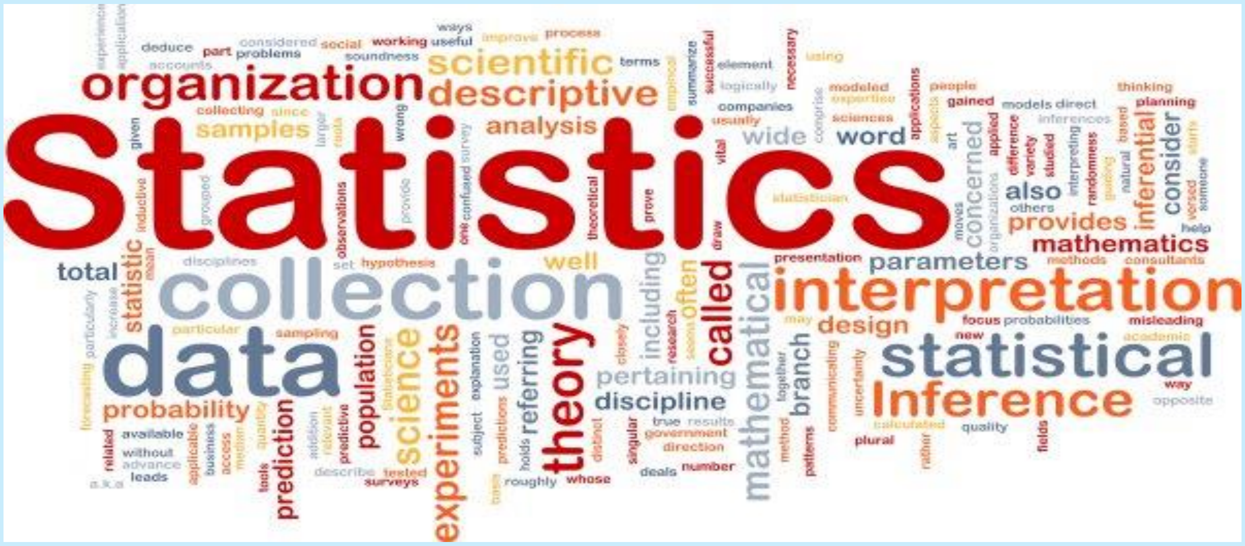
STATISTICS

1.BASIC STATISTICS

1.1 WHAT IS STATISTICS?

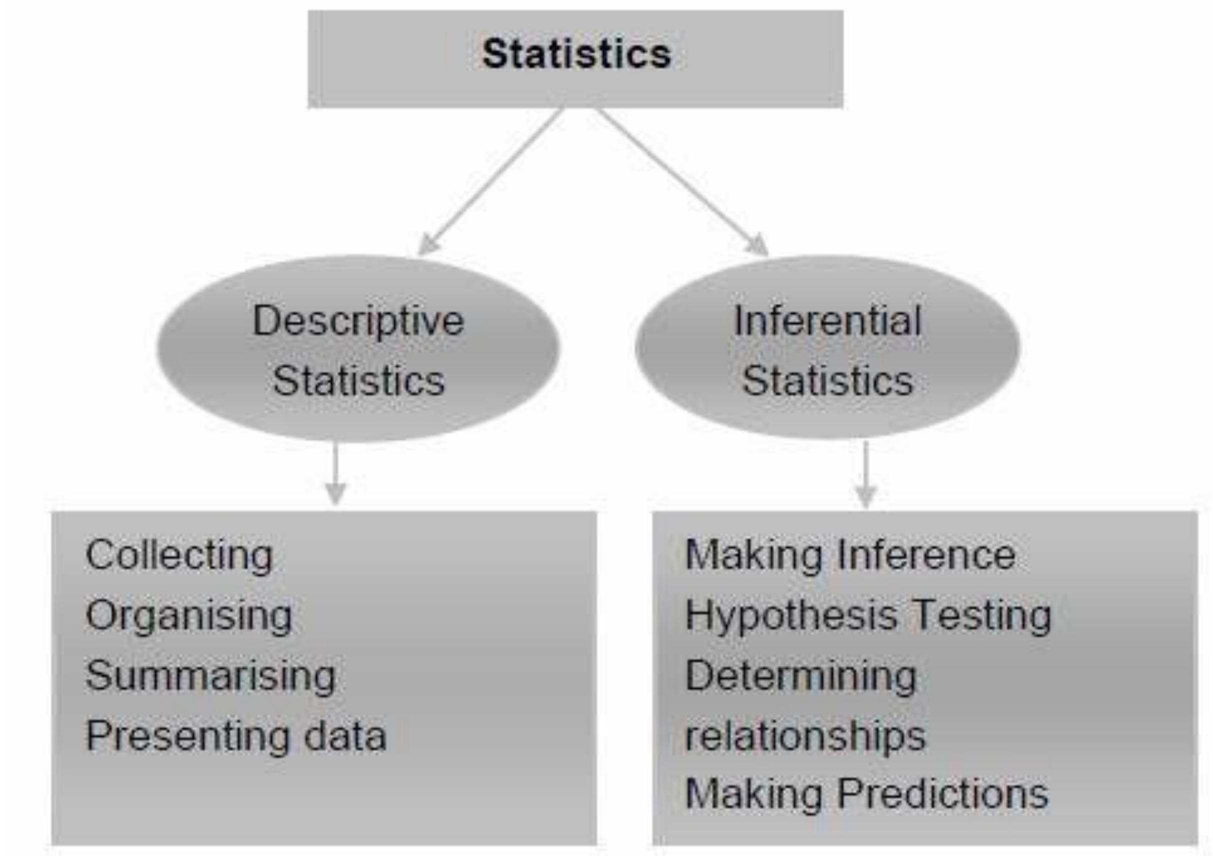


A branch of mathematics taking and transforming numbers into useful information for decision makers.

1.2 WHY STATISTICS?

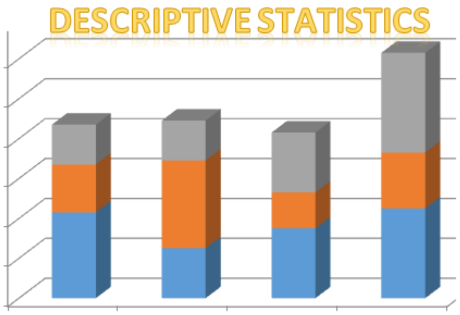
Knowledge of Statistics allows you to make better sense of the ubiquitous use of numbers. It is important for researchers and also consumers of research to understand statistics **so** that they can be informed, evaluate the credibility and usefulness of information, and make appropriate decisions.

1.3 TYPES OF STATISTICS



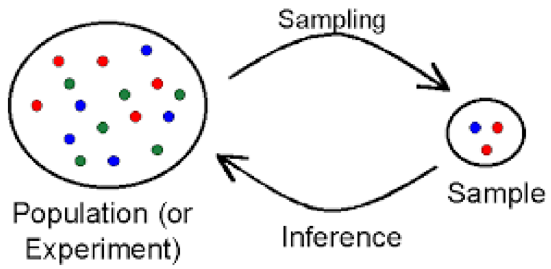
1.3.1 DESCRIPTIVE STATISTICS

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.



1.3.2 INFERENTIAL STATISTICS

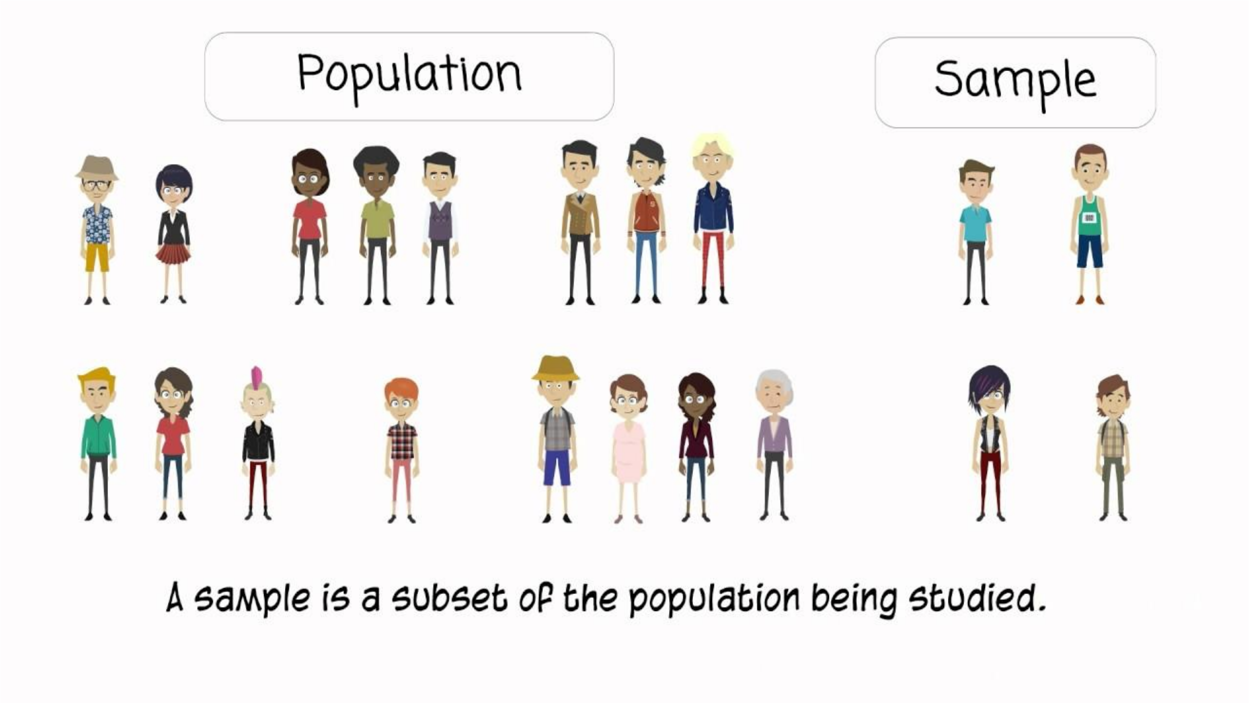
With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.



DESCRIPTIVE STATISTICS

1.4 POPULATION AND SAMPLE

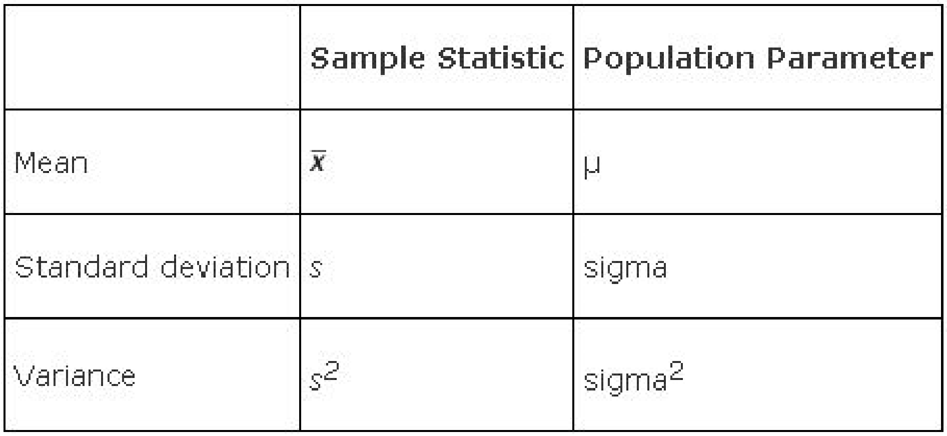
Whenever we hear the term ‘population,’ the first thing that strikes our mind is a large group of people. In the same way, in statistics population denotes a large group consisting of elements having at least one common feature. The term is often contrasted with the sample, which is nothing but a part of the population that is so selected to represent the entire group. Population represents the entirety of persons, units, objects and anything that is capable of being conceived, having certain properties. On the contrary, the sample is a finite subset of the population, that is chosen by a systematic process, to find out the characteristics of the parent set. The article presented below describes the differences between population and sample.



A sample is a subset of the population being studied.

1.4.1 PARAMETER AND STATISTIC

Parameters are numbers that summarize data for an entire population. Statistics are numbers that summarize data from a sample, i.e. some subset of the entire population.



	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	sigma
Variance	s^2	sigma^2

1.5 VARIABLE AND DATA

In statistics, a variable has two defining characteristics:

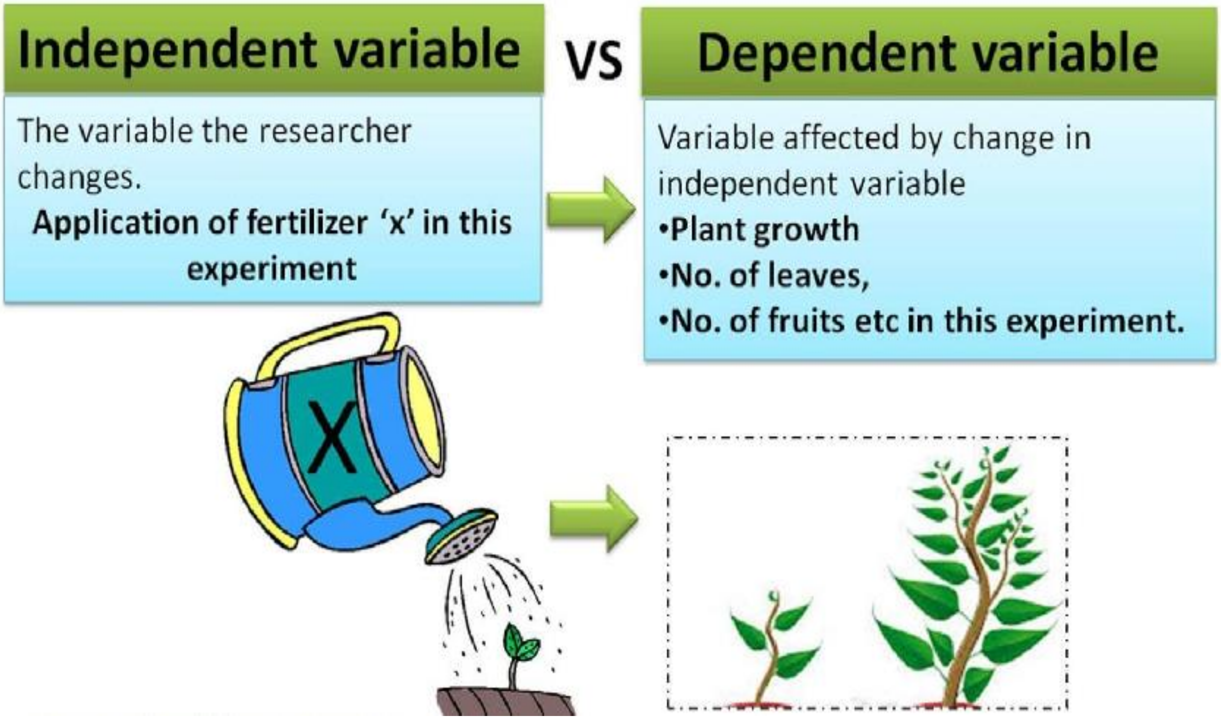
- A variable is an attribute that describes a person, place, thing, or idea.
- The value of the variable can "vary" from one entity to another.

For example, a person's hair colour is a potential variable, which could have the value of "blond" for one person and "brunette" for another.

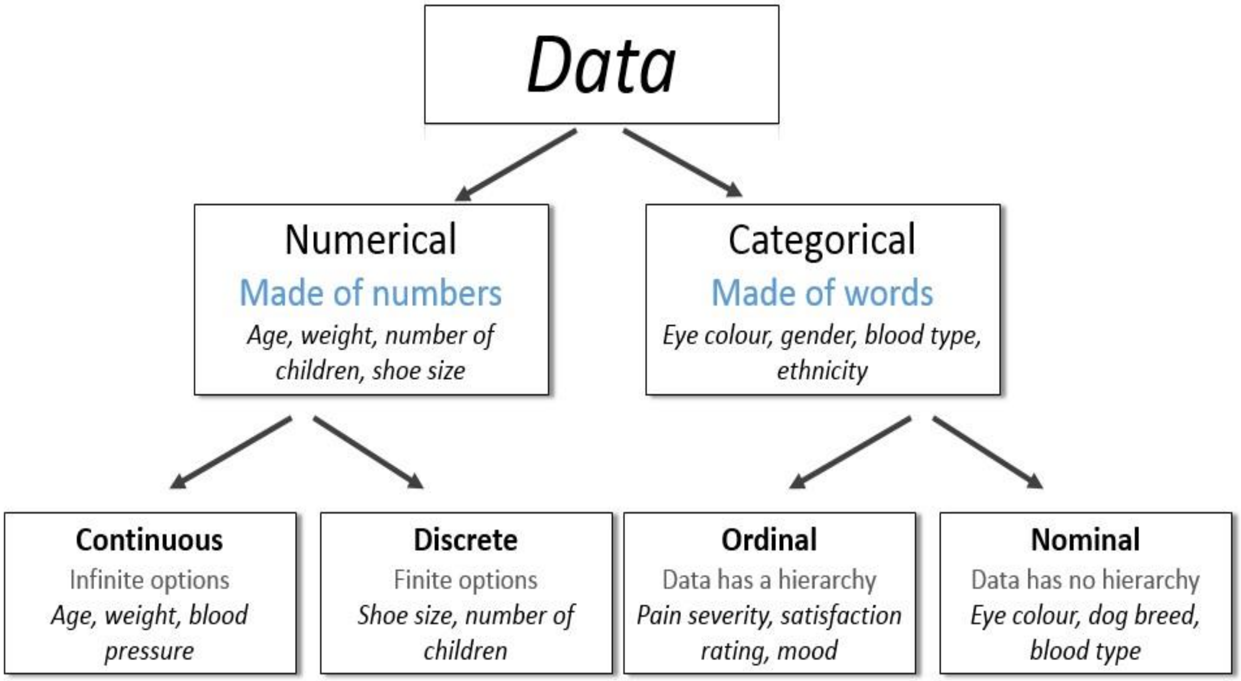
In fact, a good definition of data is "facts or figures from which conclusions can be drawn". Data can take various forms, but are often numerical. As such, data can relate to an enormous variety of aspects, for example:

- the daily weight measurements of each individual in your classroom;
- the number of movie rentals per month for each household in your neighbourhood;
- the city's temperature (measured every hour) for a one-week period.

1.5.1 TYPES OF VARIABLES


Independent variable
vs
Dependent variable
The variable the researcher changes.
Application of fertilizer 'x' in this experiment
Variable affected by change in independent variable
•Plant growth
•No. of leaves,
•No. of fruits etc in this experiment.
X

1.5.2 TYPES OF DATA


Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

1.6 MEASURES OF CENTRE - MEAN, MEDIAN AND MODE

Mean, median, and mode are different measures of centre in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

Mean: The "average" number: found by adding all data points and dividing by the number of data points.

Calculating the mean

There are many different types of mean, but usually when people say mean, they are talking about the arithmetic mean.

The arithmetic mean is the sum of all of the data points divided by the number of data points.

mean= (Sum of data) / (No. of data points).

Here's the same formula written more formally:

Mean = $\frac{\sum_i x}{n}$,

Example: The mean of 4, 1, and 7 is (4+1+7)/3 = 12/3 = 4

Median: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Finding the median

- Arrange the data points from smallest to largest.

- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list.

Example: The median of 444, 1, and 7 is 4 because when the numbers are put in order (1, 4, 7) the number 4 is in the middle.

Mode: The most frequent number—that is, the number that occurs the highest number of times.

Finding the mode

The mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

Example: The mode of {4, 222, 444, 333, 222, 2} is 222 because it occurs three times, which is more than any other number.

1.7 THE MEASURE OF VARIABLITY – RANGE, S.D AND COEFF. OF VARIATION

Range: In statistics, the range is a measure of spread: it’s the difference between the highest value and the lowest value in a data set.

Calculation of Range

$$\text{Range} = x_{\max} - x_{\min} = x_n - x_1$$

Observations are ordered in ascending order.

Variance (σ^2): Variance is a measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.

The Equation Defining Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

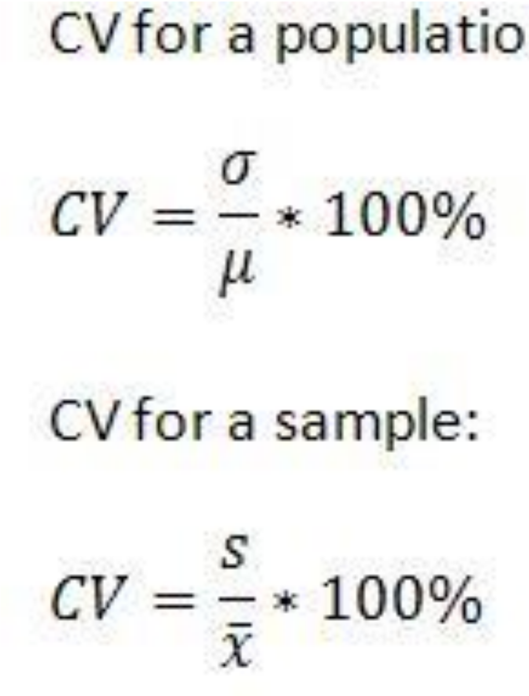
Standard Deviation (σ): The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

The Equation Defining Standard Deviation

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

Coefficient of Variation: A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. It represents the ratio of the standard deviation to the mean.

The Equation Defining CV



CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

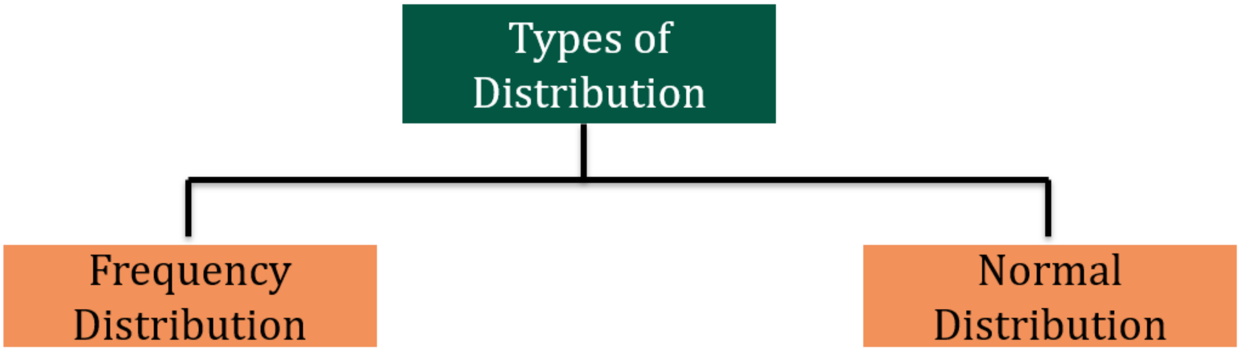
$$CV = \frac{s}{\bar{x}} * 100\%$$

The most common use of the coefficient of variation is to assess the precision of a technique. It is also used as a measure of variability when the standard deviation is proportional to the mean, and as a means to compare variability of measurements made in different units.

Less Coefficient of variance means less risk and more consistency.

More coefficient of variance means more risk and less consistency.

1.8 TYPES OF DISTRIBUTION



1.9 FREQUENCY DISTRIBUTION

A *frequency distribution* is an overview of all distinct values in some variable and the number of times they occur. These are visual displays that organize and present frequency counts so that the information can be interpreted more easily.

1.9.1 HOW DO WE SHOW A FREQUENCY DISTRIBUTION?

A frequency distribution of data can be shown in a table or graph. Some common methods of showing frequency distributions include frequency tables, histograms or bar charts.

1.9.2 FREQUENCY TABLES

A frequency table is a simple way to display the number of occurrences of a particular value or characteristic.

For example, if we have collected data about height from a sample of 50 children, we could present our findings as:

Height of Children

Height (cm) of children	**Absolute frequency**	**Relative frequency**
120 – less than 130	9	18%
130 – less than 140	10	20%
140 – less than 150	13	26%
150 – less than 160	11	22%
160 – less than 170	7	14%
Total	**50**	**100%**

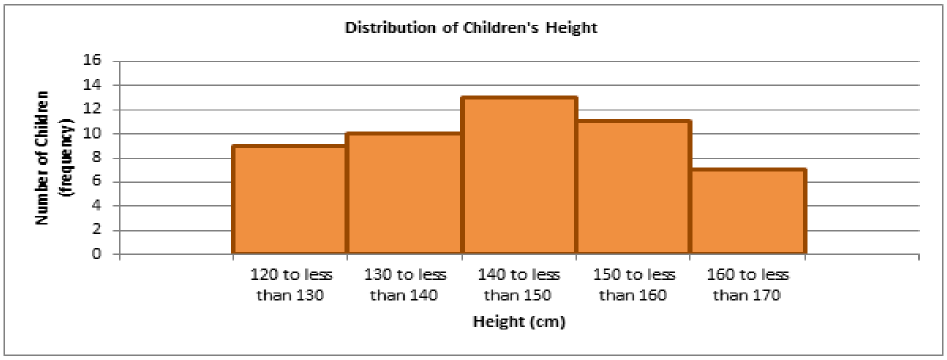
From this frequency table we can quickly identify information such as 7 children (14% of all children) are in the 160 to less than 170 cm height range, and that there are more children with heights in the 140 to less than 150 cm range (26% of all children) than any other height range.

1.9.3 FREQUENCY GRAPH

Histograms and bar charts are both visual displays of frequencies using columns plotted on a graph. The Y-axis (vertical axis) generally represents the frequency count, while the X-axis (horizontal axis) generally represents the variable being measured.

A histogram is a type of graph in which each column represents a numeric variable, in particular that which is continuous and/or grouped. It shows the distribution of all observations in a quantitative dataset. It is useful for describing the shape, center and spread to better understand the distribution of the dataset.

For example:
The histogram below shows the same information as the frequency table.



A bar chart is a type of graph in which each column (plotted either vertically or horizontally) represents a categorical variable or a discrete ungrouped numeric variable. It is used to *compare* the frequency (count) for a category or characteristic with another category or characteristic.

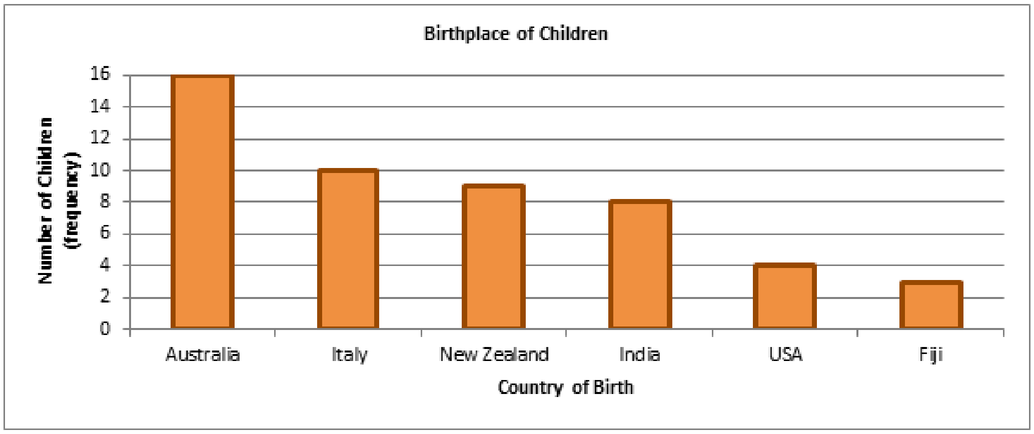
For example:
If data had been collected for 'country of birth' from a sample of children, a bar chart could be used to plot the data as 'country of birth' is a categorical variable.

Birthplace of Children

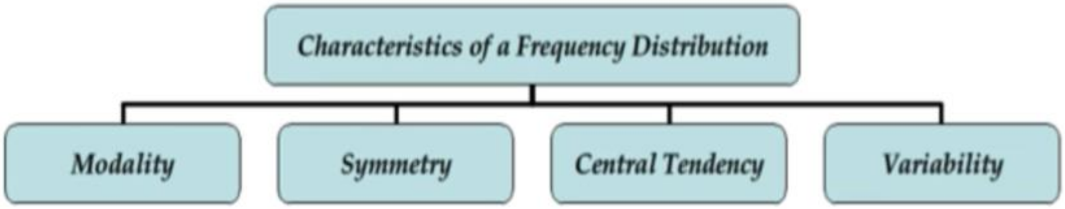
Country of Birth	**Absolute frequency**	**Relative frequency**
Australia	16	32%
Fiji	3	6%
India	8	16%
Italy	10	20%
New Zealand	9	18%

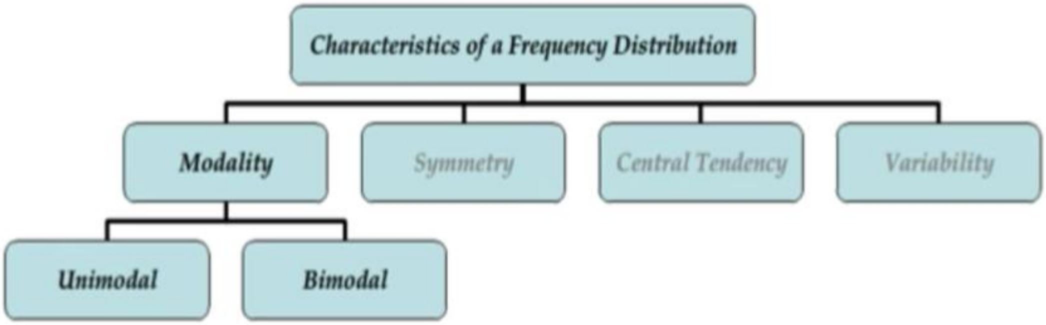
United States of America	4	8%
Total	**50**	**100%**

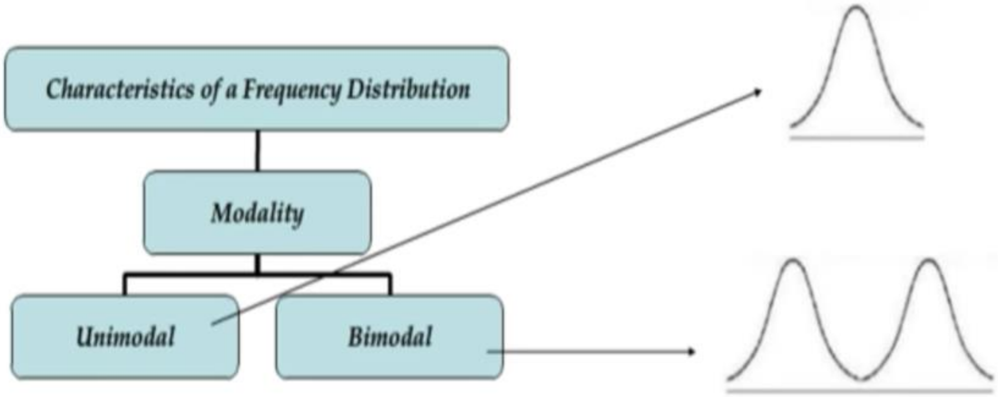
The bar chart below shows us that 'Australia' is the most commonly observed country of birth of the 50 children sampled, while 'Fiji' is the least common country of birth.



1.9.4 CHARACTERISTICS OF FREQUENCY DISTRIBUTION





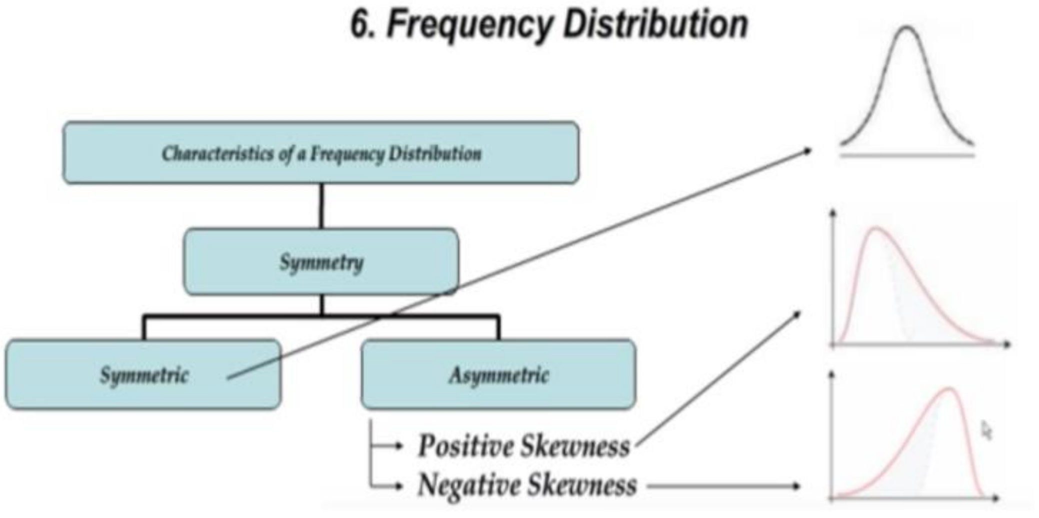


Characteristics of a Frequency Distribution

Modality

Unimodal

Bimodal



6. Frequency Distribution

Characteristics of a Frequency Distribution

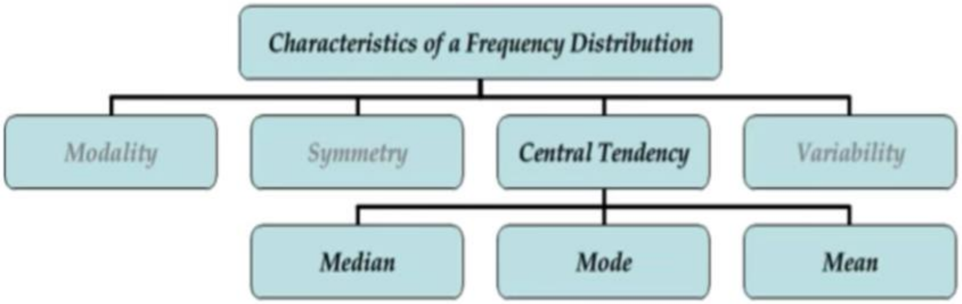
Symmetry

Symmetric

Asymmetric

Positive Skewness

Negative Skewness



Characteristics of a Frequency Distribution

Modality

Symmetry

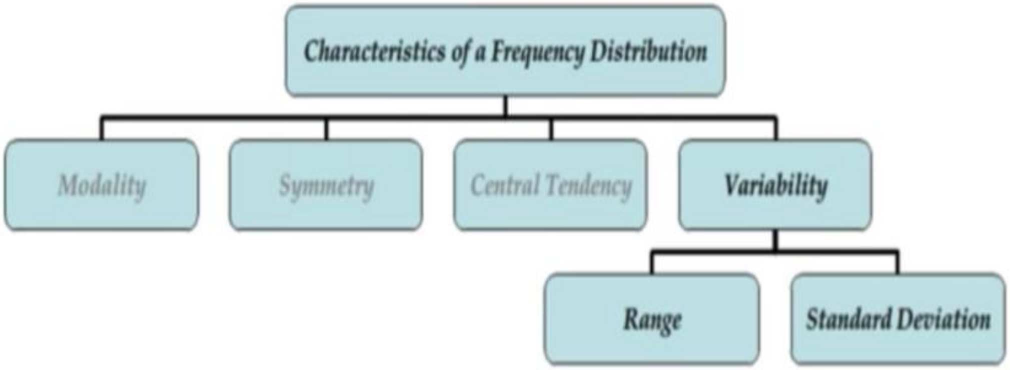
Central Tendency

Variability

Median

Mode

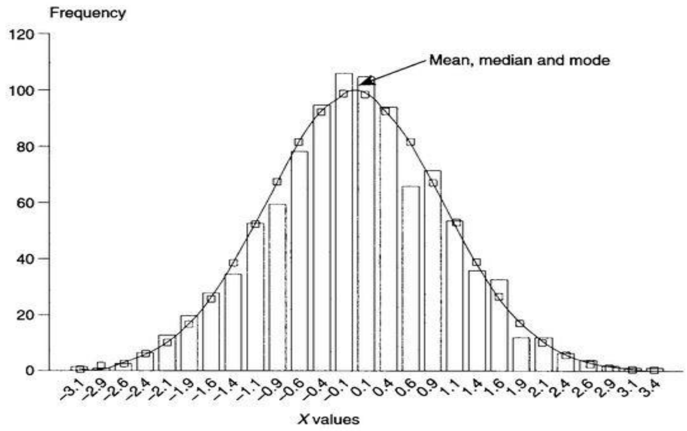
Mean



1.10 NORMAL (GAUSSIAN) DISTRIBUTION

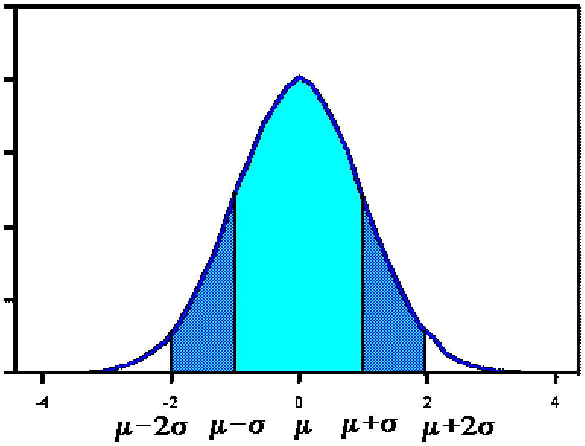
Many sets of scores have frequency distributions that are shaped roughly like a mathematical distribution called the normal distribution. Informally, it is called the "bell shaped curve".

In this figure a normal distribution, is superimposed over a frequency distribution. The normal distribution roughly matches the frequency distribution.



The normal distribution has many useful properties. One is that the standard deviation of the normal distribution cuts off specific proportions of the curve. The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation.

The **Standard Normal** curve, shown in figure, has mean 0 and standard deviation 1. If a dataset follows a normal distribution, then about 68% of the observations will fall within σ of the mean μ, which in this case is with the interval (-1,1). About 95% of the observations will fall within 2 standard deviations of the mean, which is the interval (-2,2) for the standard normal, and about 99.7% of the observations will fall within 3 standard deviations of the mean, which corresponds to the interval (-3,3) in this case. Although it may appear as if a normal distribution does not include any values beyond a certain interval, the density is actually positive for all values, $(-\infty, \infty)$.

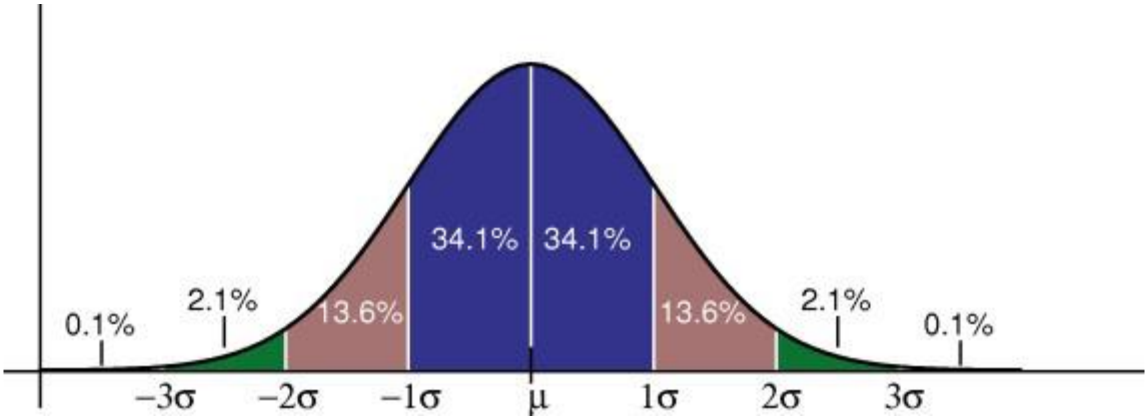


1.10.1 THE EMPRICAL RULE

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- 68% of the data falls within one standard deviation of the mean.

- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.



The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the mean; the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the mean; the normal distribution will be flatter and wider.

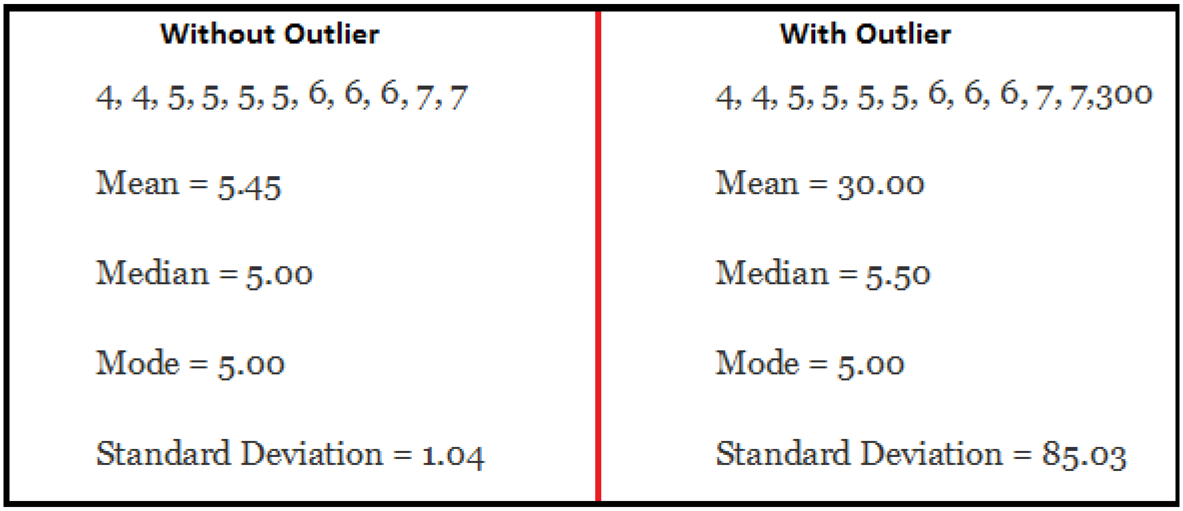
1.11 CONCEPT OF OUTLIERS

In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Effect of outliers on Central Tendency

Outliers affect the mean value and standard deviation of the data but have little effect on the median or mode of a given set of data.

For example.,



Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Notice that the outlier had a small or no effect on the median and mode of the data.

It should be noted that because outliers affect the mean and have little effect on the median, the median is often used in cases where outliers are more.

But median cannot be used as “average” in all the cases so in order to find the “average” we neglect the outliers so that it does not affect our “mean” or “average”. To find which of the values are outliers we use BOX plot to find them and eliminate them respectively.

1.12 PERCENTILES

If all you are interested in is where you stand compared to the rest of the herd, you need a statistic that reports relative standing, and that statistic is called a percentile.

The k^{th} *percentile* is a value in a data set that splits the data into two pieces: The lower piece contains *k* percent of the data, and the upper piece contains the rest of the data (which amounts to [100 – *k*] percent, because the total amount of data is 100%). *Note: k* is any number between 0 and 100.

Formula for percentile to index value

I = $(\frac{N}{100} Xn)$

Where I is the Index and n is the no. of data points

If I is an integer then I = $\frac{I+(I+1)}{2}$

If I is a decimal then round it off to the next nearest value.

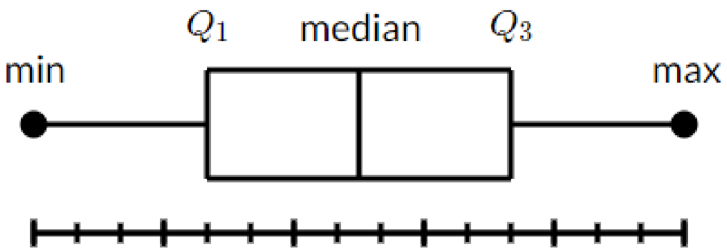
1.13 BOX PLOT

Box plots are useful for identifying outliers and for comparing distributions. A Box Plot is the visual representation of the statistical five number summary of a given data set.

A Five Number Summary includes:

- Q1 – quartile 1, the median of the lower half of the data set
- Q2 – quartile 2, the median of the entire data set
- Q3 – quartile 3, the median of the upper half of the data set
- IQR – interquartile range, the difference from Q3 to Q1

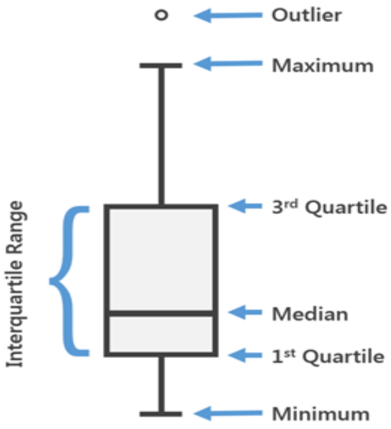
- Extreme Values – The min and max values in a data set



Steps for Sketching the Box plot

1. There are several steps in constructing a box plot. The first relies on the 25th (Q_1), 50th (Q_2) and 75th (Q_3) percentiles in the distribution of scores.
2. Finding the Inter Quartile Range IQR = Q_3 - Q_1
3. Finding the min or the Lower Whisker = Q_1 -1.5(IQR)
4. Finding the max or the Upper Whisker = Q_3 -1.5(IQR)
5. Sketch the box plot

The values Below the min and the values above the max are outliers.



1.14 COVARIANCE AND CORRELATION COEFFICIENT

Covariance indicates how two variables are related. A positive covariance means the variables are positively related, while a negative covariance means the variables are inversely related.

The formula for calculating covariance of sample

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

x = the independent variable
y = the dependent variable
n = number of data points in the sample
$\bar{x}$ = the mean of the independent variable *x*
$\bar{y}$ = the mean of the dependent variable *y*

Correlation is another way to determine how two variables are related. In addition to telling you whether variables are positively or inversely related, correlation also tells you the degree to which the variables tend to move together.

As stated above, covariance measures variables that have different units of measurement. Using covariance, you could determine whether units were increasing or decreasing, but it was impossible to measure the degree to which the variables moved together because covariance does not use one standard unit of measurement. To measure the degree to which variables move together, you must use correlation.

Formula for calculating correlation of two variables

$$r_{(x,y)} = \frac{COV(x,y)}{s_x s_y}$$

$r_{(x,y)}$ = correlation of the variables x and y
$COV(x, y)$ = covariance of the variables x and y
s_x = sample standard deviation of the random variable x
s_y = sample standard deviation of the random variable y

1.15 CENTRAL LIMIT THEOREM

The central limit theorem states that when samples from a data set with a known variance are aggregated their mean roughly equals the population mean. Said another way, CLT is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all the samples will follow an approximate normal distribution pattern, with all variances being approximately equal with some standard error which is equal to the SD of the population divided by each sample's size.

Standard Error = $\frac{\boldsymbol{\sigma}}{\sqrt{n}}$

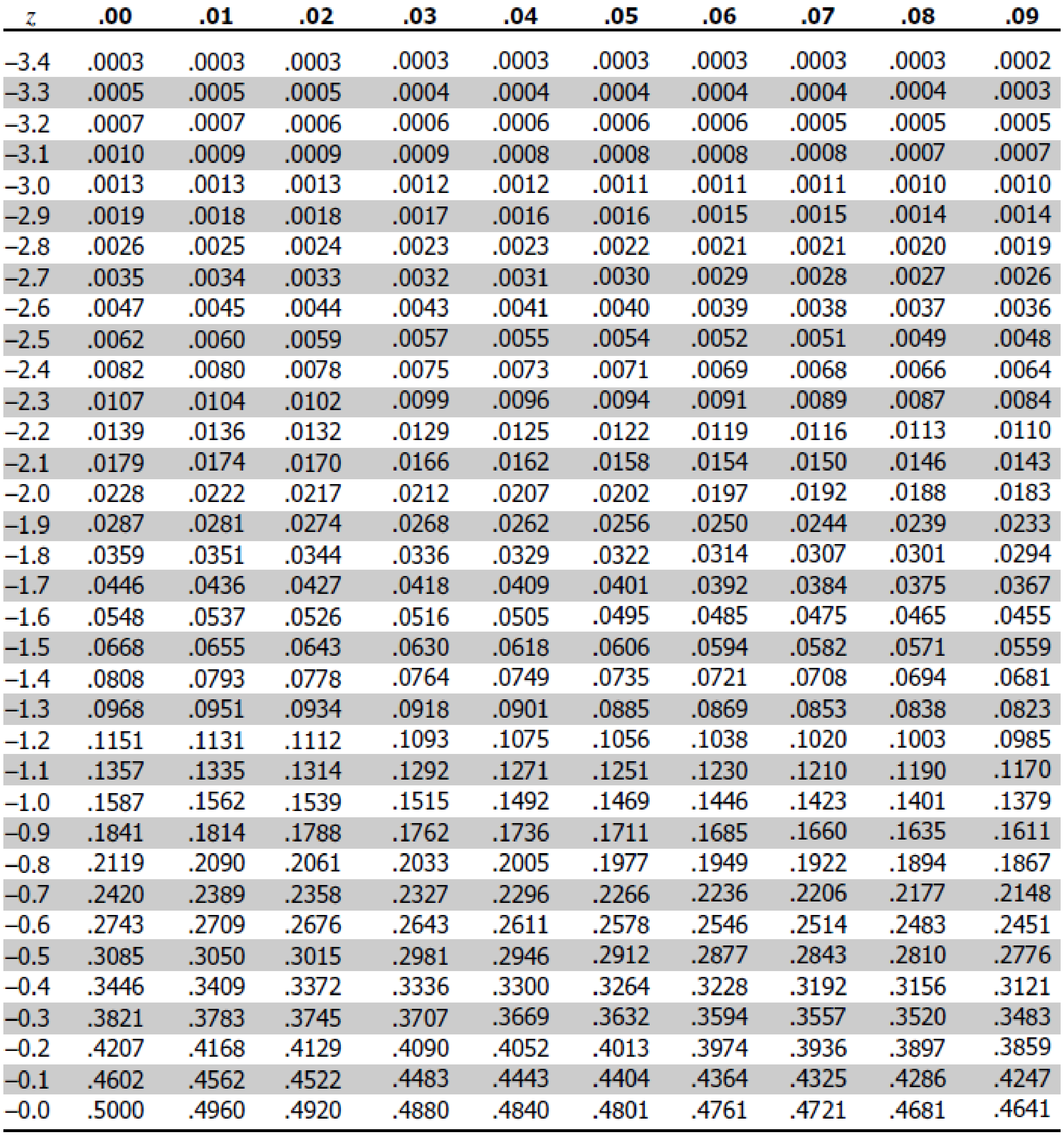
1.15.1 Z-SCORE

A **z-score** (aka, a standard **score**) indicates how many standard deviations an element is from the mean. A **z-score** can be calculated from the following formula.

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

Where you find the Z- value form the formula and find its corresponding Z- score from the Z- table.

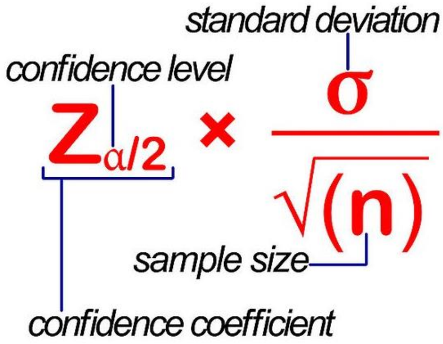


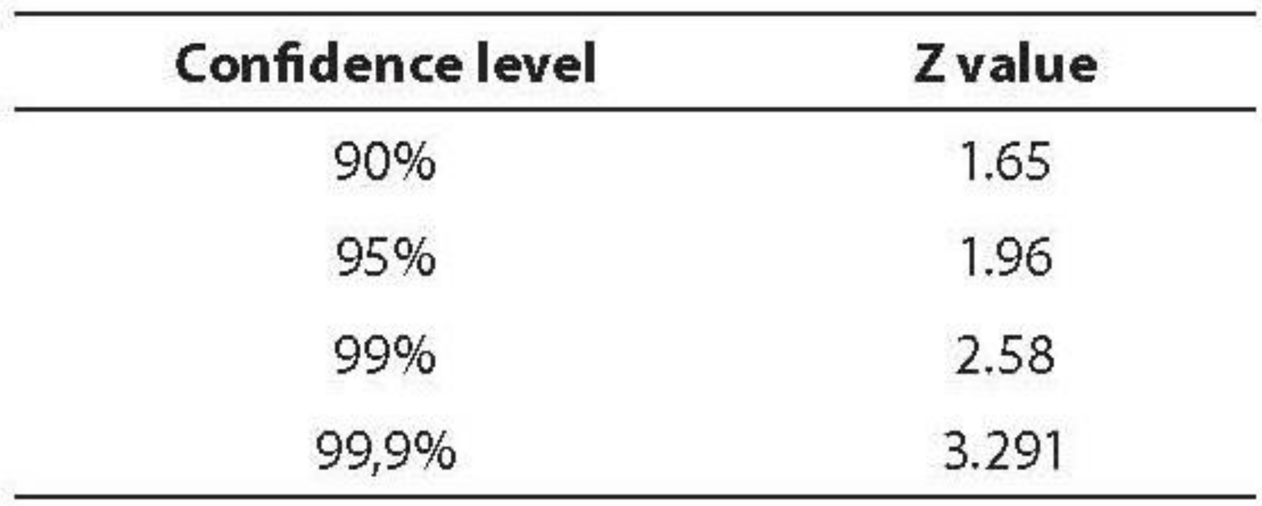
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

1.16 CONFIDENCE INTERVAL

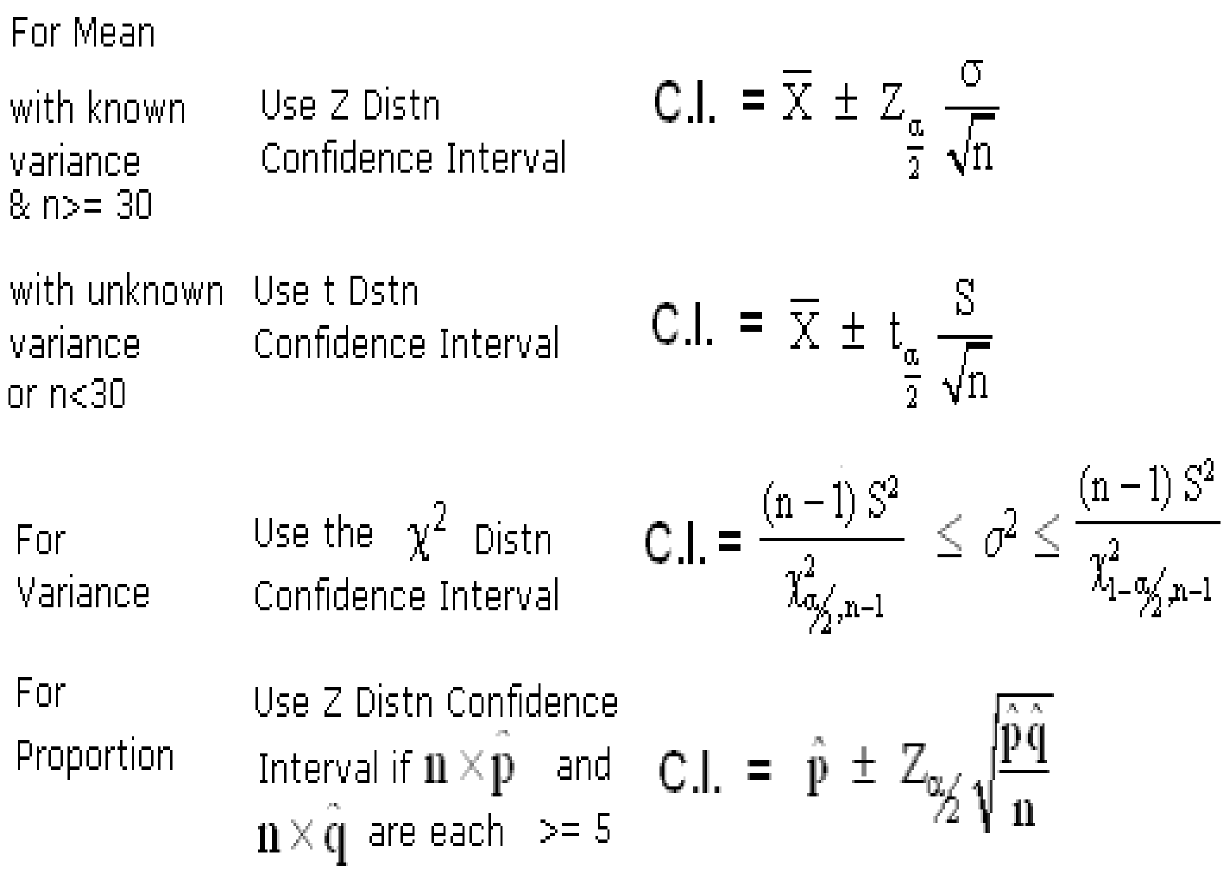
A **confidence interval** (**CI**) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. More strictly speaking, the **confidence level** represents the frequency (i.e. the proportion) of possible confidence intervals that contain the true value of the unknown population parameter.

For a population with unknown mean μ and known standard deviation σ, a confidence interval for the population mean, based on a simple random sample (SRS) of size *n* is





Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291



For Mean		
with known variance & n>= 30	Use Z Distn Confidence Interval	$C.I. = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
with unknown variance or n<30	Use t Dstn Confidence Interval	$C.I. = \bar{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$
For Variance	Use the χ^2 Distn Confidence Interval	$C.I. = \frac{(n-1)\,S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2,n-1}}$
For Proportion	Use Z Distn Confidence Interval if $n \times \hat{p}$ and $n \times \hat{q}$ are each >= 5	$C.I. = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$

1.17 MARGIN ERROR

A margin of error tells you how many percentage points your results will differ from the real population value.

Margin of error = Critical value x Standard error of the statistic

INFERENTIAL STATISTICS

2.HYPOTHESIS TESTING

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.

Steps in performing hypothesis testing

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

What is null hypothesis?

the null hypothesis is always the accepted fact. Simple examples of null hypotheses that are generally accepted as being true are:

1. DNA is shaped like a double helix.
2. There are 8 planets in the solar system (excluding Pluto).

Stating the null hypothesis

You won't be required to actually perform a real experiment or survey in elementary statistics (or even disprove a fact like

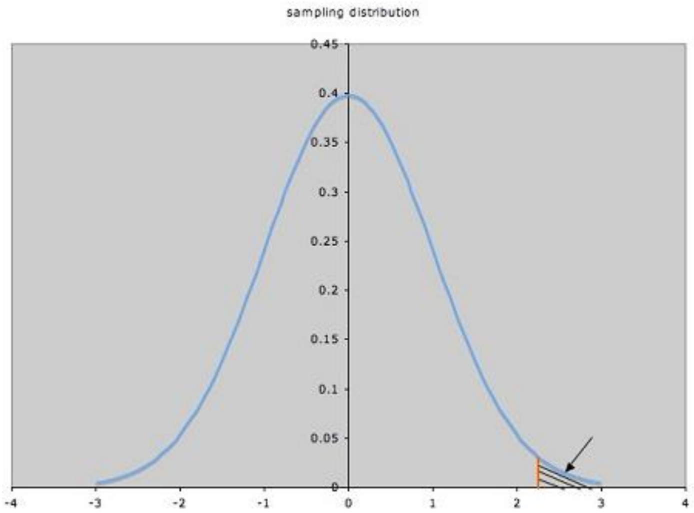
“Pluto is a planet”!), so you’ll be given word problems from real-life situations. You’ll need to figure out what your hypothesis is from the problem.

Rejecting the null hypothesis

Ten or so years ago, we believed that there were 9 planets in the solar system. Pluto was demoted as a planet in 2006. The null hypothesis of “Pluto is a planet” was replaced by “Pluto is not a planet.” Of course, rejecting the null hypothesis isn’t always that easy — the hard part is usually figuring out what your null hypothesis is in the first place.

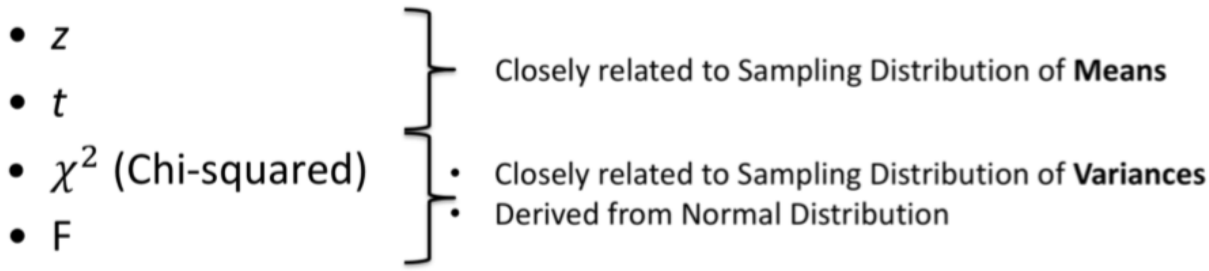
2.1 CRITICAL VALUE

A critical value is a line on a graph that splits the graph into sections. One or two of the sections is the “rejection region”; if your test value falls into that region, then you reject the null hypothesis.



A one tailed test with the rejection in one tail. The critical value is the red line to the left of that region.

2.2 ONE VARIABLE TESTS



- z
- t
- χ^2 (Chi-squared)
- F

Closely related to Sampling Distribution of **Means**

- Closely related to Sampling Distribution of **Variances**
- Derived from Normal Distribution

2.2.1 Z- TEST

Z- test conditions

- The sample size is large (n > 30),
- The data were collected in a random way, each observation must be independent of the others, the sampling distribution must be normal or approximately normal, and the population standard deviation must be known

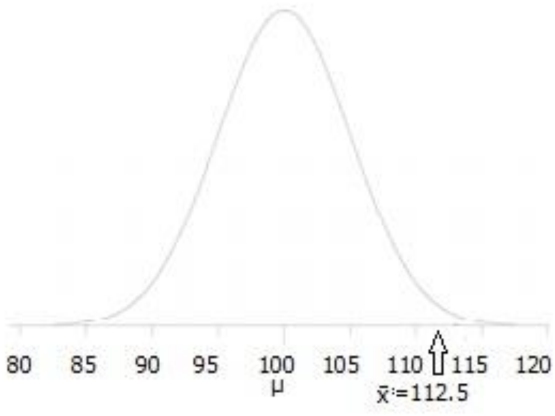
Let us understand the test with an example

A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal’s claim? The mean population IQ is 100 with a standard deviation of 15.

Step 1: State the Null hypothesis. The accepted fact is that the population mean is 100, so: H_0: μ=100

Step 2: State the Alternate Hypothesis. The claim is that the students have above average IQ scores, so:
H_1: μ > 100.
The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test.

Step 3: Draw a picture to help you visualize the problem.



Step 4: State the alpha level. If you aren't given an alpha level, use 5% (0.05).

Step 5: Find the rejection region area (given by your alpha level above) from the z-table. An area of .05 is equal to a z-score of 1.645.

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Step 6: Find the test statistic using this formula: For this set of data: z= (112.5-100) / (15/√30) =4.56.

If Step 6 is greater than Step 5, reject the null hypothesis. If it's less than Step 5, you cannot reject the null hypothesis. In this case, it is greater (4.56 > 1.645), so you can reject the null.

2.2.2 T-TEST

T- test conditions

- The sample size is not large (n < 30),
- The data were collected in a random way, each observation must be independent of the others, and the sampling distribution must be normal or approximately normal.

Let us understand the test with an example

Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.

Step 1: Write your null hypothesis statement (How to state a null hypothesis). The accepted hypothesis is that there is no difference in sales, so:
H_0: μ = $100.

Step 2: Write your alternate hypothesis. This is the one you’re testing. You think that there *is* a difference (that the mean sales increased), so: H_1: $\mu > \$100$.

Step 3: Identify the following pieces of information you’ll need to calculate the test statistic. The question should give you these items:

1. **The sample mean**($\bar{x}$). This is given in the question as $130.
2. **The population mean**(μ). Given as $100 (from past data).
3. **The sample standard deviation**(s) = $15.
4. **Number of observations**(n) = 25.

Step 4: Insert the items from above into the t score formula.

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

$t = (130 – 100) / ((15 / \sqrt{(25)})$
$t = (30 / 3) = 10$
This is your **calculated t-value**.

Step 5: Find the t-table value. You need two values to find this:

1. The alpha level: given as 5% in the question.

2. The degrees of freedom, which is the number of items in the sample (n) minus 1: 25 – 1 = 24.

Look up 24 degrees of freedom in the left column and 0.05 in the top row. The intersection is 1. 711.This is your one-tailed critical t-value.

What this critical value means is that we would expect most values to fall under 1.711. If our calculated t-value (from Step 4) falls within this range, the null hypothesis is likely true.

Step 5: Compare Step 4 to Step 5. The value from Step 4 **does not** fall into the range calculated in Step 5, so we can reject the null hypothesis. The value of 10 falls into the rejection region (the left tail). In other words, it’s highly likely that the mean sale is greater. The sales training was probably a success.

2.2.3 X² TEST (CHI SQUARE TEST)

Type 1: GOODNESS OF FIT

Set up the hypothesis for Chi-Square goodness of fit test:

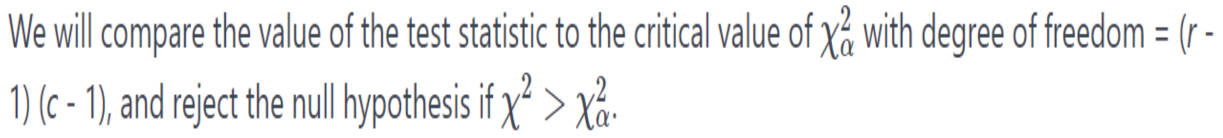
Null hypothesis: In Chi-Square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and the expected value.

Alternative hypothesis: In Chi-Square goodness of fit test, the alternative hypothesis assumes that there is a significant difference between the observed and the expected value.

Compute the value of Chi-Square goodness of fit test using the following formula:

$$\chi^2 = \sum (O - E)^2 / E$$

Where CHI square here is goodness of fit.



We will compare the value of the test statistic to the critical value of χ^2_α with degree of freedom = (r - 1) (c - 1), and reject the null hypothesis if $\chi^2 > \chi^2_\alpha$.

Type 2: INDEPENDENCE TEST

It tests the independence of two categorical variables. We can summarize two categorical variables within a two-way table, also called a $r \times c$ contingency table, where r = number of rows, c = number of columns. Our question of interest is “Are the two variables independent?” This question is set up using the following hypothesis statements:

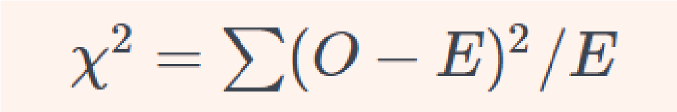
Null Hypothesis

The two categorical variables are independent

Alternative Hypothesis

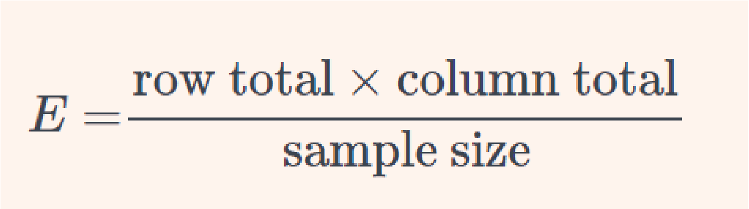
The two categorical variables are dependent

Chi-Square Test Statistic

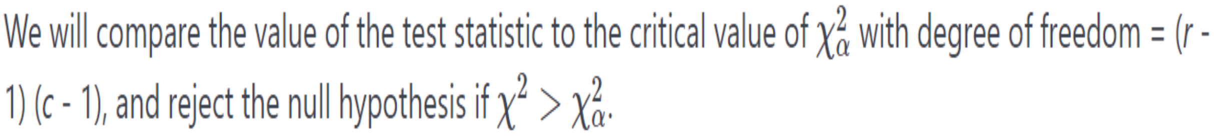


$$\chi^2 = \sum (O - E)^2 / E$$

where *O* represents the observed frequency. *E* is the expected frequency under the null hypothesis and computed by:

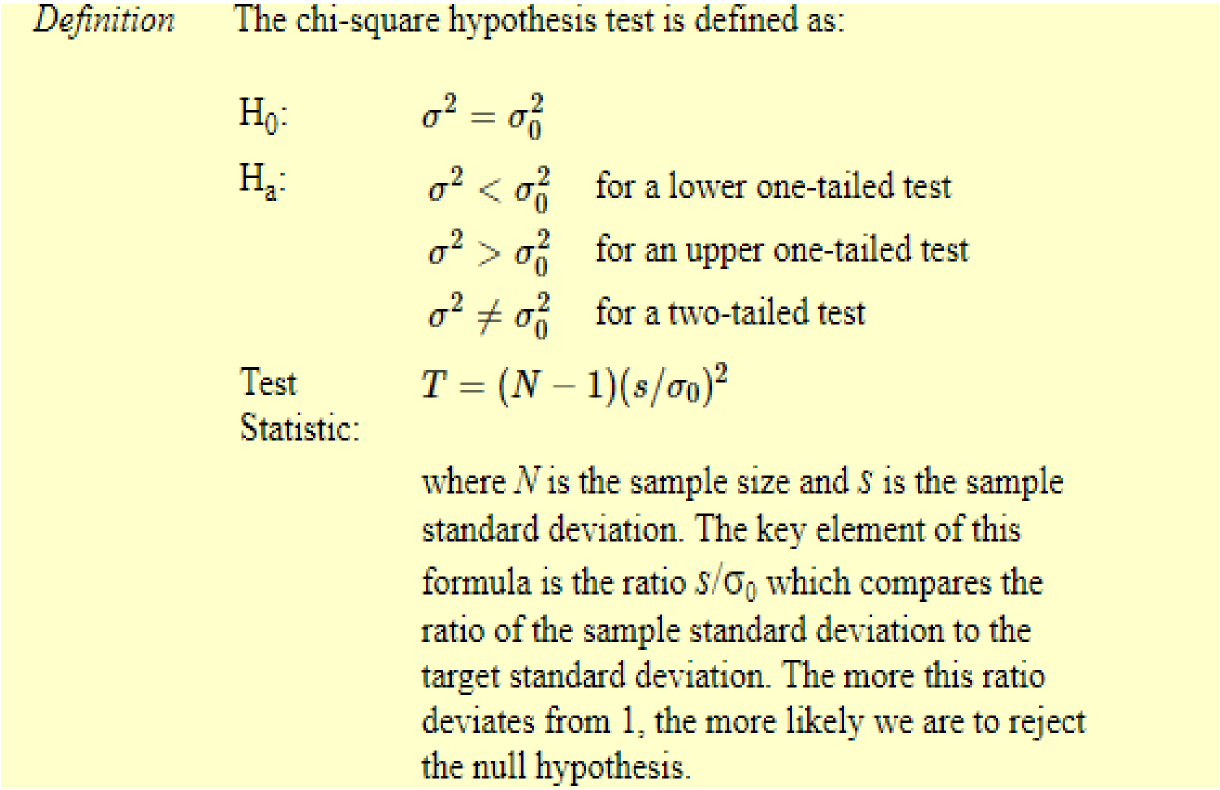


$$E = \frac{\text{row total} \times \text{column total}}{\text{sample size}}$$



We will compare the value of the test statistic to the critical value of χ^2_α with degree of freedom = (*r* - 1) (*c* - 1), and reject the null hypothesis if $\chi^2 > \chi^2_\alpha$.

Type 3: TEST HYPOTHESIS ABOUT VARIANCE



Definition The chi-square hypothesis test is defined as:

H$_0$: $\sigma^2 = \sigma_0^2$

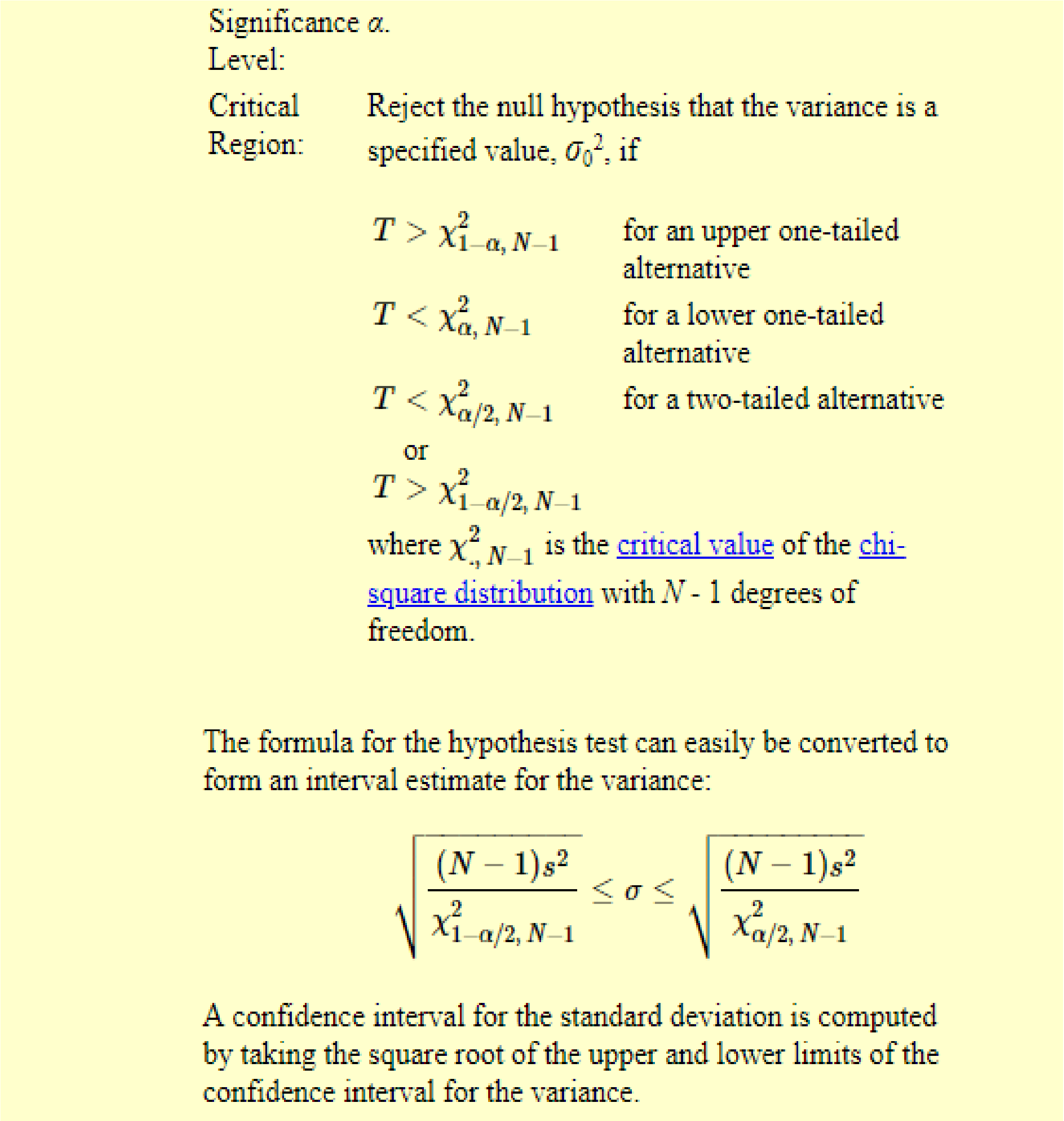
H$_a$: $\sigma^2 < \sigma_0^2$ for a lower one-tailed test

$\sigma^2 > \sigma_0^2$ for an upper one-tailed test

$\sigma^2 \neq \sigma_0^2$ for a two-tailed test

Test Statistic: $T = (N-1)(s/\sigma_0)^2$

where *N* is the sample size and *s* is the sample standard deviation. The key element of this formula is the ratio s/σ_0 which compares the ratio of the sample standard deviation to the target standard deviation. The more this ratio deviates from 1, the more likely we are to reject the null hypothesis.



Significance Level: α.

Critical Region: Reject the null hypothesis that the variance is a specified value, σ_0^2, if

$T > \chi^2_{1-\alpha,\, N-1}$	for an upper one-tailed alternative
$T < \chi^2_{\alpha,\, N-1}$	for a lower one-tailed alternative
$T < \chi^2_{\alpha/2,\, N-1}$ or $T > \chi^2_{1-\alpha/2,\, N-1}$	for a two-tailed alternative

where $\chi^2_{\cdot,\, N-1}$ is the critical value of the chi-square distribution with N - 1 degrees of freedom.

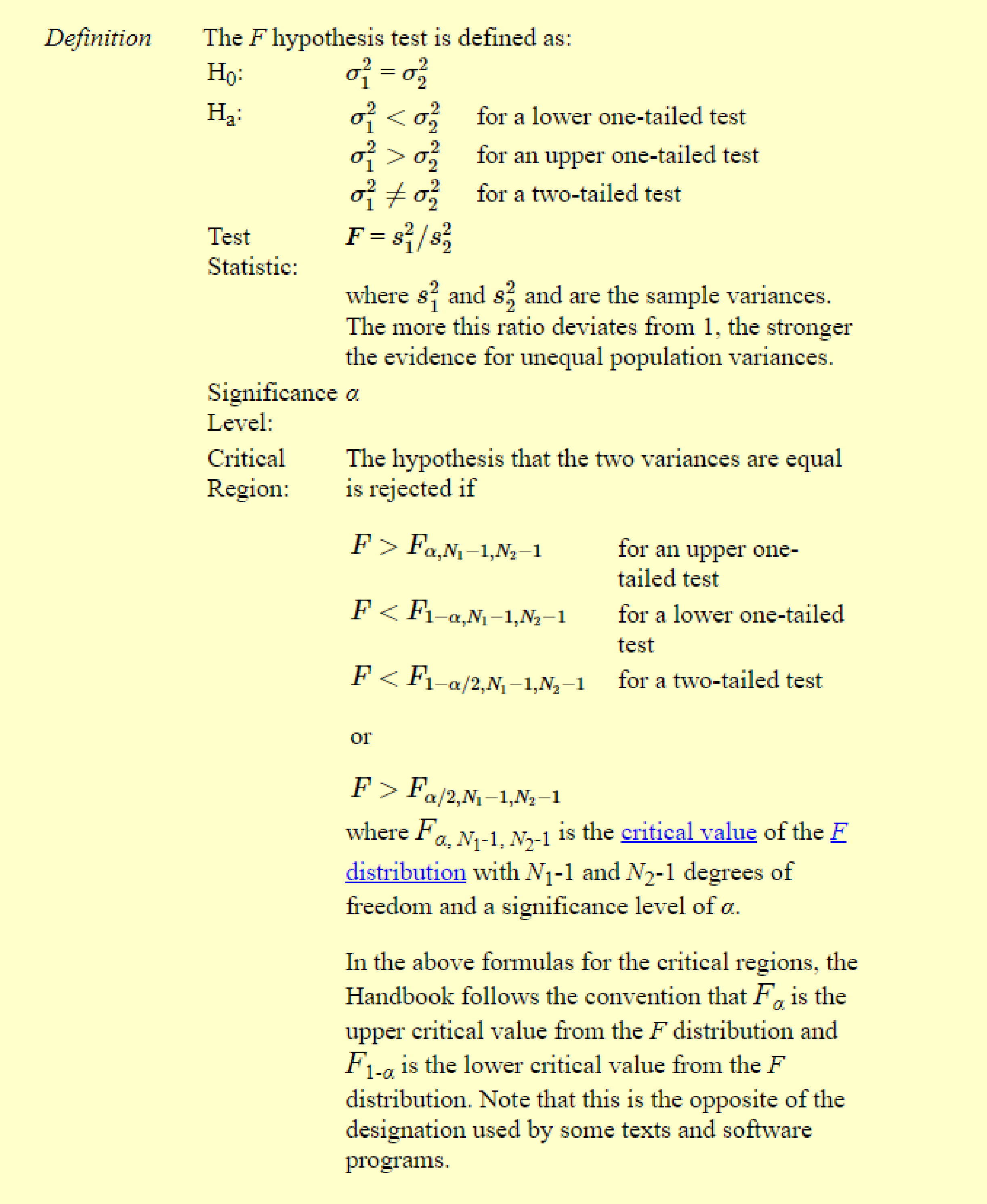
The formula for the hypothesis test can easily be converted to form an interval estimate for the variance:

$$\sqrt{\frac{(N-1)s^2}{\chi^2_{1-\alpha/2,\, N-1}}} \le \sigma \le \sqrt{\frac{(N-1)s^2}{\chi^2_{\alpha/2,\, N-1}}}$$

A confidence interval for the standard deviation is computed by taking the square root of the upper and lower limits of the confidence interval for the variance.

2.2.4 F-DISTRIBUTION TEST

An *F*-test is used to test if the variances of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the variances are not equal. The one-tailed version only tests in one direction, that is the variance from the first population is either greater than or less than (but not both) the second population variance. The choice is determined by the problem.



Definition The F hypothesis test is defined as:

H$_0$: $\sigma_1^2 = \sigma_2^2$

H$_a$: $\sigma_1^2 < \sigma_2^2$ for a lower one-tailed test

$\sigma_1^2 > \sigma_2^2$ for an upper one-tailed test

$\sigma_1^2 \neq \sigma_2^2$ for a two-tailed test

Test Statistic: $F = s_1^2/s_2^2$

where s_1^2 and s_2^2 and are the sample variances. The more this ratio deviates from 1, the stronger the evidence for unequal population variances.

Significance Level: α

Critical Region: The hypothesis that the two variances are equal is rejected if

$F > F_{\alpha,N_1-1,N_2-1}$ for an upper one-tailed test

$F < F_{1-\alpha,N_1-1,N_2-1}$ for a lower one-tailed test

$F < F_{1-\alpha/2,N_1-1,N_2-1}$ for a two-tailed test

or

$F > F_{\alpha/2,N_1-1,N_2-1}$

where $F_{\alpha, N_1\text{-}1, N_2\text{-}1}$ is the critical value of the F distribution with N_1-1 and N_2-1 degrees of freedom and a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that F_α is the upper critical value from the F distribution and $F_{1\text{-}\alpha}$ is the lower critical value from the F distribution. Note that this is the opposite of the designation used by some texts and software programs.

2.2.5 ANOVA

The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.

A pharmaceutical company tested 3 formulations of a migraine relief drug. 27 volunteers were randomly grouped in 3 groups. Each group was given a different drug formulation. The participants took the drug when they had the next migraine attack and recorded the pain on a scale of 1 to 10, 1 being no pain and 10 being extreme pain 30 minutes after taking the medicine.

We want to understand if the differences are due to within group differences or between group differences.

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

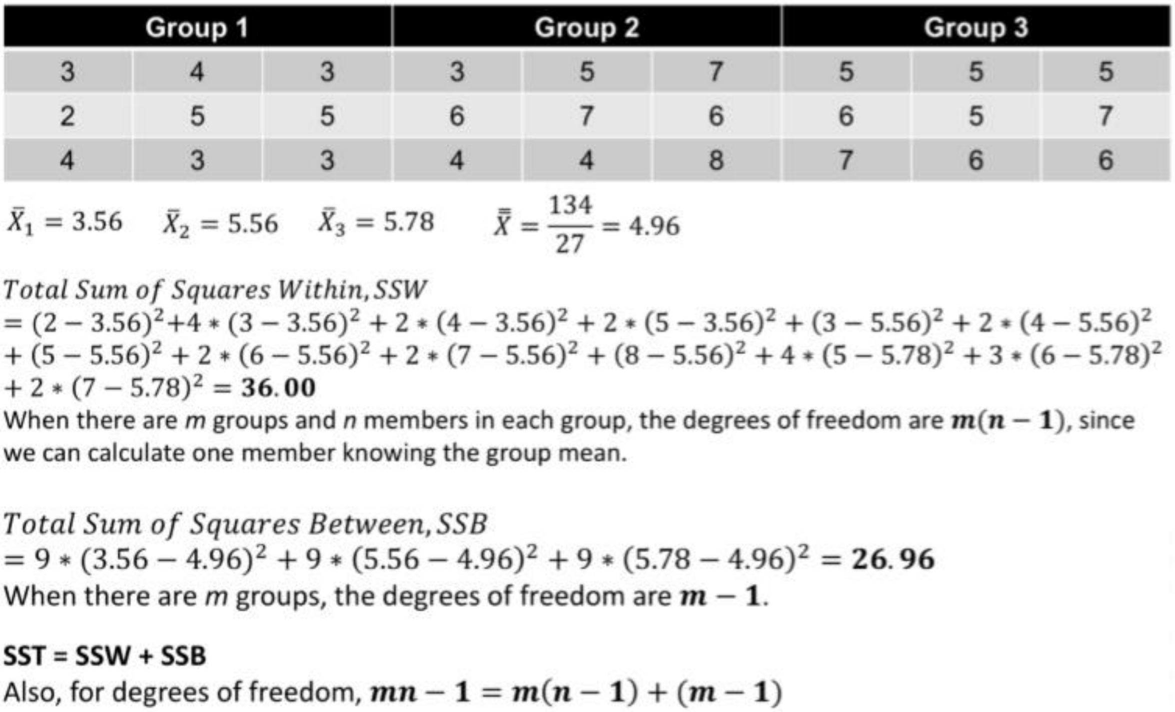
$\bar{X}_1 = 3.56$ $\bar{X}_2 = 5.56$ $\bar{X}_3 = 5.78$

$$\bar{\bar{X}} = \frac{134}{27} = 4.96$$

$$\begin{aligned}&\mathit{Total\ Sum\ of\ Squares, SST}\\&= (2-4.96)^2 + 5*(3-4.96)^2 + 4*(4-4.96)^2 + 7*(5-4.96)^2 + 5*(6-4.96)^2\\&+ 4*(7-4.96)^2 + (8-4.96)^2 = \mathbf{62.96}\end{aligned}$$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{mn-1}$, since we can calculate one member knowing the overall mean.

How much of this variation is coming from within the groups and how much from between the groups?



Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56 \quad \bar{X}_2 = 5.56 \quad \bar{X}_3 = 5.78 \qquad \bar{\bar{X}} = \frac{134}{27} = 4.96$

$$Total\ Sum\ of\ Squares\ Within, SSW$$
$$= (2-3.56)^2 + 4*(3-3.56)^2 + 2*(4-3.56)^2 + 2*(5-3.56)^2 + (3-5.56)^2 + 2*(4-5.56)^2 + (5-5.56)^2 + 2*(6-5.56)^2 + 2*(7-5.56)^2 + (8-5.56)^2 + 4*(5-5.78)^2 + 3*(6-5.78)^2 + 2*(7-5.78)^2 = \mathbf{36.00}$$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{m(n-1)}$, since we can calculate one member knowing the group mean.

$$Total\ Sum\ of\ Squares\ Between, SSB$$
$$= 9*(3.56-4.96)^2 + 9*(5.56-4.96)^2 + 9*(5.78-4.96)^2 = \mathbf{26.96}$$

When there are *m* groups, the degrees of freedom are $\boldsymbol{m-1}$.

SST = SSW + SSB

Also, for degrees of freedom, $\boldsymbol{mn-1 = m(n-1) + (m-1)}$

What is the null hypothesis?

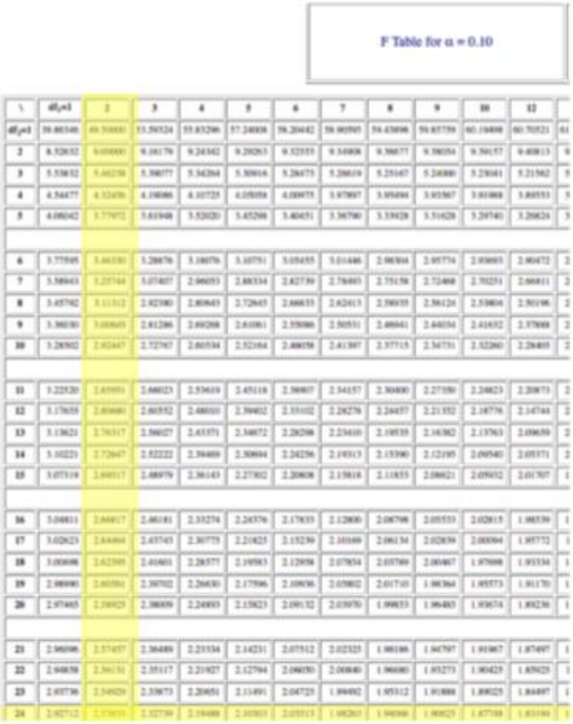
The population means of the 3 groups from which the samples were taken have the same mean, i.e., the drug formulations do not have an impact on relieving migraine headache. $\mu_1 = \mu_2 = \mu_3$. Let us also have a significance level, $\alpha = 0.10$.

What is the alternate hypothesis?

The drug formulations have an impact on migraine pain relief.

$$F - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.96}{2}}{\frac{36}{24}} = 8.9876$$

If numerator is much bigger than the denominator, it means variation **between** means has bigger impact than variation **within**, thus rejecting the null hypothesis.



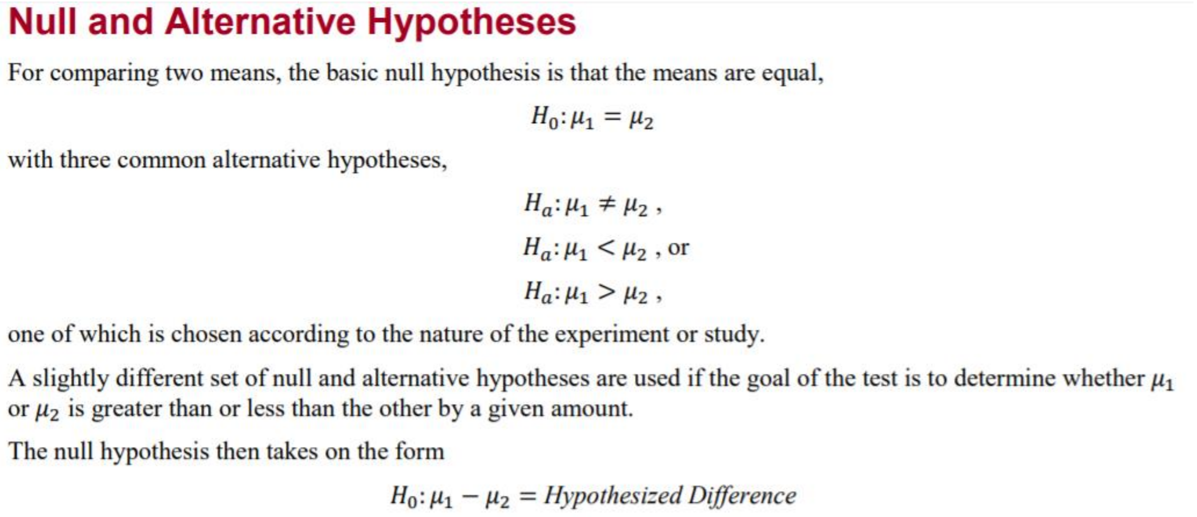
F Table for α = 0.10

\	df1=1	2	3	4	5	6	7	8	9	10	12
df2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50196
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632	2.37888
10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731	2.32260	2.28405
11	3.22520	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.30400	2.27350	2.24823	2.20873
12	3.17655	2.80680	2.60552	2.48010	2.39402	2.33102	2.28278	2.24457	2.21352	2.18776	2.14744
13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.23410	2.19535	2.16382	2.13763	2.09659
14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.15390	2.12195	2.09540	2.05371
15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621	2.05932	2.01707
16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.12800	2.08798	2.05533	2.02815	1.98539
17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839	2.00094	1.95772
18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467	1.97698	1.93334
19	2.98990	2.60561	2.39702	2.26630	2.17596	2.10936	2.05802	2.01710	1.98364	1.95573	1.91170
20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.03970	1.99853	1.96485	1.93674	1.89236
21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797	1.91967	1.87497
22	2.94858	2.56131	2.35117	2.21927	2.12794	2.06050	2.00840	1.96680	1.93273	1.90425	1.85925
23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888	1.89025	1.84497
24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625	1.87748	1.83194

The *df* are 2 for numerator and 24 for denominator.

F_c, the critical F-statistic, therefore, is 2.53833. 8.9876 is way higher than this and hence we reject the null hypothesis. That means the drug formulations do have an impact on migraine pain relief.

2.3 TWO VARIABLE T- TESTS FOR EQUAL AND UNEQUAL VARIANCES



Null and Alternative Hypotheses

For comparing two means, the basic null hypothesis is that the means are equal,

$$H_0\colon \mu_1 = \mu_2$$

with three common alternative hypotheses,

$$H_a\colon \mu_1 \neq \mu_2 \ ,$$

$$H_a\colon \mu_1 < \mu_2 \ , \text{or}$$

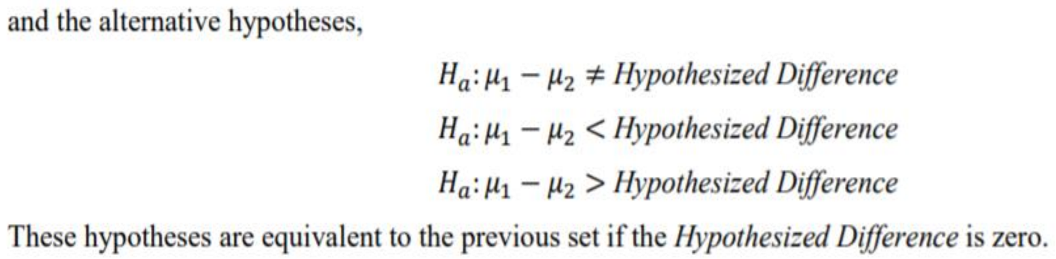
$$H_a\colon \mu_1 > \mu_2 \ ,$$

one of which is chosen according to the nature of the experiment or study.

A slightly different set of null and alternative hypotheses are used if the goal of the test is to determine whether μ_1 or μ_2 is greater than or less than the other by a given amount.

The null hypothesis then takes on the form

$$H_0\colon \mu_1 - \mu_2 = \textit{Hypothesized Difference}$$



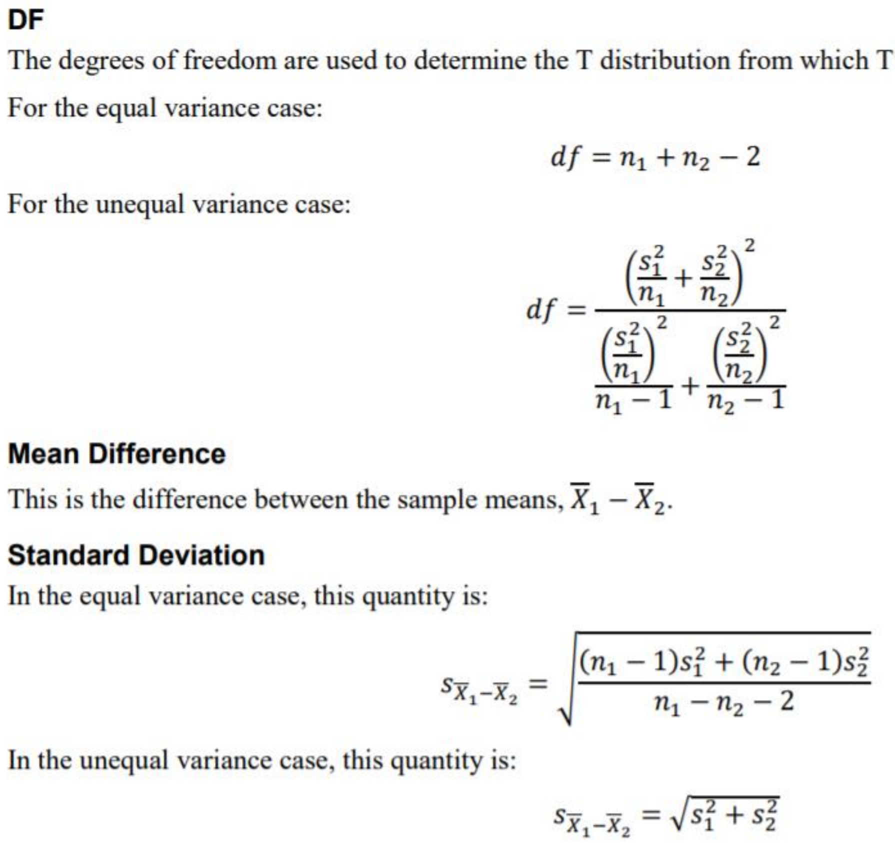
and the alternative hypotheses,

$$H_a\colon \mu_1 - \mu_2 \neq \textit{Hypothesized Difference}$$

$$H_a\colon \mu_1 - \mu_2 < \textit{Hypothesized Difference}$$

$$H_a\colon \mu_1 - \mu_2 > \textit{Hypothesized Difference}$$

These hypotheses are equivalent to the previous set if the *Hypothesized Difference* is zero.



DF

The degrees of freedom are used to determine the T distribution from which T* is generated.

For the equal variance case:

$$df = n_1 + n_2 - 2$$

For the unequal variance case:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Mean Difference

This is the difference between the sample means, $\overline{X}_1 - \overline{X}_2$.

Standard Deviation

In the equal variance case, this quantity is:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - n_2 - 2}}$$

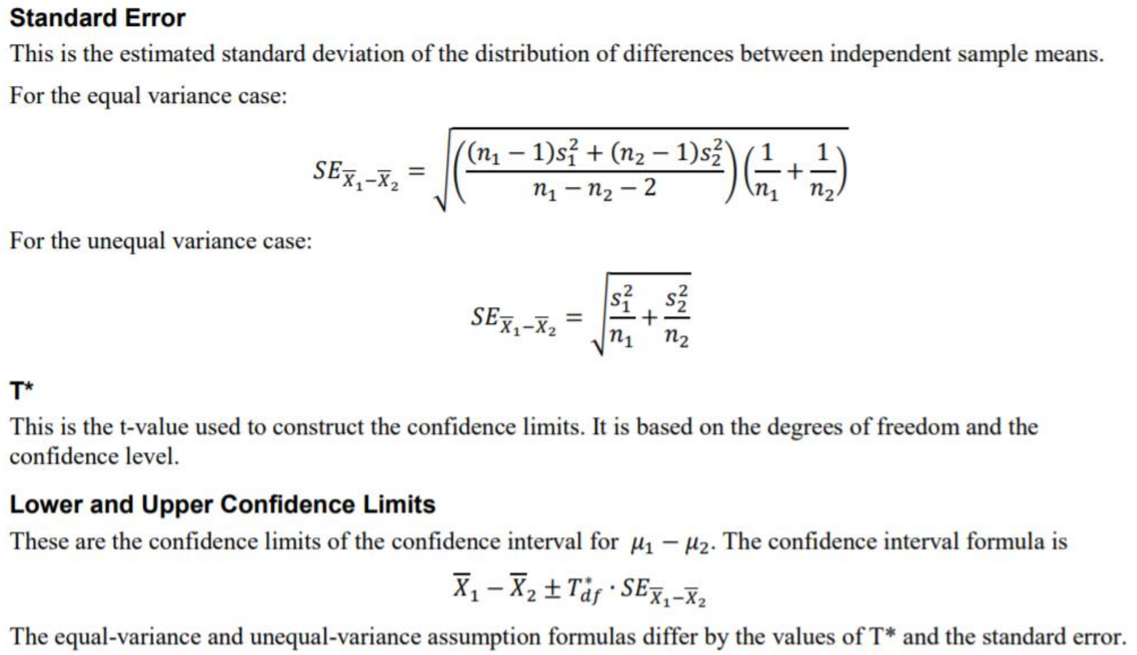
In the unequal variance case, this quantity is:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s_1^2 + s_2^2}$$

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2 - Hypothesized\ Difference}{SE_{\overline{X}_1 - \overline{X}_2}}$$

For unequal variance this is the formula for t- statistic but for equal vaiances the Hypothesized difference = 0 and so the equation becomes

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2}{SE_{\overline{X}_1 - \overline{X}_2}}$$



Standard Error

This is the estimated standard deviation of the distribution of differences between independent sample means.

For the equal variance case:

$$SE_{\overline{X}_1-\overline{X}_2} = \sqrt{\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1-n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

For the unequal variance case:

$$SE_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

T*

This is the t-value used to construct the confidence limits. It is based on the degrees of freedom and the confidence level.

Lower and Upper Confidence Limits

These are the confidence limits of the confidence interval for $\mu_1 - \mu_2$. The confidence interval formula is

$$\overline{X}_1 - \overline{X}_2 \pm T^*_{df} \cdot SE_{\overline{X}_1-\overline{X}_2}$$

The equal-variance and unequal-variance assumption formulas differ by the values of T* and the standard error.

Two-Sample T-Test Assumptions

The assumptions of the two-sample t-test are:

1. The data are continuous (not discrete).

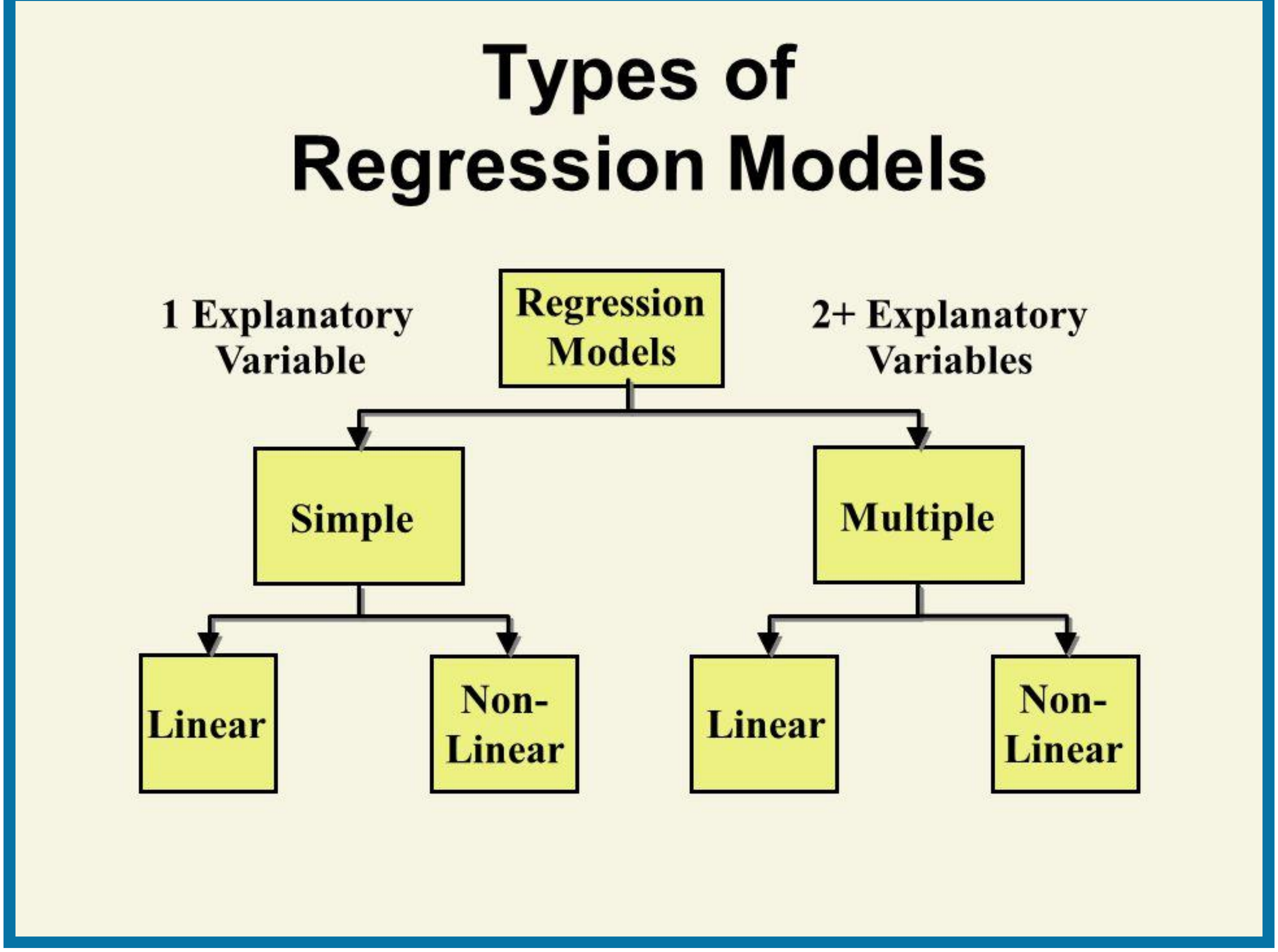
2. The data follow the normal probability distribution.

3. The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)

4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired t-test).

5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

3. TYPES OF REGRESSION MODELS



Types of Regression Models

1 Explanatory Variable

Regression Models

2+ Explanatory Variables

Simple

Multiple

Linear

Non-Linear

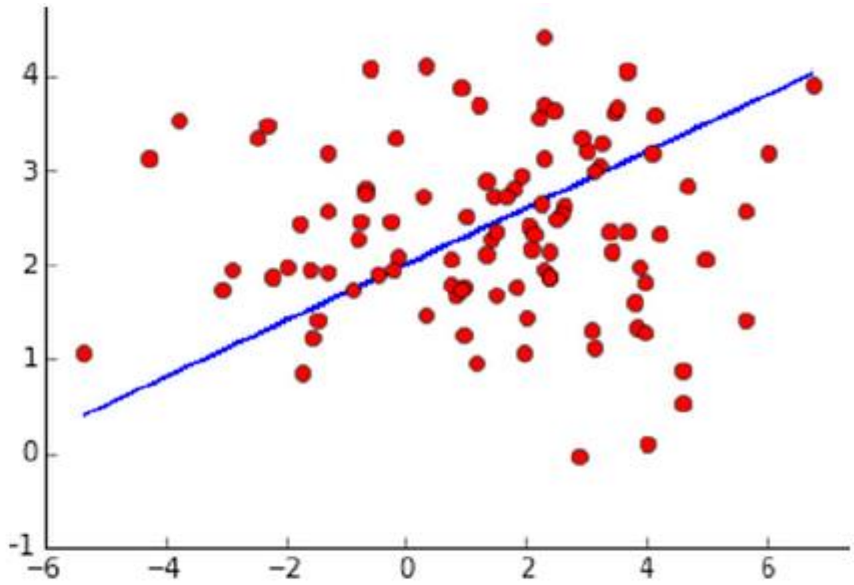
Linear

Non-Linear

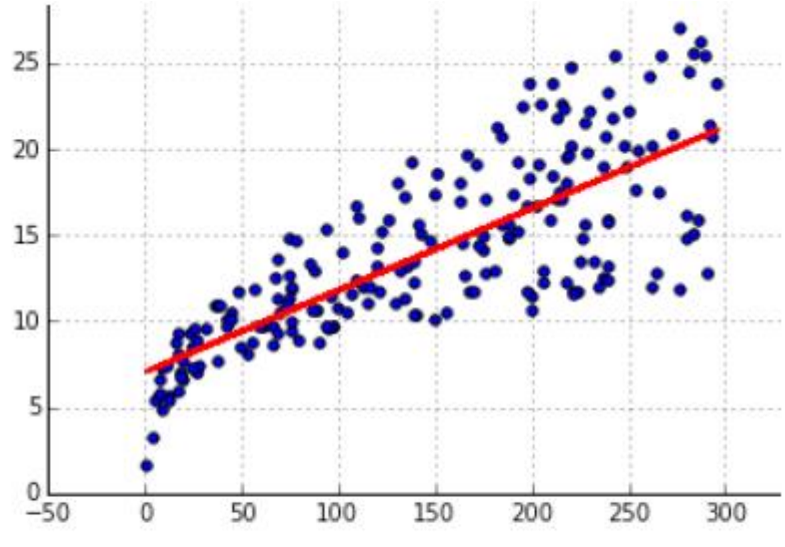
3.1 SIMPLE LINEAR REGRESSION

In simple words linear regression is predicting the value of a variable Y(dependent variable) based on some variable X(independent variable) provided there is a linear relationship between X and Y.

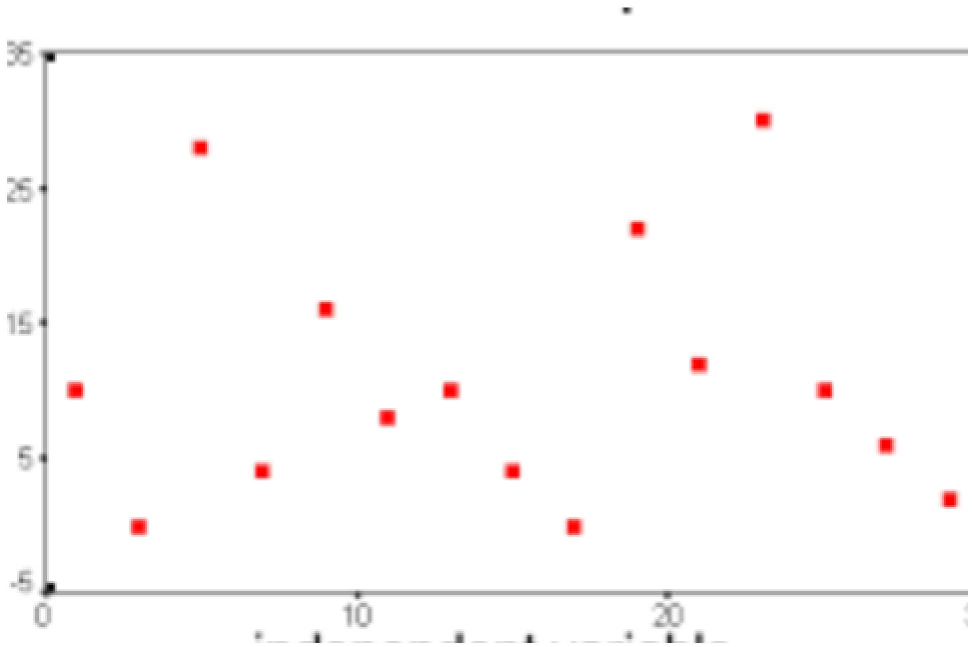
This linear relationship between the 2 variables can be represented by a straight line (called ***regression line***).



Now to determine if there is a linear relationship between 2 variables we can simply plot a scatter plot of variable Y with variable X .If the plotted points are randomly scattered that it can be inferred that the variables are not related.

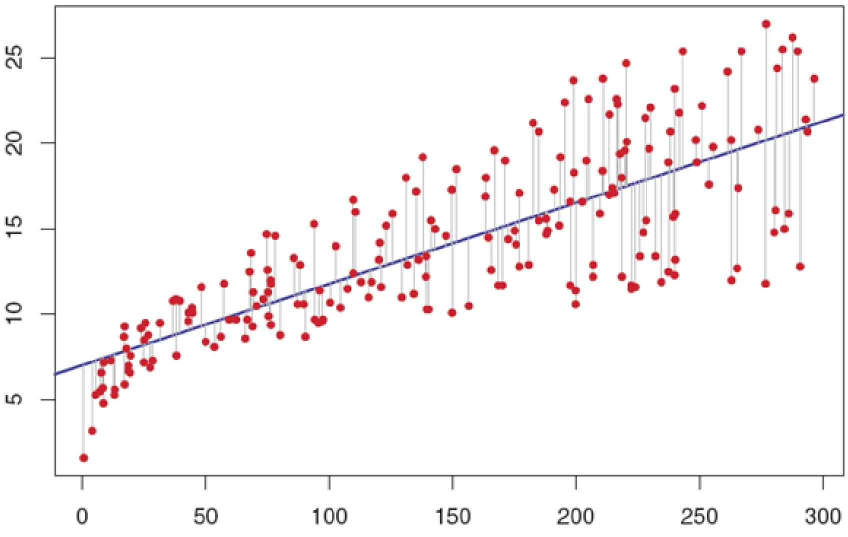


There is linear relationship between the variables. Good model.



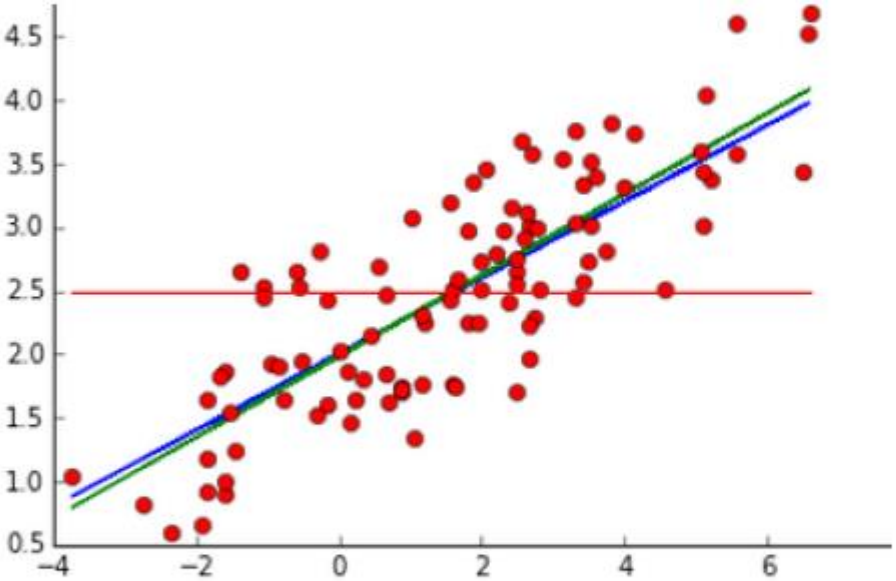
There is no linear relationship between the variables. Bad model.

When regression line is drawn some points will lie on the regression line other points will lie in the close vicinity of it. This is because our regression line is a **probabilistic model** and our prediction is approximate. So there will be some errors/deviations from actual/observed value of variable Y.



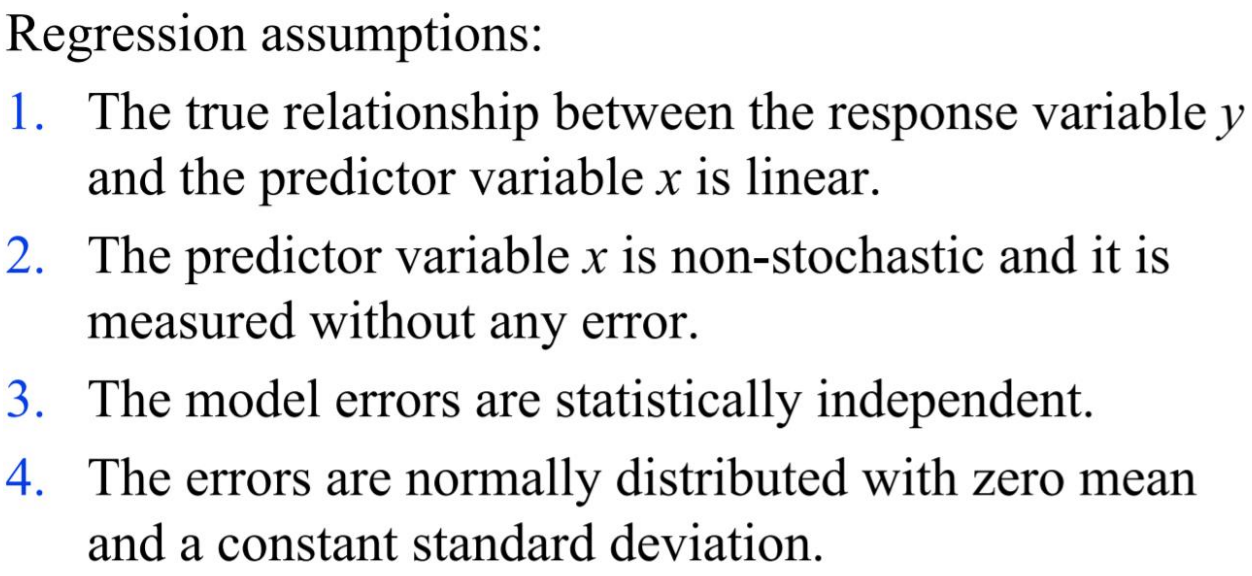
Deviations from actual point to the drawn regression line

But when the linear relationship exist between X and Y we can plot more than one line through these points. Now how do we know which one is the best fit?



To help us choose the best line we use the concept of “least squares”.

3.1.1 REGRESSION ASSUMPTIONS



Regression assumptions:

1. The true relationship between the response variable y and the predictor variable x is linear.
2. The predictor variable x is non-stochastic and it is measured without any error.
3. The model errors are statistically independent.
4. The errors are normally distributed with zero mean and a constant standard deviation.

3.1.2 LEAST SQUARES

Choose the line that minimizes the sum of squares of the errors.

$$\hat{y} = b_0 + b_1 x$$

where, $\hat{y}$ is the predicted response when the predictor variable is x. The parameter b_0 and b_1 are fixed regression parameters to be determined from the data.

Given n observation pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the estimated response $\hat{y}_i$ for the ith observation is:

$$\hat{y}_i = b_0 + b_1 x_i$$

The error is:

$$e_i = y_i - \hat{y}_i$$

The best linear model minimizes the sum of squared errors (SSE):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Regression parameters that give minimum error variance are:

$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x}$$

where,

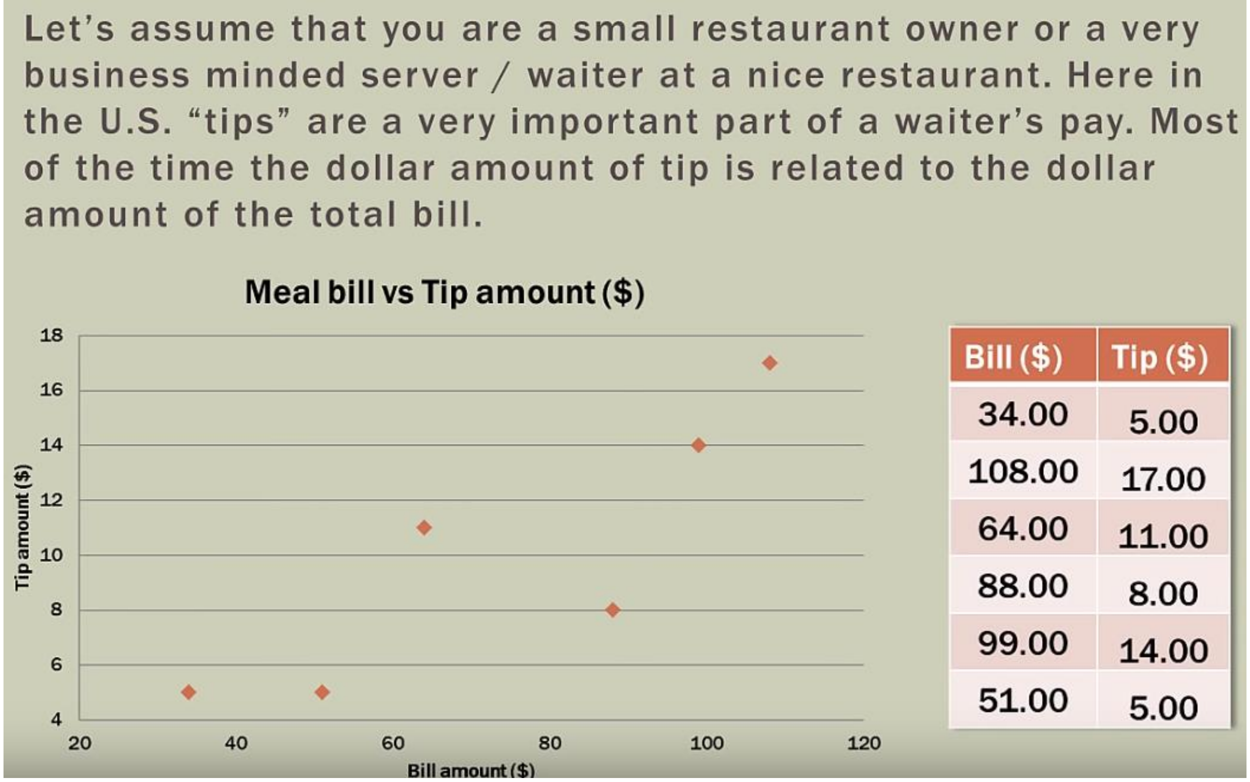
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\Sigma xy = \sum_{i=1}^{n} x_i y_i \qquad \Sigma x^2 = \sum_{i=1}^{n} x_i^2$$

Positive linear relationship so slope is positive, If there is a negative relationship then slope b0 is negative.

3.1.3 DRAWING THE REGRESSION LINE

Lets consider drawing the regression line with an example.

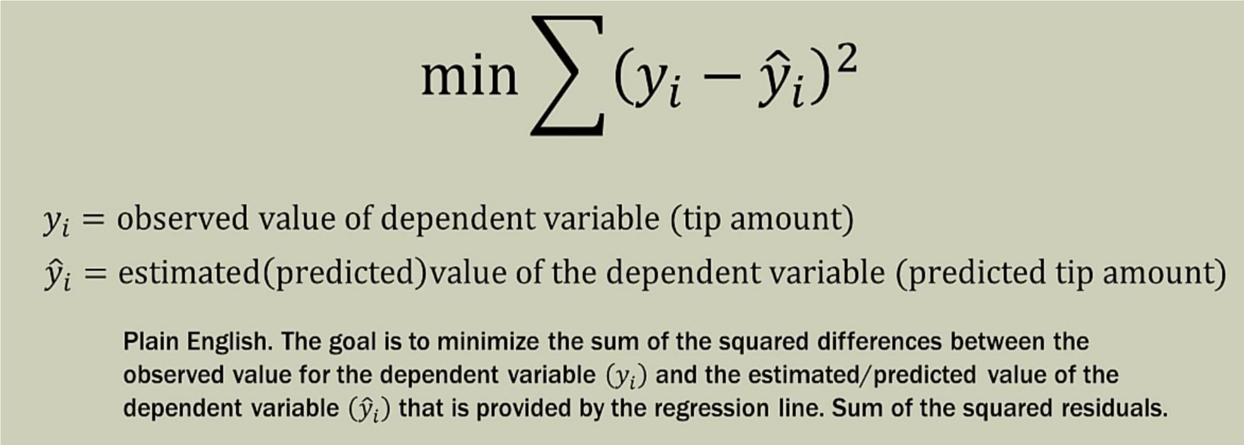


Let’s assume that you are a small restaurant owner or a very business minded server / waiter at a nice restaurant. Here in the U.S. “tips” are a very important part of a waiter’s pay. Most of the time the dollar amount of tip is related to the dollar amount of the total bill.

Meal bill vs Tip amount ($)

Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

LEAST SUM OF SQUARES CRITERION

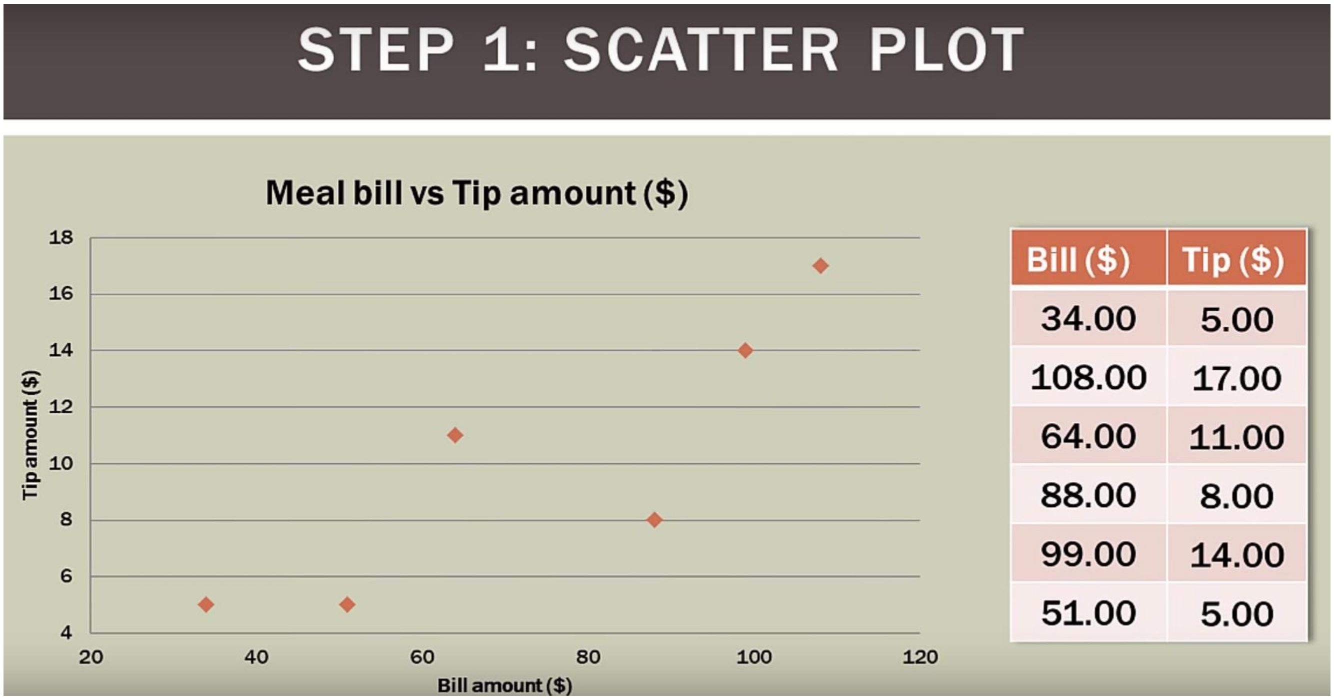


$$\min \sum (y_i - \hat{y}_i)^2$$

y_i = observed value of dependent variable (tip amount)

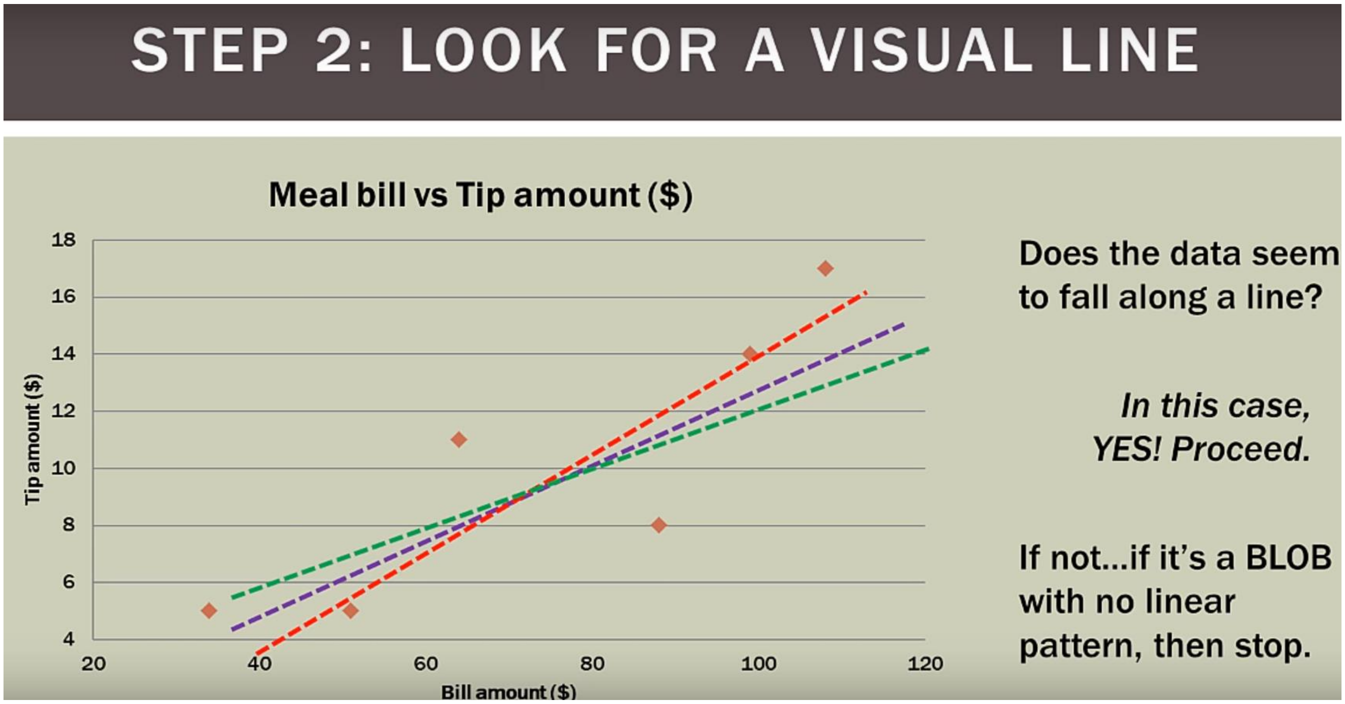
$\hat{y}_i$ = estimated(predicted)value of the dependent variable (predicted tip amount)

Plain English. The goal is to minimize the sum of the squared differences between the observed value for the dependent variable (y_i) and the estimated/predicted value of the dependent variable $(\hat{y}_i)$ that is provided by the regression line. Sum of the squared residuals.



STEP 1: SCATTER PLOT

Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00



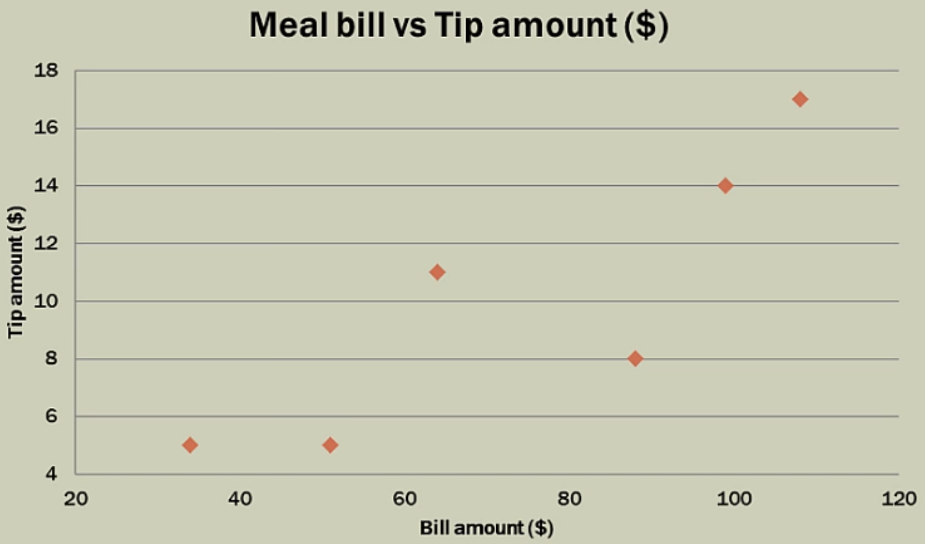
STEP 2: LOOK FOR A VISUAL LINE

Does the data seem to fall along a line?

In this case, YES! Proceed.

If not…if it's a BLOB with no linear pattern, then stop.

STEP 3: CORRELATION (OPTIONAL)



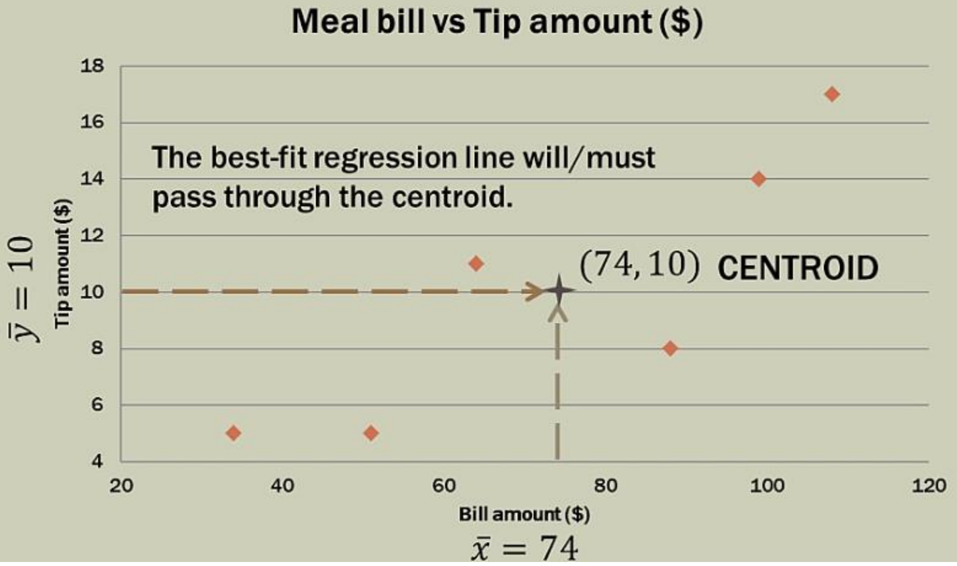
What is the correlation coefficient, r?

In this case,
$r = 0.866.$

Is the relationship strong?

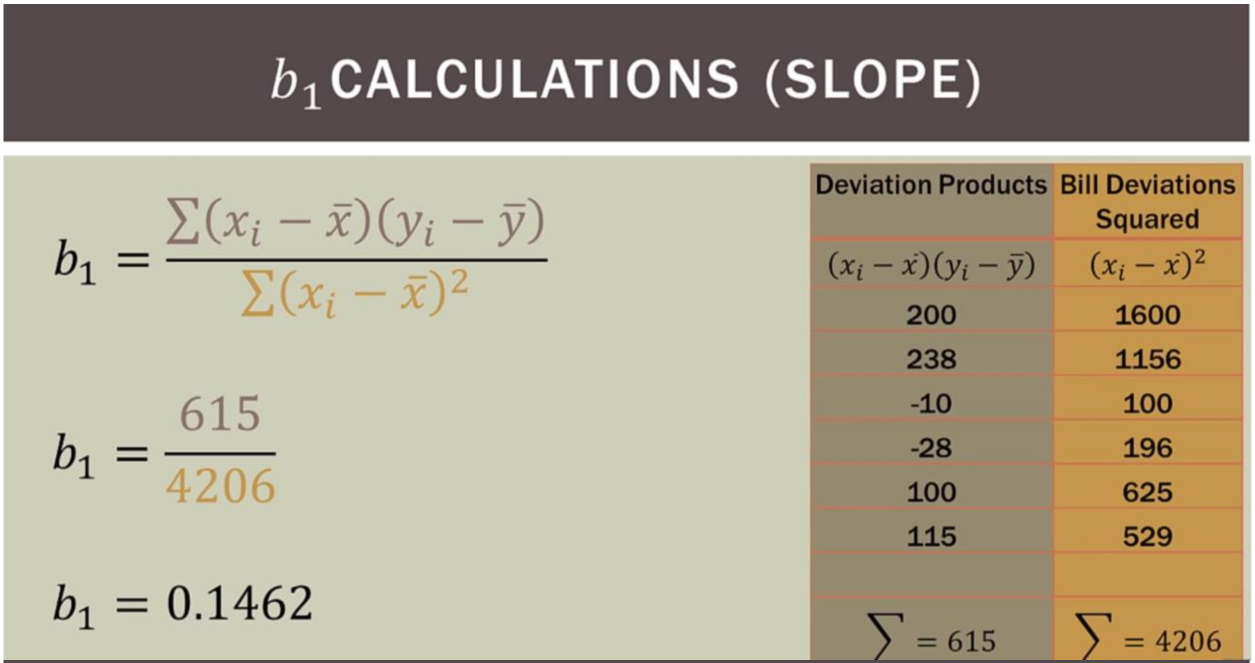
In this case,
YES

STEP 4: DESCRIPTIVE STATISTICS / CENTROID



The best-fit regression line will/must pass through the centroid.

Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$



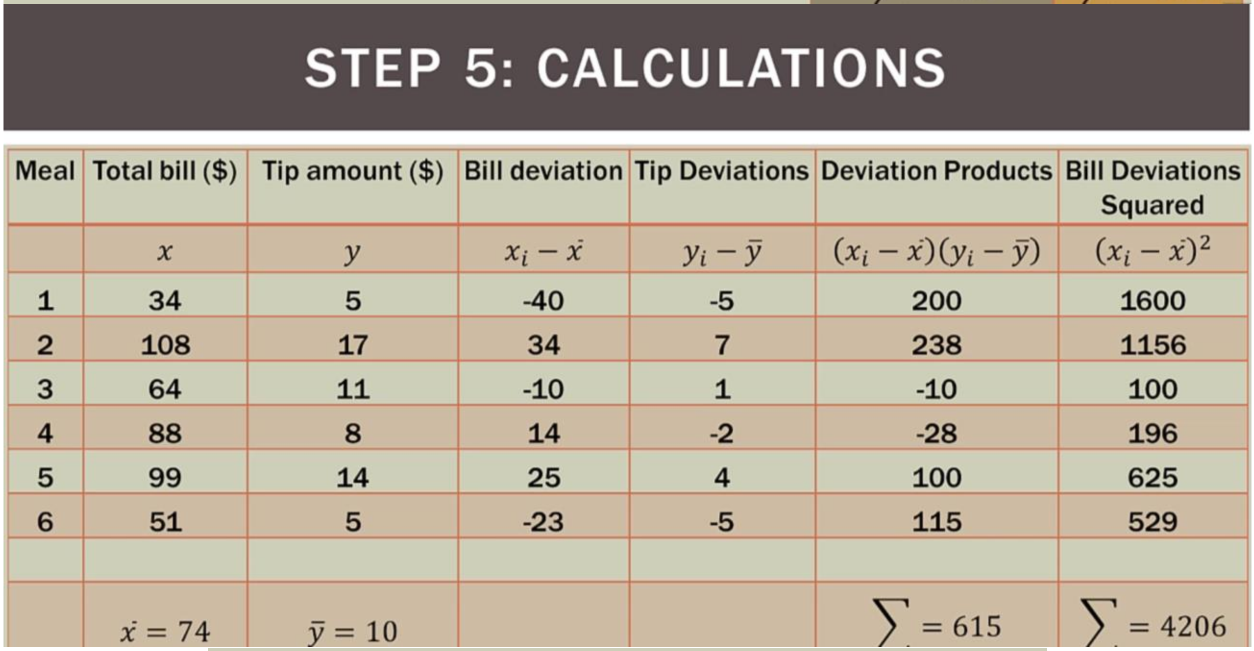
b_1 CALCULATIONS (SLOPE)

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_1 = \frac{615}{4206}$$

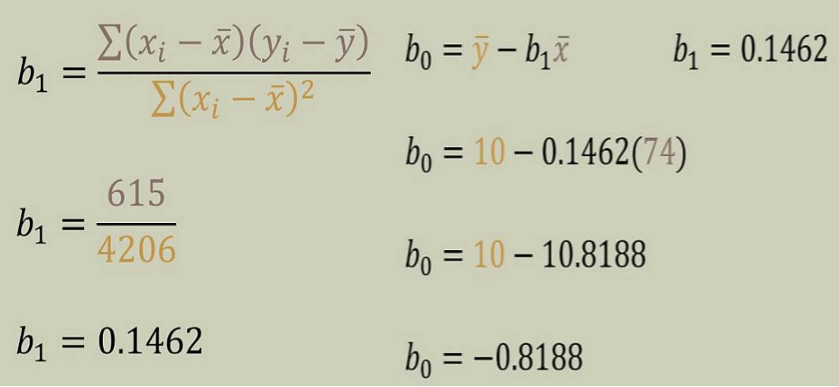
$$b_1 = 0.1462$$

Deviation Products	Bill Deviations Squared
$(x_i - \dot{x})(y_i - \bar{y})$	$(x_i - \dot{x})^2$
200	1600
238	1156
-10	100
-28	196
100	625
115	529
$\sum = 615$	$\sum = 4206$



STEP 5: CALCULATIONS

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviations Squared
	x	y	$x_i - \dot{x}$	$y_i - \bar{y}$	$(x_i - \dot{x})(y_i - \bar{y})$	$(x_i - \dot{x})^2$
1	34	5	-40	-5	200	1600
2	108	17	34	7	238	1156
3	64	11	-10	1	-10	100
4	88	8	14	-2	-28	196
5	99	14	25	4	100	625
6	51	5	-23	-5	115	529
	$\dot{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$



$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = 0.1462$$

$$b_0 = 10 - 0.1462(74)$$

$$b_1 = \frac{615}{4206}$$

$$b_0 = 10 - 10.8188$$

$$b_1 = 0.1462$$

$$b_0 = -0.8188$$

YOUR REGRESSION LINE

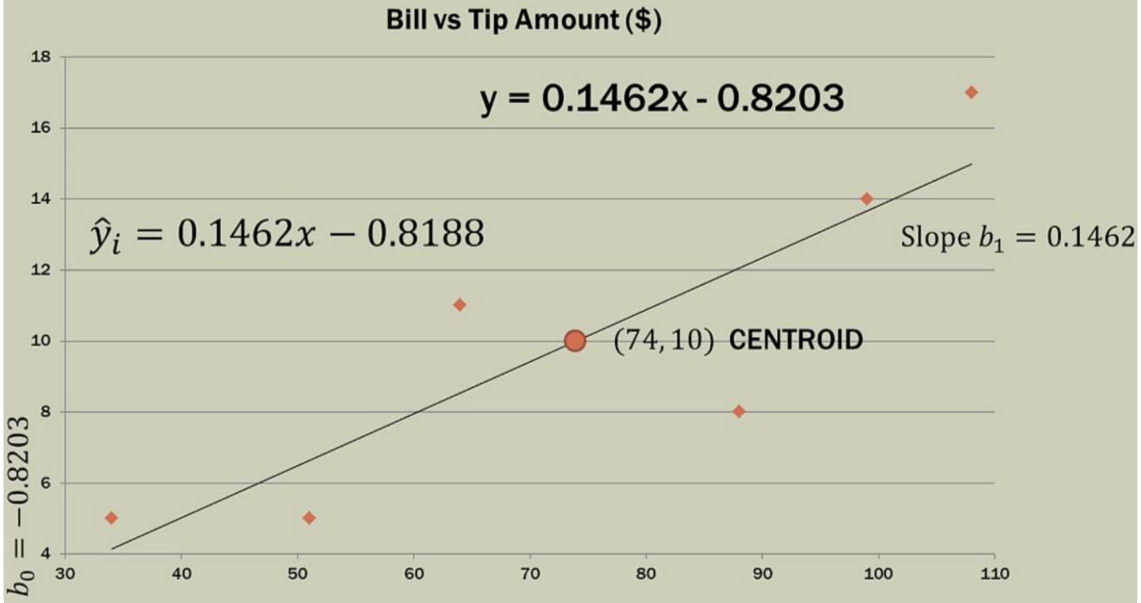
$\hat{y}_i = b_0 + b_1 x_i$ $\qquad b_0 = -0.8188$ $\qquad b_1 = 0.1462$

intercept $\qquad$ slope

$$\hat{y}_i = -0.8188 + 0.1462x$$

OR

$$\hat{y}_i = 0.1462x - 0.8188$$



$$\hat{y}_i = 0.1462x - 0.8188$$

For every $1 the bill amount (x) increases, we would expect the tip amount to increase by $0.1462 or about 15-cents.

If the bill amount (x) is zero, then the expected/predicted tip amount is $-0.8188 or negative 82-cents! Does this make sense? NO. The intercept may or may not make sense in the "real world.

3.1.4 ERRORS

$$\text{SST} = \text{SSR} + \text{SSE}$$

Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

$$\text{SST} = \sum(\text{Y}_\text{i} - \overline{\text{Y}})^2 \quad \text{SSR} = \sum(\hat{\text{Y}}_\text{i} - \overline{\text{Y}})^2 \quad \text{SSE} = \sum(\text{Y}_\text{i} - \hat{\text{Y}}_\text{i})^2$$

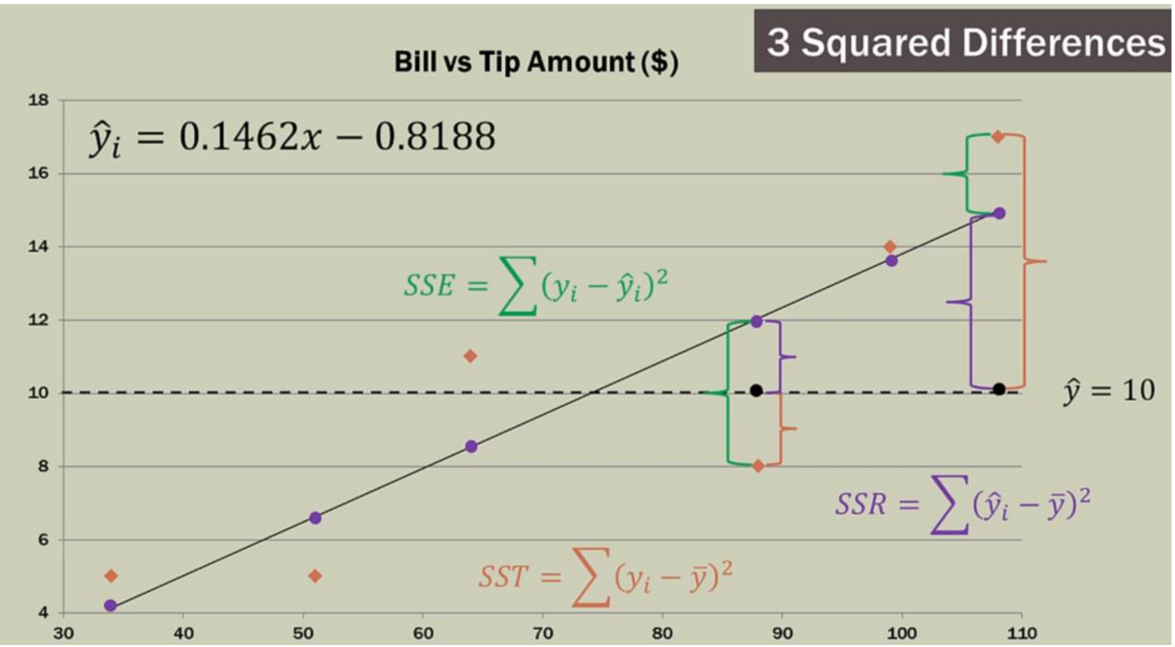
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

Where R^2, $0 < R^2 < 1$.

MSE = $\frac{\text{SSE}}{n}$, Where MSE stands for mean square error

n is the no of sample size.

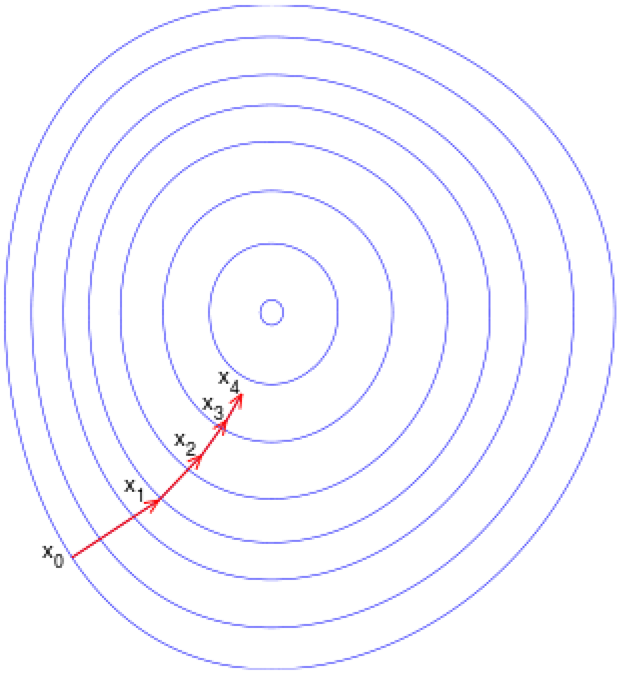
Lets now see these errors in our example:



Where the cost function here is the SSE

3.2 GRADIENT DESCENT

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the *negative* of the gradient (or approximate gradient) of the function at the current point. If instead one takes steps proportional to the *positive* of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent.



3.3 POLYNOMIAL REGRESSION

Sometimes, a plot of the residuals versus a predictor may suggest there is a nonlinear relationship. One way to try to account for such a relationship is through a **polynomial regression** model. Such a model for a single predictor, *X*, is:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_hX_h+\epsilon,$$

where *h* is called the **degree** of the polynomial. For lower degrees, the relationship has a specific name (i.e., *h* = 2 is called **quadratic**, *h* = 3 is called **cubic**, *h* = 4 is called **quartic**, and so on). Although this model allows for a nonlinear relationship between *Y* and *X*, polynomial regression is still considered linear regression since it is linear in the regression coefficients, $\beta_1,\beta_2,\ldots,\beta_h$!

3.3.1 UNDERFITTING, RIGHT FIT AND OVERFITTING

Let us consider that we are designing a machine learning model. A model is said to be a good machine learning model, if it generalizes any new input data from the problem domain in a proper way. This helps us to make predictions in the future data, that data model has never seen.

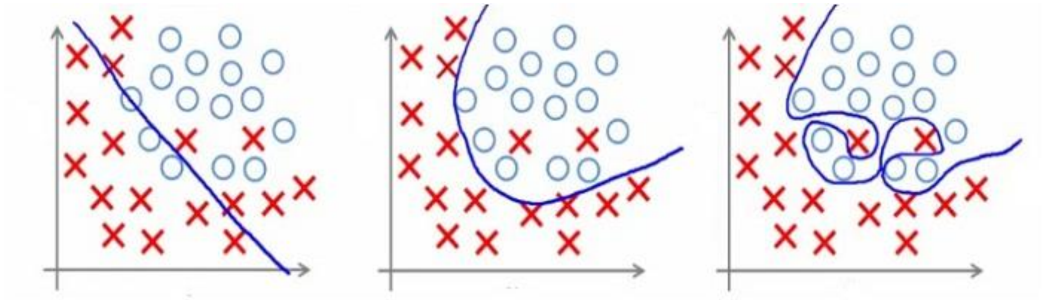
Now, suppose we want to check how well our machine learning model learns and generalizes to the new data. For that we have overfitting and underfitting, which are majorly responsible for the poor performances of the machine learning algorithms.

Underfitting:

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. *(It’s just like trying to fit undersized pants!)* Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.

Overfitting:

A statistical model is said to be overfitted, when we train it with a lot of data *(just like fitting ourselves in an oversized pants!)*. When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too much of details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data.



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

The commonly used methodologies are:

- **Cross- Validation:** A standard way to find out-of-sample prediction error is to use 5-fold cross validation.
- **Early Stopping:** Its rules provide us the guidance as to how many iterations can be run before learner begins to over-fit.
- **Pruning:** Pruning is extensively used while building related models. It simply removes the nodes which add little predictive power for the problem in hand.
- **Regularization:** It introduces a cost term for bringing in more features with the objective function. Hence it tries to push the coefficients for many variables to zero and hence reduce cost term.

3.3.2 MULTICOLLINEARITY

Multicollinearity is the occurrence of high intercorrelations among independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model. In general, multicollinearity can lead to wider confidence intervals and less reliable probability values (P values) for the independent variables.

3.3.4 REGULARIZATION

If we increase the number of terms in data then we should penalize the correlation coefficient so that it does not overfit.

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, ***this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.***

A simple relation for linear regression looks like this. Here Y represents the learned relation and *β represents the coefficient estimates for different variables or predictors(X).*

Y ≈ β0 + β1X1 + β2X2 + …+ βpXp

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Now, this will adjust the coefficients based on your training data. *If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero*

Ridge Regression:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Above image shows ridge regression, where the ***RSS is modified by adding the shrinkage quantity.*** Now, the coefficients are estimated by minimizing this function. Here, ***λ is the tuning parameter that decides how much we want to penalize the flexibility of our model.*** The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept β0, This intercept is a measure of the mean value of the response when xi1 = xi2 = …= xip = 0.

When λ = 0, the penalty term has no effect, and the estimates produced by ridge regression will be equal to least squares. However, ***as λ→∞, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero***. As can be seen, selecting a good value of λ is critical. Cross validation comes in handy for this purpose. The coefficient estimates produced by this method are ***also known as the L2 norm***.

Ridge reduces all the non- significant columns.

Lasso Regression:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Lasso is another variation, in which the above function is minimized. Its clear that *this variation differs from ridge regression only in penalizing the high coefficients*. It uses |βj|(modulus)instead of squares of β, as its penalty. In statistics, this is *known as the L1 norm*.

Lasso cancels out all the non – significant columns.

Elastic net regression:

It is a hybrid of both Ridge and Lasso Regression which overcomes the limitations of Lasso regression.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Lasso stands for **L**east **A**bsolute **S**hrinkage **S**elector **O**perator

3.4 LOGISTIC REGRESSION

In statistics, the **logistic model** (or **logit model**) is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, **logistic regression** or **logit regression** is estimating the parameters of a logistic model.

The two possible dependent variable values are often labelled as "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick.

3.4.1 ASSUMPTIONS FOR LOGISTIC REGRESSION

First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.

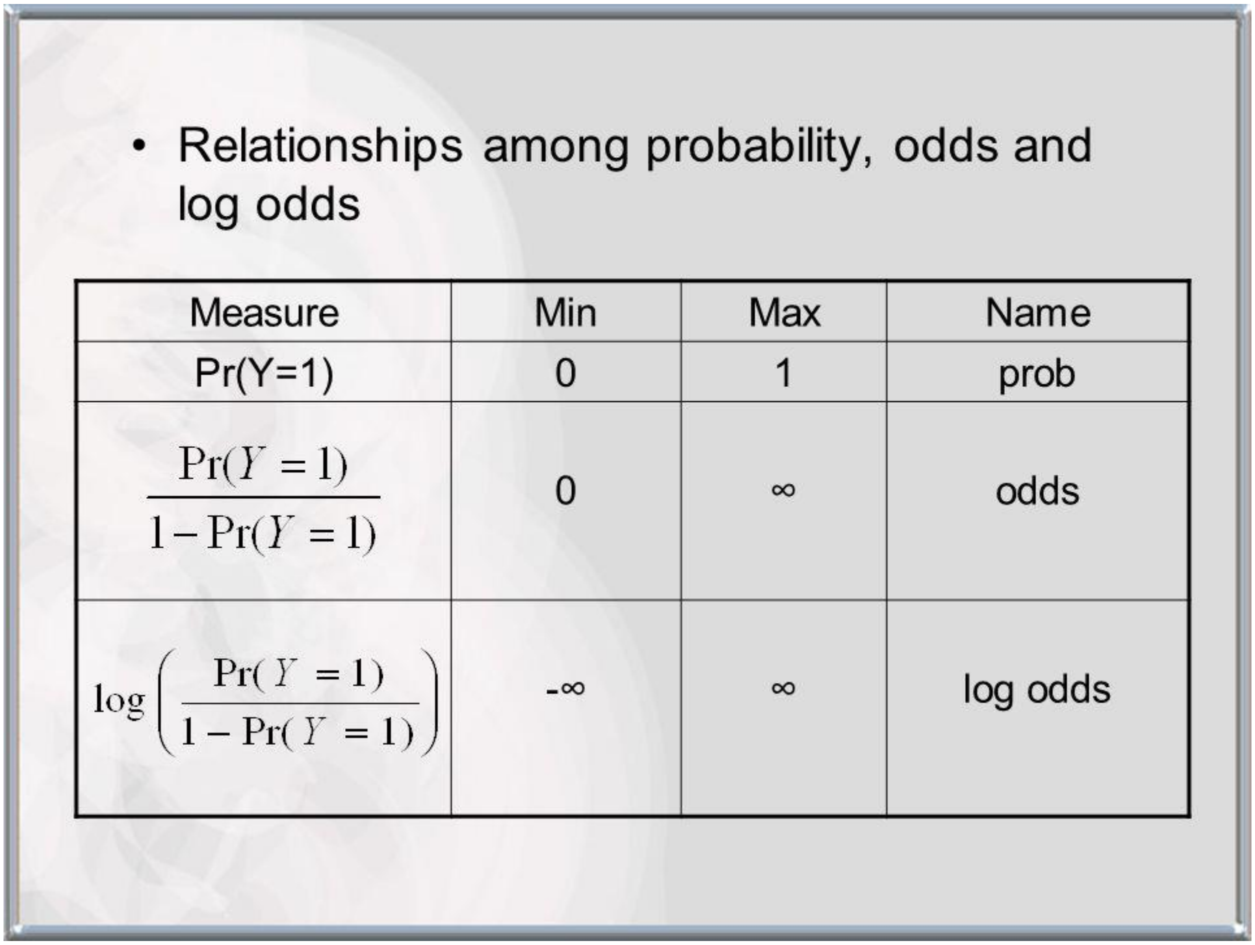
Second, logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.

Third, logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.

Fourth, logistic regression assumes linearity of independent variables and log odds. although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

Finally, logistic regression typically requires a large sample size.

3.4.2 LOGISTIC REGRESSION DEFENITION



- Relationships among probability, odds and log odds

Measure	Min	Max	Name
Pr(Y=1)	0	1	prob
$\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$	0	∞	odds
$\log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right)$	-∞	∞	log odds

We can see from the table that the **ODDS** are always positive.

$$odds = \frac{P}{1-P}$$

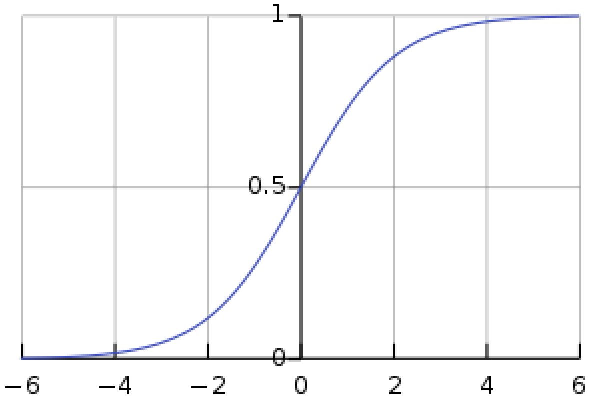
$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find

logit(P) = a + bX,

3.4.2 SIGMOID FUNCTION

A **sigmoid function** is a mathematical function having a characteristic "S"-shaped curve or **sigmoid curve**.



$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \qquad \text{OR} \qquad P = \frac{1}{1+e^{-(a+bX)}}$$

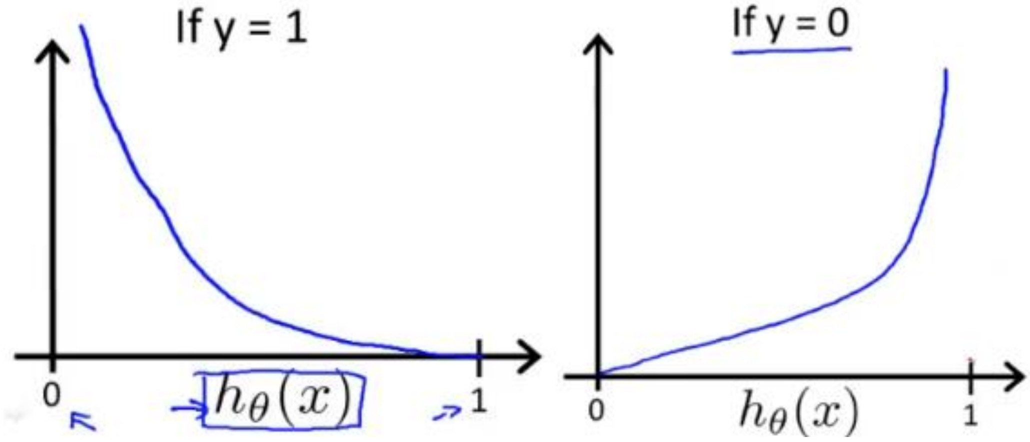
3.4.3 COST FUNCTION

Instead of Mean Squared Error, we use a cost function called Cross-Entropy, also known as Log Loss. Cross-entropy loss can be divided into two separate cost functions: one for y=1 and one for y=0.

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\begin{aligned} \mathrm{Cost}(h_\theta(x), y) &= -\log(h_\theta(x)) & \text{if } y = 1 \\ \mathrm{Cost}(h_\theta(x), y) &= -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{aligned}$$

The benefits of taking the logarithm reveal themselves when you look at the cost function graphs for y=1 and y=0. These smooth monotonic functions (always increasing or always decreasing) make it easy to calculate the minimize cost.



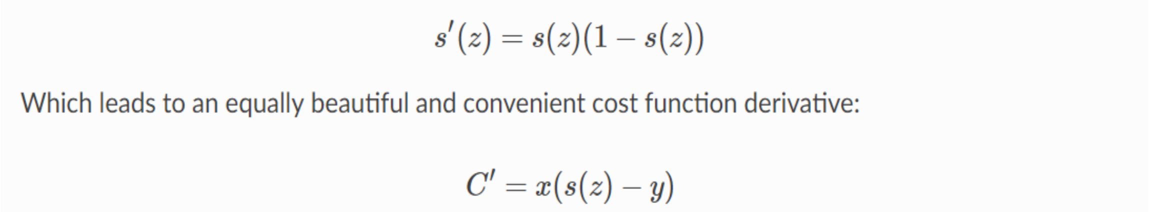
Above functions compressed into one

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

3.4.4 GRADIENT DESCEND

To minimize our cost, we use Gradient Descent just like before in Linear Regression.

One of the neat properties of the sigmoid function is its derivative is easy to calculate.



$$s'(z) = s(z)(1 - s(z))$$

Which leads to an equally beautiful and convenient cost function derivative:

$$C' = x(s(z) - y)$$

ERROR METRICES

4. CONFUSION MATRIX

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix.

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

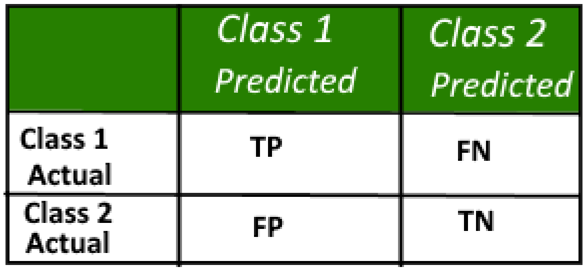
Confusion Matrix:

A confusion matrix is a summary of prediction results on a classification problem.

The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.

It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.



	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Here,

- Class 1: Positive
- Class 2: Negative

4.1 DEFINITION OF THE TERMS WITH EXAMPLE:

- Positive (P) : Observation is positive (for example: is an apple).
- Negative (N) : Observation is not positive (for example: is not an apple).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Business case scenario:

Voltas pvt. Ltd., has signed a contract for mall construction and their deadline comes the next month. They are current;y working on their CCTV panel installation and they should complete this before the end of this week. The Supervisor is responsible for the completion of this part. Now define the error metrices for this case.

True Positive: The service men have arrived and the installation is going as planned. The supervisor reports to his manager the same.

False Positive: The service men have not yet arrived and the work in site is delayed. The supervisor is afraid to tell the exact situation to the manager and tells his manager the work is in progress and everything is all right.

False negative: The service men have arrived and the job is going as planned. But due to some reasons the supervisor reports the manager that the workers have not arrived yet.

True negative: The service men have not yet arrived and the job is delayed. The supervisor reports the actual situation and the manager takes effective to restore the same.

Here if u notice the **FP** makes more damage to the revenue of the company than **FN** which in this case would not affect the company much.

Classification Rate/Accuracy:

Classification Rate or Accuracy is given by the relation:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Recall:

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Precision:

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (small number of FP).

Precision is given by the relation:

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F-measure:

Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$\textbf{\textit{F - measure}} = \frac{2*Recall*Precision}{Recall + Precision}$$

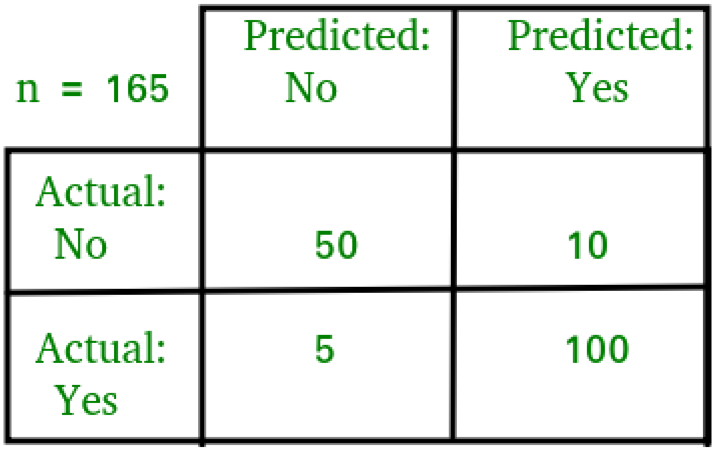
Let’s consider an example now, in which we have infinite data elements of class B and a single element of class A and the model is predicting class A against all the instances in the test data.

Here,

Precision: 0.0
Recall: 1.0

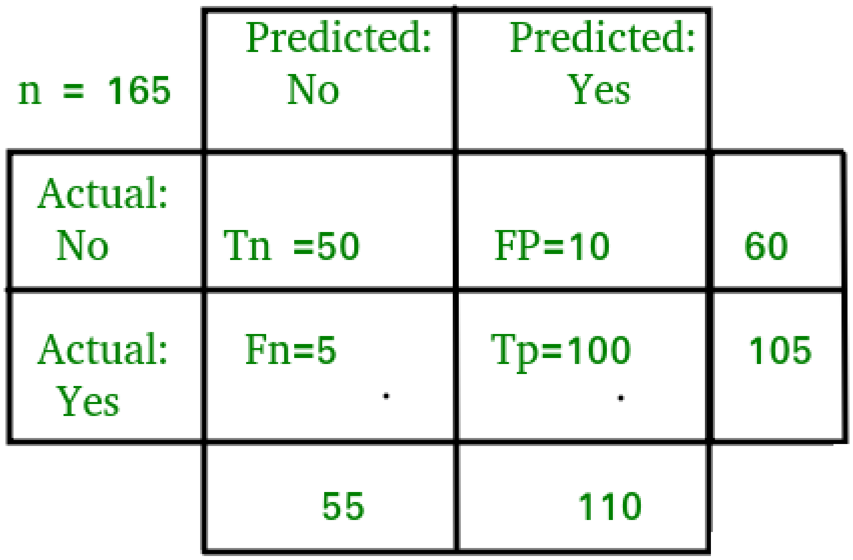
Now:
Arithmetic mean: 0.5
Harmonic mean: 0.0
When taking the arithmetic mean, it would have 50% correct. Despite being the worst possible outcome! While taking the harmonic mean, the F-measure is 0.

Example to interpret confusion matrix:



n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

For the simplification of the above confusion matrix i have added all the terms like TP,FP,etc and the row and column totals in the following image:



n = 165	Predicted: No	Predicted: Yes	
Actual: No	Tn =50	FP=10	60
Actual: Yes	Fn=5	Tp=100	105
	55	110	

Now,

Classification Rate/Accuracy:

Accuracy = (TP + TN) / (TP + TN + FP + FN)= (100+50) /(100+5+10+50)= 0.90

Recall: Recall gives us an idea about when it's actually yes, how often does it predict yes.

Recall=TP / (TP + FN) =100/ (100+5) =0.95

Precision: Precision tells us about when it predicts yes, how often is it correct.

Precision = TP / (TP + FP) =100/ (100+10) =0.91

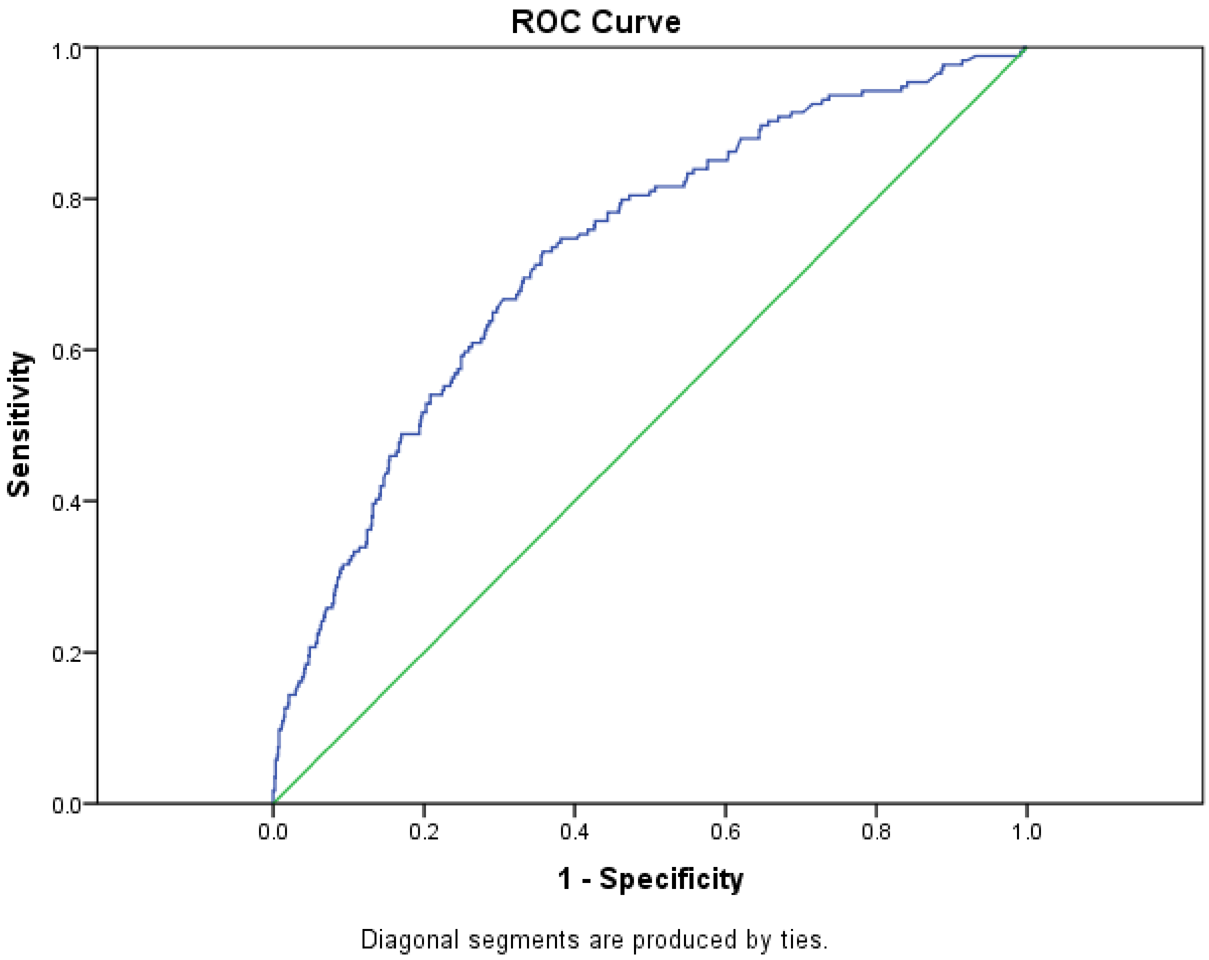
F-measure:

F- Measure=(2*Recall*Precision) / (Recall + Precision) = (2*0.95*0.91)/ (0.91+0.95) =0.92

5. ROC CURVE

An incredibly useful tool in evaluating and comparing predictive models is the ROC curve. ROC stands for Receiver Operating Characteristic.

The ROC curve plots out the sensitivity and specificity for every possible decision rule cutoff between 0 and 1 for a model.



This plot tells you a few different things.

A model that predicts at chance will have an ROC curve that looks like the diagonal green line. That is not a discriminating model.

The further the curve is from the diagonal line, the better the model is at discriminating between positives and negatives in general.

There are useful statistics that can be calculated from this curve, like the Area Under the Curve (AUC) .

Although ROCs are often used for evaluating and interpreting logistic regression models, they’re not limited to logistic regression. A common usage in medical studies is to run a ROC to see how much better a single continuous predictor (a “biomarker”) can predict disease status compared to chance.