17, 18, 19, 21, 21, 23

- Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17 + 18}{2} = 17.5$$

What is mode?

- Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.
  - If there is only one number that appears maximum number of times, the data has one mode, and is called Uni-modal.
  - If there are two numbers that appear maximum number of times, the data has two modes, and is called Bi-modal.
  - If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called Multi-modal.

What is mode?

- The data:

  11, 15, 16, 17, 17, 18, 19, 21, 21, 23

- Mode is given by the number that occurs maximum number of times.
- Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

Note:

- Since Median and Mode does not take all the data points for calculations, these are robust to outliers, i.e. these are not effected by outliers.
- At the same time, Mean shifts towards the outlier as it considers all the data points. This means if the outlier is big, mean overestimates the data and if it is small, the data is underestimated.
- If the distribution is symmetrical, Mean = Median = Mode. Normal distribution is an example.

Mid-range

- The midrange is a type of average, or mean. Electronic gadgets are sometimes classified as “midrange”, meaning they’re in the middle-price bracket.

  The formula to find the midrange = (high + low) / 2.
- Sample problem: Current cell phone prices in a mobile phone store range from ₹40 (the cheapest) to ₹550 (the most expensive). Find the midrange.

  Step 1: Add the lowest value to the highest: ₹550 + ₹40 = ₹590.

  Step 2: Divide Step 1 by two: ₹590 / 2 = ₹295.
- The mid priced phones would be priced at around ₹295.

Dispersion

- Dispersion refers to measures of how spread out our data is.
- Typically they’re statistics for which values near zero signify not spread out at all and for which large values (whatever that means) signify very spread out.

Dispersion - Types

- Absolute Deviation from Mean
- Variance
- Standard Deviation
- Range
- Quartiles
- Skewness
- Kurtosis

Reference: https://towardsdatascience.com

Mean Absolute Deviation

- The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set.
- It is calculated as,

$$Mean\ \ Absolute\ \ Deviation\ \ = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|$$

Variance

- In statistics, the variance is a measure of how far individual (numeric) values in a dataset are from the mean or average value.
- The variance is often used to quantify spread or dispersion. Spread is a characteristic of a sample or population that describes how much variability there is in it.
- A high variance tells us that the values in our dataset are far from their mean. So, our data will have high levels of variability.
- On the other hand, a low variance tells us that the values are quite close to the mean. In this case, the data will have low levels of variability.

Variance

- To calculate the variance in a dataset, we first need to find the difference between each individual value and the mean. The variance is the average of the squares of those differences. We can express the variance with the following math expression:

$$\sigma^2 = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \mu)^2$$

- In this equation, xi stands for individual values or observations in a dataset. μ stands for the mean or average of those values. n is the number of values in the dataset.
- The term xi - μ is called the deviation from the mean. So, the variance is the mean of square deviations. That's why we denoted it as σ^2.

Variance

Say we have a dataset [3, 5, 2, 7, 1, 3]. To find its variance, we need to calculate the mean which is:

$$(3+5+2+7+1+3)/6 = 3.5$$

Then, we need to calculate the sum of the square deviation from the mean of all the observations. Here's how:

$$(3-3.5)^2 + (5-3.5)^2 + (2-3.5)^2 + (7-3.5)^2 + (1-3.5)^2 + (3-3.5)^2 = 23.5$$

To find the variance, we just need to divide this result by the number of observations like this:

$$23.5/6 = 3.916666667$$

Variance

- That's all. The variance of our data is 3.916666667. The variance is difficult to understand and interpret, particularly how strange its units are.
- For example, if the observations in our dataset are measured in pounds, then the variance will be measured in square pounds.
- So, we can say that the observations are, on average, 3.916666667 square pounds far from the mean 3.5.
- Fortunately, the standard deviation comes to fix this problem

Standard Deviation

- The standard deviation measures the amount of variation or dispersion of a set of numeric values.
- Standard deviation is the square root of variance σ2 and is denoted as σ.
- So, if we want to calculate the standard deviation, then all we just have to do is to take the square root of the variance as follows:

$$Std\ \ Deviation\ \ = \sqrt{Variance\ \ } = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

Standard Deviation

- Again, we need to distinguish between the population standard deviation, which is the square root of the population variance (σ2) and the sample standard deviation, which is the square root of the sample variance (S2).
- We'll denote the sample standard deviation as S:

$$S=\sqrt{S2}$$

Standard Deviation

- There are six steps for finding the standard deviation:
  - List each score and find their mean.
  - Subtract the mean from each score to get the deviation from the mean.
  - Square each of these deviations.
  - Add up all of the squared deviations.
  - Divide the sum of the squared deviations by N – 1.
  - Find the square root of the number you found.

Standard Deviation

- Low values of standard deviation tell us that individual values are closer to the mean.
- High values, on the other hand, tell us that individual observations are far away from the mean of the data.
- Values that are within one standard deviation of the mean can be thought of as fairly typical, whereas values that are three or more standard deviations away from the mean can be considered much more atypical. They're also known as outliers.

Standard Deviation

If we're trying to estimate the standard deviation of the population using a sample of data, then we'll be better served using **n - 1** degrees of freedom. Here's a math expression that we typically use to estimate the population variance:

$$\sigma_x = \sqrt{\frac{\sum_{i=0}^{n-1} (x_i - \mu_x)^2}{n-1}}$$

Note that this is the square root of the sample variance with **n - 1** degrees of freedom. This is equivalent to say:

$$S_{n-1} = \sqrt{S_{n-1}^2}$$

Mean deviation

- The mean deviation is defined as a statistical measure which is used to calculate the average deviation from the mean value of the given data set.
- The mean deviation of the data values can be easily calculated using the below procedure.
  - Step 1: Find the mean value for the given data values
  - Step 2: Now, subtract mean value form each of the data value given (Note: Ignore the minus symbol)
  - Step 3: Now, find the mean of those values obtained in step 2.

Mean deviation

- The formula to calculate the mean deviation for the given data set is given below.
- Mean Deviation = $[\Sigma |X - \mu|]/N$
- Here,
  - Σ represents the addition of values
  - X represents each value in the data set
  - M represents the mean value of the data set
  - N represents the number of data values
  - || represents the absolute value, which ignores the “-” symbol

Example

- Determine the mean deviation for the data values 5, 3,7, 8, 4, 9.
- Solution:

  Given data values are 5, 3, 7, 8, 4, 9.

  We know that the procedure to calculate the mean deviation.
- First, find the mean for the given data:

  Mean, $\mu = (5+3+7+8+4+9)/6$

  $\mu = 36/6$

  $\mu = 6$

  Therefore, the mean value is 6.

Example

- Now, subtract each mean from the data value, and ignore the minus symbol if any (Ignore"-")

  5 – 6 = 1, 3 – 6 = 3, 7 – 6 = 1, 8 – 6 = 2, 4 – 6 = 2 9 – 6 = 3
- Now, the obtained data set is 1, 3, 1, 2, 2, 3.

  Finally, find the mean value for the obtained data set

  Therefore, the mean deviation is

  = (1+3 + 1+ 2+ 2+3) /6

  = 12/6

  = 2
- Hence, the mean deviation for 5, 3, 7, 8, 4, 9 is 2.

Range

- Range is the difference between the Maximum value and the Minimum value in the data set.
- It is given as,

$$\text{range} = \text{maximum} - \text{minimum}$$

What is Hypothesis ?

- A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation. For example:
  - A new medicine you think might work.
  - A way of teaching you think might be better.
  - A possible location of new species.
  - A fairer way to administer standardized tests.
- It can really be anything at all as long as you can put it to the test.

What is Hypothesis Statement?

- If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:
- "If I…(do this to an independent variable)….then (this will happen to the dependent variable)."
- For example:
  - If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
  - If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
  - If I (give exams at noon instead of 7) then (student test scores will improve).
  - If I (look in this certain location) then (I am more likely to find new species).

Good Hypothesis

- A good hypothesis statement should:
  - Include an “if” and “then” statement (according to the University of California).
  - Include both the independent and dependent variables.
  - Be testable by experiment, survey or other scientifically sound technique.
  - Be based on information in prior research (either yours or someone else’s).
  - Have design criteria (for engineering or programming projects).

What is Hypothesis Testing?

$$z = \frac{\hat{p} - p}{\sqrt{p\,q/n}}$$

- Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.
- You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance.
- If your results may have happened by chance, the experiment won't be repeatable and so has little use.

What is Hypothesis Testing?

Hypothesis testing can be one of the most confusing aspects for students, mostly because before you can even perform a test, you have to know what your null hypothesis is. Often, those tricky word problems that you are faced with can be difficult to decipher. But it’s easier than you think; all you need to do is:

- Figure out your null hypothesis,
- State your null hypothesis,
- Choose what kind of test you need to perform,
- Either support or reject the null hypothesis.

Null Hypothesis

- If you trace back the history of science, the null hypothesis is always the accepted fact. Simple examples of null hypotheses that are generally accepted as being true are:
  - DNA is shaped like a double helix.
  - There are 8 planets in the solar system (excluding Pluto).
  - Taking Vioxx can increase your risk of heart problems (a drug now taken off the market).

How to state Null Hypothesis?

- You won't be required to actually perform a real experiment or survey in elementary statistics (or even disprove a fact like "Pluto is a planet"!), so you'll be given word problems from real-life situations.
- You'll need to figure out what your hypothesis is from the problem. This can be a little trickier than just figuring out what the accepted fact is.
- With word problems, you are looking to find a fact that is nullifiable (i.e. something you can reject).

Hypothesis Testing Example

- *A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.*
- The hypothesis statement in this question is that the researcher believes the average recovery time is more than 8.2 weeks. It can be written in mathematical terms as:

$$H1: \mu > 8.2$$

- Next, you'll need to state the null hypothesis (See: How to state the null hypothesis). That's what will happen if the researcher is wrong. In the above example, if the researcher is wrong then the recovery time is less than or equal to 8.2 weeks. In math, that's:

$$H0\ \mu \leq 8.2$$

Making it practically

- In business, you are most likely constantly running experiments, trying to improve something.
- For instance, implementing new processes, running marketing campaigns, taking a poll, conducting surveys as well as many others.
- Now, you implemented a change and are capturing data from which you analyze the before and after.
- How can you tell the difference is not just because of chance? Enter the importance of statistical significance!

T Test

- There are multiple statistical hypothesis tests out there. Each test aims to find if there is a difference in one of many statistical properties.
- Statistical properties include the standard deviation, average or variance for example.
- The T-Test is used to determine if the mean (average) of two groups are truly different.
- It is also called the Student's T-Test. Not because it's used in college! But rather, because its inventor, William Sealy Gosset, used the pseudonym Student.

When to use T Test?

- You use the T-Test when you will be comparing the means of two samples. If you have more than 2 samples, you will have to run a pairwise T-Test on all samples or use another statistical hypothesis method called Anova.
- When you don’t know the population’s mean and standard deviation. In the T-Test, you are comparing 2 samples of an unknown population.
- A sample is a randomly chosen set of data points from a population. If you do know the population’s mean and standard deviation, you would run a Z-Test instead.

When to use T Test?

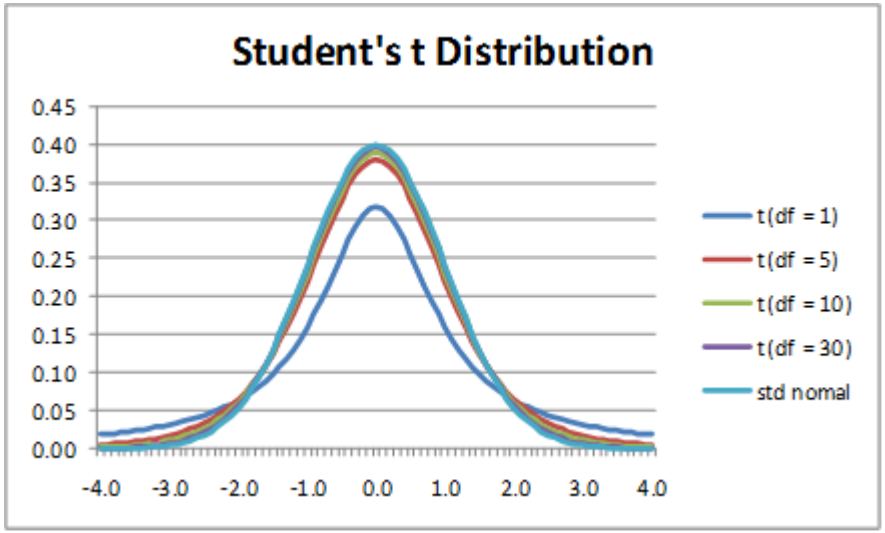
- When you have a small number of samples. The T-Test is commonly used when you have less than 30 samples in each of the groups you are running the T-Test on.
- If you have less than 30 samples in each of the groups, you run the T-Test if you can assume the population follows a normal distribution.
- As mentioned previously, the T-Test is commonly used on smaller sample sizes. You use the T-Test if the samples follow a normal distribution. Why is this allowed? You can thank the Central Limit Theorem for this.

Types of T Tests

- Independent Sample T-Test.
  - In this type of test, you are comparing the average of two independent unrelated groups. Meaning, you are comparing samples from two different populations and are testing whether or not they have a different average.
- Paired Sample T-Test.
  - In this test, you compare the average of two samples taken from the same population but at different points in time. A simple example would be when you would like to test the means of before and after observations taken from the same target.
- One-Sample T-Test
  - Here we test if the average of a single group is different from a known average or hypothesized average.

T Tests

- The t test (also called Student's T Test) compares two averages (means) and tells you if they are different from each other.
- The t test also tells you how significant the differences are; In other words it lets you know if those differences could have happened by chance.



Example:

- Let’s say you have a cold and you try a naturalistic remedy. Your cold lasts a couple of days.
- The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week.
- You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy.
- What you really want to know is, are these results repeatable?
- A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

Another Example:

- Student's T-tests can be used in real life to compare means. For example, a drug company may want to test a new cancer drug to find out if it improves life expectancy.
- In an experiment, there's always a control group (a group who are given a placebo, or "sugar pill").
- The control group may show an average life expectancy of +5 years, while the group taking the new drug might have a life expectancy of +6 years.
- It would seem that the drug might work. But it could be due to a fluke.
- To test this, researchers would use a Student's t-test to find out if the results are repeatable for an entire population.

What is T-score?

- The t score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups.
- The smaller the t score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different from each other as they are within each other.
- When you run a t test, the bigger the t-value, the more likely it is that the results are repeatable.
  - A large t-score tells you that the groups are different.
  - A small t-score tells you that the groups are similar.

T-value and P-value

- How big is “big enough”? Every t-value has a p-value to go with it. A p-value is the probability that the results from your sample data occurred by chance.
- P-values are from 0% to 100%. They are usually written as a decimal.
- For example, a p value of 5% is 0.05. Low p-values are good; They indicate your data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance.
- In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

How to perform 2-sample test?

- Lets us say we have to test whether the height of men in the population is different from height of women in general. So we take a sample from the population and use the t-test to see if the result is significant.
  - Determine a null and alternate hypothesis.
  - Collect sample data
  - Determine a confidence interval and degrees of freedom
  - Calculate the t-statistic
  - Calculate the critical t-value from the t distribution
  - Compare the critical t-values with the calculated t statistic

Step-1

- Determine a null and alternate hypothesis.
- In general, the null hypothesis will state that the two populations being tested have no statistically significant difference.
- The alternate hypothesis will state that there is one present. In this example we can say that:

Null Hypothesis : Height of men & women are the same

Alternate Hypothesis : Height of men & women are the different

Step-2

- Collect sample data
- Next step is to collect data for each population group.
- In our example we will collect 2 sets of data, one with the height of women and one with the height of men.
- The sample size should ideally be the same but it can be different.
- Lets say that the sample sizes are nx and ny.

Step-3

- Determine a confidence interval and degrees of freedom
- This is what we call alpha (α). The typical value of α is 0.05.
- This means that there is 95% confidence that the conclusion of this test will be valid.
- The degree of freedom can be calculated by the the following formula:

$$df = n_x + n_y - 2$$

Step-4

- Calculate the t-statistic
- t-statistic can be calculated using the below formula:

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

M = mean
n = number of scores per group

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

x = individual scores
M = mean
n= number of scores in group

- where, Mx and My are the mean values of the two samples of male and female.
- Nx and Ny are the sample space of the two samples S is the standard deviation

Step-5

- Calculate the critical t-value from the t distribution
- To calculate the critical t-value, we need 2 things, the chosen value of alpha and the degrees of freedom.
- The formula of critical t-value is complex but it is fixed for a fixed pair of degree of freedom and value of alpha.
- We therefore use a table to calculate the critical t-value:

Step-5

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Step-6

- Compare the critical t-values with the calculated t statistic
- If the calculated t-statistic is greater than the critical t-value, the test concludes that there is a statistically significant difference between the two populations. Therefore, you reject the null hypothesis that there is no statistically significant difference between the two populations.
- In any other case, there is no statistically significant difference between the two populations. The test fails to reject the null hypothesis and we accept the alternate hypothesis which says that the height of men and women are statistically different.

Categorical Variable

- Categorical variables fall into a particular category of those variables that can be divided into finite categories.
- These categories are generally names or labels.
- These variables are also called qualitative variables as they depict the quality or characteristics of that particular variable.
- For example, the category “Movie Genre” in a list of movies could contain the categorical variables – “Action”, “Fantasy”, “Comedy”, “Romance”, etc.

Categorical Variable

- There are broadly two types of categorical variables:
  - Nominal Variable: A nominal variable has no natural ordering to its categories. They have two or more categories. For example, Marital Status (Single, Married, Divorcee); Gender (Male, Female, Transgender), etc.
  - Ordinal Variable: A variable for which the categories can be placed in an order. For example, Customer Satisfaction (Excellent, Very Good, Good, Average, Bad), and so on
- When the data we want to analyze contains this type of variable, we turn to the chi-square test, denoted by x^2, to test our hypothesis.

Chi-Square Test

- What is the Chi-square goodness of fit test?
  - The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.
- When can I use the test?
  - You can use the test when you have counts of values for a categorical variable.
- Is this test the same as Pearson's Chi-square test?
  - Yes.

Example:

- Let’s learn the use of chi-square with an intuitive example.
- A research scholar is interested in the relationship between the placement of students in the statistics department of a reputed University and their C.G.P.A (their final assessment score).
- He obtains the placement records of the past five years from the placement cell database (at random).
- He records how many students who got placed fell into each of the following C.G.P.A. categories – 9-10, 8-9, 7-8, 6-7, and below 6.

Example:

- If there is no relationship between the placement rate and the C.G.P.A., then the placed students should be equally spread across the different C.G.P.A. categories (i.e. there should be similar numbers of placed students in each category).
- However, if students having C.G.P.A more than 8 are more likely to get placed, then there would be a large number of placed students in the higher C.G.P.A. categories as compared to the lower C.G.P.A. categories. In this case, the data collected would make up the observed frequencies.
- So the question is, are these frequencies being observed by chance or do they follow some pattern?
- Here enters the chi-square test! The chi-square test helps us answer the above question by comparing the observed frequencies to the frequencies that we might expect to obtain purely by chance.

Assumptions of test

- The x^2 assumes that the data for the study is obtained through random selection, i.e. they are randomly picked from the population
- The categories are mutually exclusive i.e. each subject fits in only one category. For e.g.- from our above example – the number of people who lunched in your restaurant on Monday can't be filled in the Tuesday category
- The data should be in the form of frequencies or counts of a particular category and not in percentages
- The data should not consist of paired samples or groups or we can say the observations should be independent of each other
- When more than 20% of the expected frequencies have a value of less than 5 then Chi-square cannot be used. To tackle this problem: Either one should combine the categories only if it is relevant or obtain more data

Example

- This is a non-parametric test. We typically use it to find how the observed value of a given event is significantly different from the expected value. In this case, we have categorical data for one independent variable, and we want to check whether the distribution of the data is similar or different from that of the expected distribution.
- Let’s consider the above example where the research scholar was interested in the relationship between the placement of students in the statistics department of a reputed University and their C.G.P.A.
- In this case, the independent variable is C.G.P.A with the categories 9-10, 8-9, 7-8, 6-7, and below 6.

Example

- In this case, the independent variable is C.G.P.A with the categories 9-10, 8-9, 7-8, 6-7, and below 6.
- The statistical question here is: whether or not the observed frequencies of placed students are equally distributed for different C.G.P.A categories (so that our theoretical frequency distribution contains the same number of students in each of the C.G.P.A categories).
- We will arrange this data by using the contingency table which will consist of both the observed and expected values as below:

Example

	C.G.P.A					
	10-9	9-8	8-7	7-6	Below 6	Total
Observed Frequency of Placed students	30	35	20	10	5	100
Expected Frequency of Placed students	20	20	20	20	20	100

Example

- After constructing the contingency table, the next task is to compute the value of the chi-square statistic. The formula for chi-square is given as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where,

x^2 = Chi-Square value

O_i = Observed frequency

E_i = Expected frequency

Steps

- Step 1: Subtract each expected frequency from the related observed frequency. For example, for the C.G.P.A category 10-9, it will be “30-20 = 10”. Apply similar operation for all the categories
- Step 2: Square each value obtained in step 1, i.e. $(O-E)^2$. For example: for the C.G.P.A category 10-9, the value obtained in step 1 is 10. It becomes 100 on squaring. Apply similar operation for all the categories
- Step 3: Divide all the values obtained in step 2 by the related expected frequencies i.e. $(O-E)^2/E$. For example: for the C.G.P.A category 10-9, the value obtained in step 2 is 100. On dividing it with the related expected frequency which is 20, it becomes 5. Apply similar operation for all the categories
- Step 4: Add all the values obtained in step 3 to get the chi-square value. In this case, the chi-square value comes out to be 32.5
- Step 5: Once we have calculated the chi-square value, the next task is to compare it with the critical chi-square value. We can find this in the below chi-square table against the degrees of freedom (number of categories – 1) and the level of significance:

Example

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Example

- In this case, the degrees of freedom are 5-1 = 4. So, the critical value at 5% level of significance is 9.49.
- Our obtained value of 32.5 is much larger than the critical value of 9.49. Therefore, we can say that the observed frequencies are significantly different from the expected frequencies.
- In other words, C.G.P.A is related to the number of placements that occur in the department of statistics.

ANOVA Test

- ANOVA, which stands for Analysis of Variance, is a statistical test used to analyze the difference between the means of more than two groups.
- A one-way ANOVA uses one independent variable, while a two-way ANOVA uses two independent variables.

When to use?

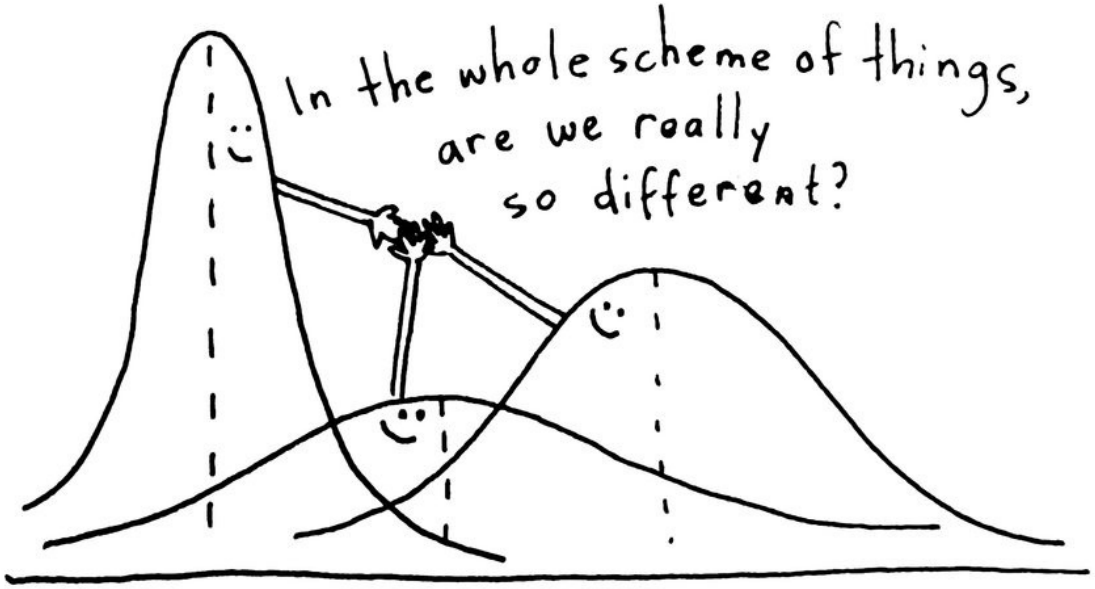
- Use a one-way ANOVA when you have collected data about one categorical independent variable and one quantitative dependent variable. The independent variable should have at least three levels (i.e. at least three different groups or categories).
- ANOVA tells you if the dependent variable changes according to the level of the independent variable. For example:
  - Your independent variable is social media use, and you assign groups to low, medium, and high levels of social media use to find out if there is a difference in hours of sleep per night.
  - Your independent variable is brand of soda, and you collect data on Coke, Pepsi, Sprite, and Fanta to find out if there is a difference in the price per 100ml.
  - You independent variable is type of fertilizer, and you treat crop fields with mixtures 1, 2 and 3 to find out if there is a difference in crop yield.

When to use?

- The null hypothesis (H_0) of ANOVA is that there is no difference among group means.
- The alternate hypothesis (H_a) is that at least one group differs significantly from the overall mean of the dependent variable.
- If you only want to compare two groups, use a t-test instead.

ANOVA

- We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not.



How it works?

- ANOVA determines whether the groups created by the levels of the independent variable are statistically different by calculating whether the means of the treatment levels are different from the overall mean of the dependent variable.
- If any of the group means is significantly different from the overall mean, then the null hypothesis is rejected.

How it works?

- ANOVA uses the F-test for statistical significance. This allows for comparison of multiple means at once, because the error is calculated for the whole set of comparisons rather than for each individual two-way comparison (which would happen with a t-test).
- The F-test compares the variance in each group mean from the overall group variance.
- If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.

Assumptions of ANOVA

- The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:
  - Independence of observations: the data were collected using statistically-valid methods, and there are no hidden relationships among observations. If your data fail to meet this assumption because you have a confounding variable that you need to control for statistically, use an ANOVA with blocking variables.
  - Normally-distributed response variable: The values of the dependent variable follow a normal distribution.
  - Homogeneity of variance: The variation within each group being compared is similar for every group. If the variances are different among the groups, then ANOVA probably isn't the right fit for the data.

ANCOVA Test

- ANCOVA is a blend of analysis of variance (ANOVA) and regression.
- It is similar to factorial ANOVA, in that it can tell you what additional information you can get by considering one independent variable (factor) at a time, without the influence of the others. It can be used as:
  - An extension of multiple regression to compare multiple regression lines,
  - An extension of analysis of variance.

ANCOVA Test

- ANCOVA can explain within-group variance. It takes the unexplained variances from the ANOVA test and tries to explain them with confounding variables (or other covariates).
- You can use multiple possible covariates. However, more you enter, the fewer degrees of freedom you'll have. Entering a weak covariate isn't a good idea as it will reduce the statistical power.
- The lower the power, the less likely you'll be able to rely on the results from your test.
- Strong covariates have the opposite effect: it can increase the power of your test.

Steps of ANCOVA

- General steps are:
  - Run a regression between the independent and dependent variables.
  - Identify the residual values from the results.
  - Run an ANOVA on the residuals.

Assumptions of ANCOVA

- Assumptions are basically the same as the ANOVA assumptions. Check that the following are true before running the test:
  - Independent variables (minimum of two) should be categorical variables.
  - The dependent variable and covariate should be continuous variables (measured on an interval scale or ratio scale.)
  - Make sure observations are independent. In other words, don’t put people into more than one group.

What is correlation ?

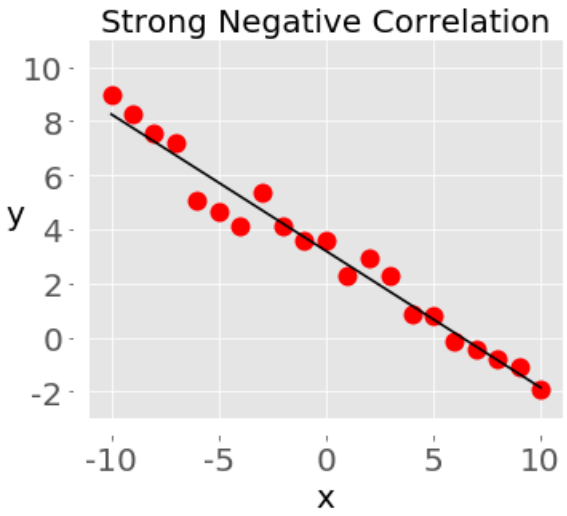
- Statistics and data science are often concerned about the relationships between two or more variables (or features) of a dataset. Each data point in the dataset is an observation, and the features are the properties or attributes of those observations.
- Every dataset you work with uses variables and observations. For example, you might be interested in understanding the following:
  - How the height of basketball players is correlated to their shooting accuracy
  - Whether there's a relationship between employee work experience and salary
  - What mathematical dependence exists between the population density and the gross domestic product of different countries

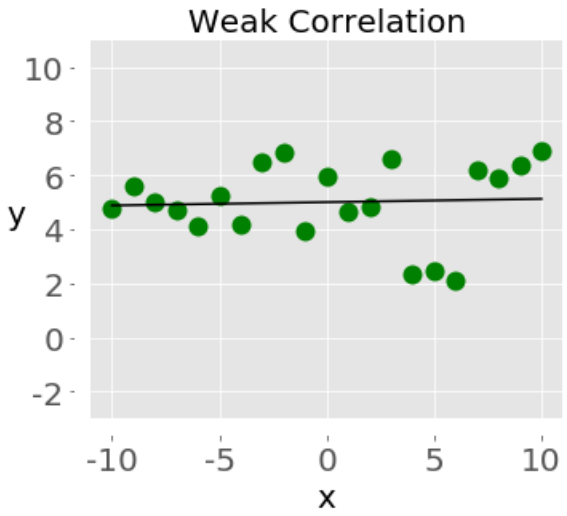
What is correlation ?

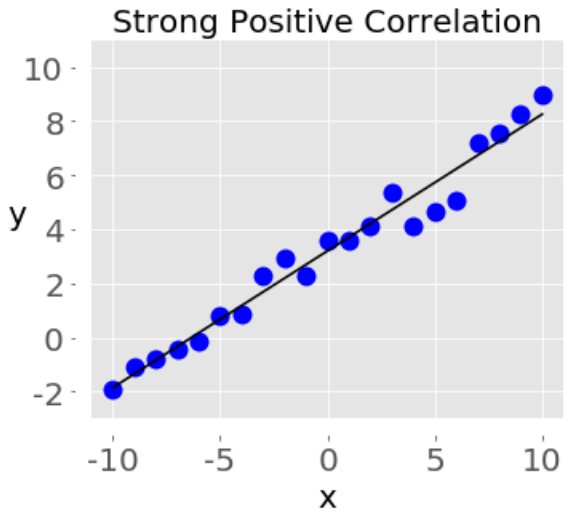
Name	Years of Experience	Annual Salary
Ann	30	120,000
Rob	21	105,000
Tom	19	90,000
Ivy	10	82,000

- In this table, each row represents one observation, or the data about one employee (either Ann, Rob, Tom, or Ivy). Each column shows one property or feature (name, experience, or salary) for all the employees.

Forms of correlation



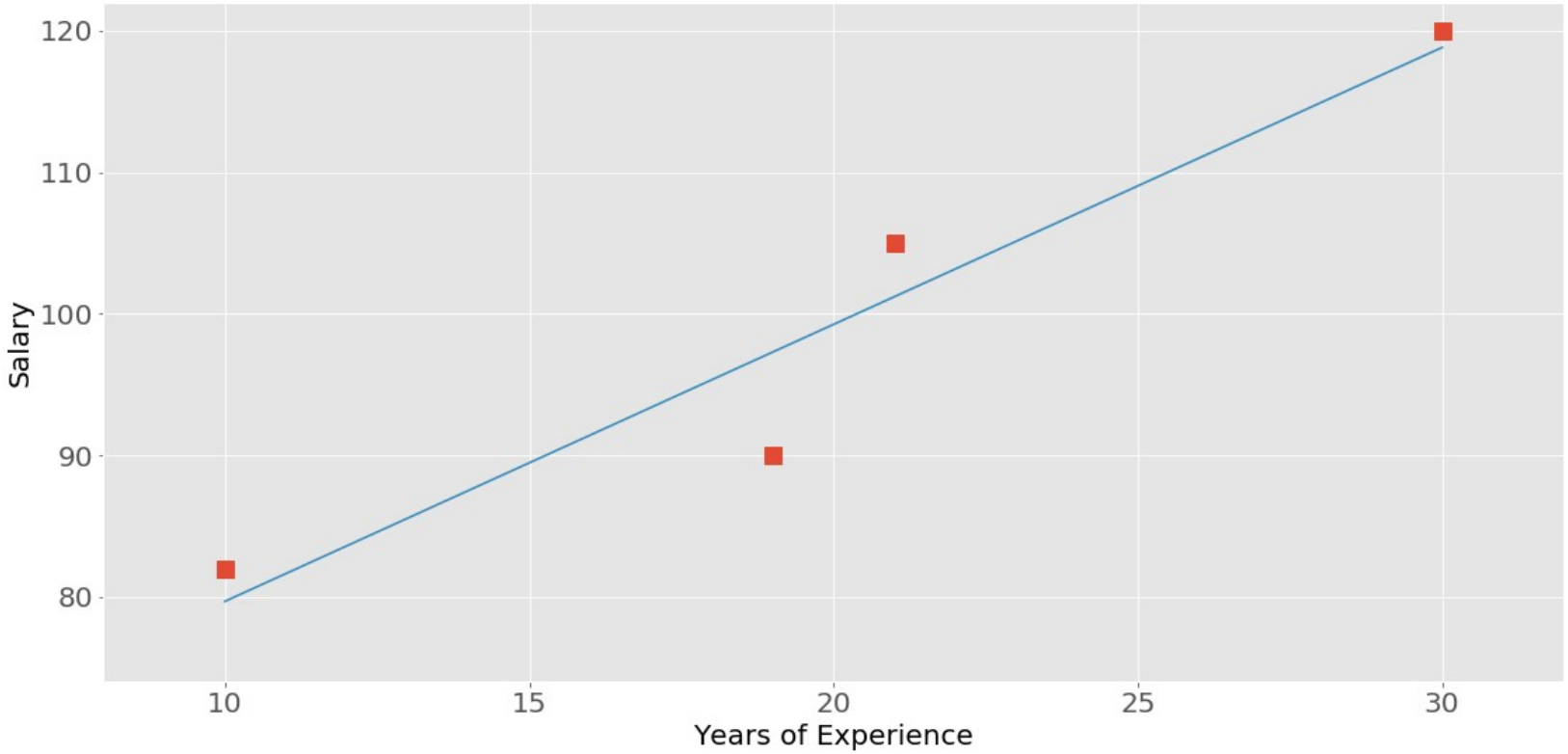




Forms of correlation

- Negative correlation (red dots): In the plot on the left, the y values tend to decrease as the x values increase. This shows strong negative correlation, which occurs when large values of one feature correspond to small values of the other, and vice versa.
- Weak or no correlation (green dots): The plot in the middle shows no obvious trend. This is a form of weak correlation, which occurs when an association between two features is not obvious or is hardly observable.
- Positive correlation (blue dots): In the plot on the right, the y values tend to increase as the x values increase. This illustrates strong positive correlation, which occurs when large values of one feature correspond to large values of the other, and vice versa.

Example: Employee table



Correlation Techniques

- There are several statistics that you can use to quantify correlation. We will be learning about three correlation coefficients:
  - Pearson's r
  - Spearman's rho
  - Kendall's tau
- Pearson's coefficient measures linear correlation, while the Spearman and Kendall coefficients compare the ranks of data.
- There are several NumPy, SciPy, and Pandas correlation functions and methods that you can use to calculate these coefficients.
- You can also use Matplotlib to conveniently illustrate the results.

What sort of correlation ?

- The values on the main diagonal of the correlation matrix (upper left and lower right) are equal to 1.
- The upper left value corresponds to the correlation coefficient for x and x, while the lower right value is the correlation coefficient for y and y. They are always equal to 1.
- However, what you usually need are the lower left and upper right values of the correlation matrix.
- These values are equal and both represent the Pearson correlation coefficient for x and y. In this case, it’s approximately 0.76.

Linear Correlation

- Linear correlation measures the proximity of the mathematical relationship between variables or dataset features to a linear function.
- If the relationship between the two features is closer to some linear function, then their linear correlation is stronger and the absolute value of the correlation coefficient is higher.

Pearson Correlation

- Consider a dataset with two features: x and y. Each feature has n values, so x and y are n-tuples. Say that the first value x_1 from x corresponds to the first value y_1 from y, the second value x_2 from x to the second value y_2 from y, and so on. Then, there are n pairs of corresponding values: (x_1, y_1), (x_2, y_2), and so on. Each of these x-y pairs represents a single observation.
- The Pearson (product-moment) correlation coefficient is a measure of the linear relationship between two features. It's the ratio of the covariance of x and y to the product of their standard deviations. It's often denoted with the letter r and called Pearson's r. You can express this value mathematically with this equation:

Pearson Correlation

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Pearson Correlation

- The Pearson correlation coefficient can take on any real value in the range $-1 \leq r \leq 1$.
- The maximum value $r = 1$ corresponds to the case when there's a perfect positive linear relationship between x and y. In other words, larger x values correspond to larger y values and vice versa.
- The value $r > 0$ indicates positive correlation between x and y.
- The value $r = 0$ corresponds to the case when x and y are independent.
- The value $r < 0$ indicates negative correlation between x and y.
- The minimal value $r = -1$ corresponds to the case when there's a perfect negative linear relationship between x and y. In other words, larger x values correspond to smaller y values and vice versa.

Pearson Correlation

Pearson's r Value	Correlation Between x and y
equal to 1	perfect positive linear relationship
greater than 0	positive correlation
equal to 0	independent
less than 0	negative correlation
equal to -1	perfect negative linear relationship

Regression

- Regression analysis is a very widely used statistical tool to establish a relationship model between two variables.
- One of these variable is called predictor variable whose value is gathered through experiments.
- The other variable is called response variable whose value is derived from the predictor variable.

Types of Regression

- Bi-variate (Linear)
- Multi-variate (Multiple)

Bivariate Regression

- In Bivariate Regression these two variables are related through an equation, where exponent (power) of both these variables is 1.
- Mathematically a linear relationship represents a straight line when plotted as a graph.
- A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.
- The general mathematical equation for a linear regression is –

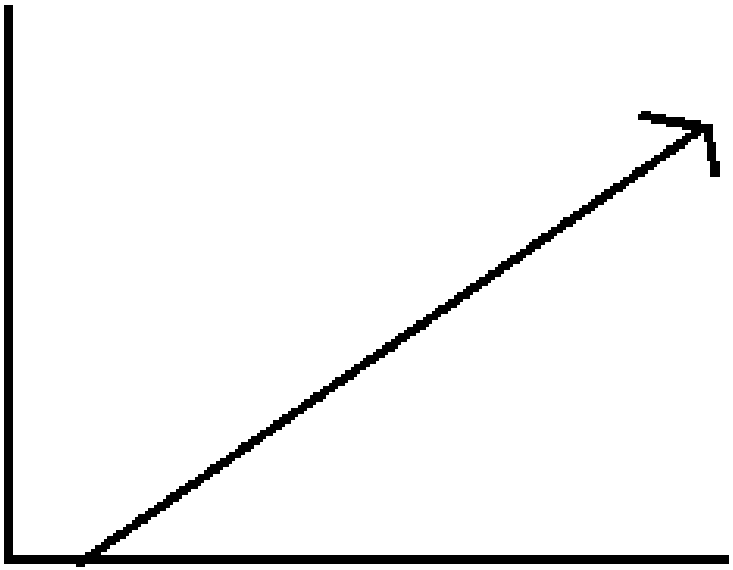
```
y = ax + b
```

y is the response variable.

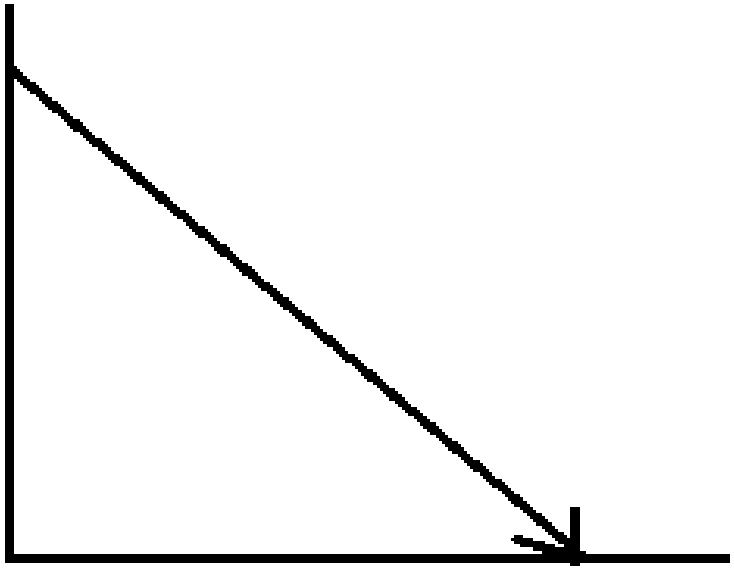
x is the predictor variable.

a and b are constants which are called the coefficients.

Bivariate Regression



Positive Linear Relationship



Negative Linear Relationship

Applications

- **Trend lines**: A trend line represents the variation in some quantitative data with passage of time (like GDP, oil prices, etc.). These trends usually follow a linear relationship. Hence, linear regression can be applied to predict future values.
- **Economics**: To predict consumption spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.
- **Finance**: Capital price asset model uses linear regression to analyze and quantify the systematic risks of an investment.
- **Biology**: Linear regression is used to model causal relationships between parameters in biological systems.

Steps to establish Linear Regression

- A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.
- The steps to create the relationship is –
  - Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
  - Create the object of Linear Regression Class.
  - Train the algorithm with dataset of X and y.
  - Get a summary of the relationship model to know the average error in prediction. Also called residuals.
  - To predict the weight of new persons, use the predict() function.

Example – Input data

- Below is the sample data representing the observations –

  # Values of height

  151, 174, 138, 186, 128, 136, 179, 163, 152, 131

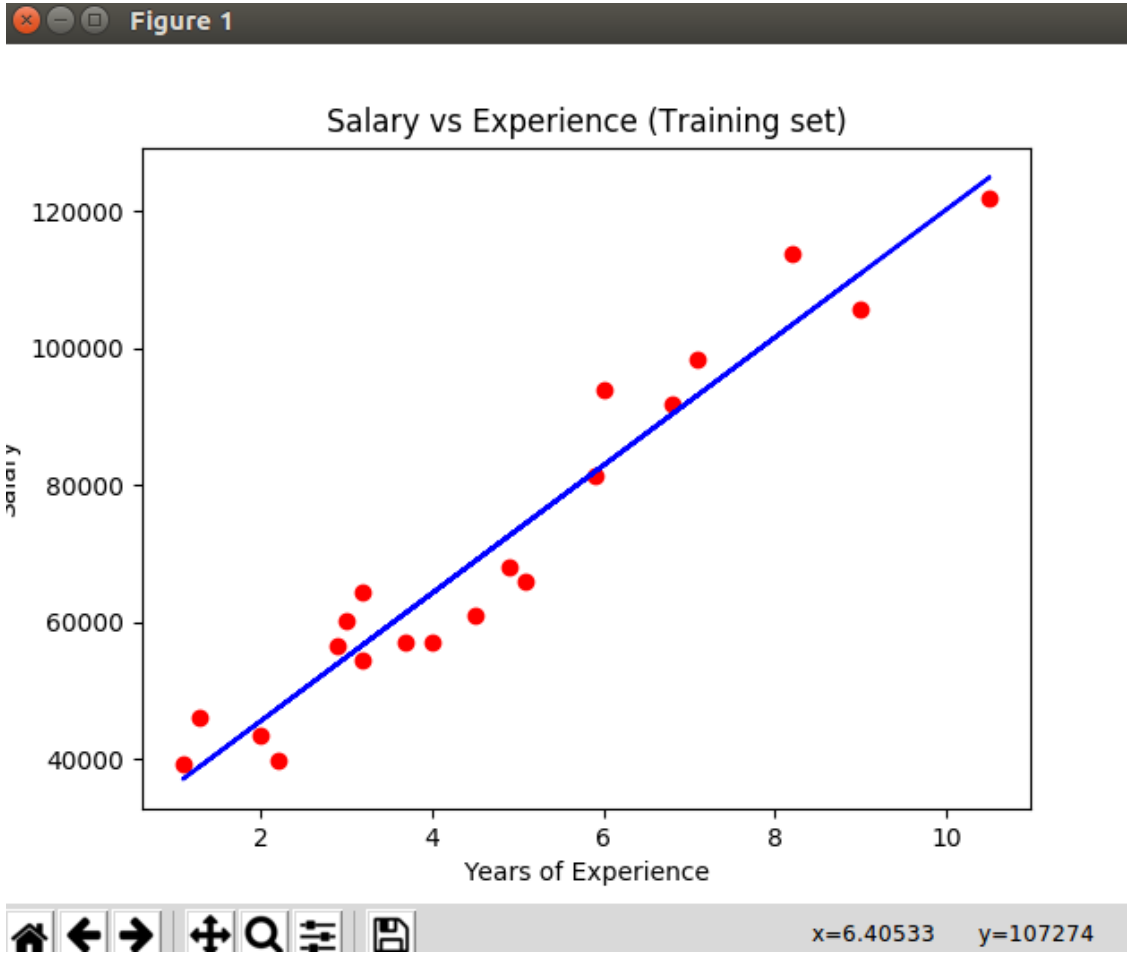
  # Values of weight.

  63, 81, 56, 91, 47, 57, 76, 72, 62, 48

Working with the csv file

	A	B
1	YearsExperience	Salary
2	1.1	39343
3	1.3	46205
4	1.5	37731
5	2	43525
6	2.2	39891
7	2.9	56642
8	3	60150
9	3.2	54445
10	3.2	64445
11	3.7	57189
12	3.9	63218
13	4	55794

Visualizing the regression



Multivariate (Multiple) Regression

- Multiple regression is an extension of linear regression into relationship between more than two variables.
- In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.
- The general mathematical equation for multiple regression is –

```
y = a + b1x1 + b2x2 +...bnxn
```

- Following is the description of the parameters used –
  - y is the response variable.
  - a, b1, b2...bn are the coefficients.
  - x1, x2, ...xn are the predictor variables.

Example – Input data

- Consider the data set "mtcars.csv". It gives a comparison between different car models in terms of mileage per gallon (mpg), cylinder displacement("disp"), horse power("hp"), weight of the car("wt") and some more parameters.
- The goal of the model is to establish the relationship between "mpg" as a response variable with "disp","hp" and "wt" as predictor variables.
- We create a subset of these variables from the mtcars data set for this purpose.

mtcars.csv

	A	B	C	D	E	F	G	H	I	J	K
1	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
2	21	6	160	110	3.9	2.62	16.46	0	1	4	4
3	21	6	160	110	3.9	2.875	17.02	0	1	4	4
4	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
5	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
6	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
7	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
8	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
9	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
10	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
11	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
12	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
13	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
14	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
15	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
16	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com