INTRODUCTION TO STATISTICS

What is statistics?

Statistics can be understood in two ways, these are plural and single statistics. Plurals statistics is more concerned with numbers, figures and data. Single statistics is understood as techniques and methods involved through major aspects of measurement or statistical analysis.

There are two important key aspects of statistics and these are Descriptive and Inferential statistics. The major understanding of involves both or either of the two aspects.

IMPORTANTANCE OF STATISTICS

Statistical literacy is important and necessary due to the fact that it makes it possible for someone to meaningfully and intelligently read, understand, analyse and evaluate research figures and findings.

An understanding of statistics enables researchers to undertake research on their own with minimum difficulty.

COMMON USES OF STATISTICS

- It involves solving practical problems. This is called practical application. This can be used in research, consultancy, management and many more. The examples of practical application includes;
  Quality control, this is used to test goods before they are released on the market.
- Market research depends on statistics to produce a product that will represent the whole population. In order to know that your adverts have an impact on the market, you have to carry out a research to collect statistics.
  Opinion research, it is not only used in predicting elections popularity, but also in other predictions of margins of error.

ROLES AND FUNCTIONS OF A STATISTICIAN

Statisticians have three major concerns or pre-occupations.

- ✓ Acquisition or collection of data.
- ✓ Selection of the best method for making inferences.
- ✓ Determination or evaluation of the goodness of the inferences.

Acquisition of data: using sample data (sample survey) the statistician knows that by manipulating the sample, it is possible to affect the quantity and cost the information that one wants. It will also have implications. Probability sampling will give high quality data. Larger samples will yield high quality data. However it is more costly.

Selection of the best method for making inferences; after the acquisition of data, the statistician has to find an appropriate method for making inferences. He/she should decide on which test to use to test his output e.g. parametric test or non parametric test.

Determination or evaluation of the goodness of the inferences; in making inferences we are concerned with issues of making the predictions and estimation but also to seek the upper limit. After collecting data and gone through all what has been discussed, there are a number of things you can do with this data. This can be organising, summarising and describing it and make inferences.

DESCRIPTIVE AND INFERENTIAL STATISTICS

Descriptive statistics is concerned with organising, summarising and describing of data. This is also the way of describing data in such a way that you can ascertain the main characteristics with minimum efforts. For instance you can organise data through graphical techniques as well as using numerical techniques. Organising of data when it is in its raw form consists of listing and grouping it in the form of frequency so that you can find out how often a particular value occurs. Then proceed to summarising either graphically or numerically. When summarising graphically you can do this through;

- Pie charts
- Bar graphs
- Histograms
- Frequency histogram or frequency polygons

But if you choose to summarise the data numerically, you can use the measure of central tendency which will bring out the main or major characteristics of the entire set of data numerically. These mostly include the mean, mode and median.

We are also interested in knowing how spread out are the main characteristics and also how near they are to the central position. It is also interested in measures of dispersion. These measures of dispersion include the range, the inter-quartile range (quartile deviation), variance and standard deviation.

MEASURES OF RELATIVE STANDING

Under this, we are interested in measures of relative standing. These include measures like percentile scores (points) / percentiles and percentile ranks.

A well presented table of results makes it possible for the researcher to ascertain the particular characteristic in relation to the other. In other words it makes it possible to ascertain the position of one point in relation to the other in a given score. Describing statistics are the first step before the inferential statistics can be applied. A well applied frequency polygon will indicate the point about a given situation hence will indicate whether the distribution is normal or not.

INFERENTIAL STATISTICS

Inferential statistics is concerned with things like testing the hypothesis, estimating of population value and predictions. It has one important purpose of estimating population values on the basis of

sample value. Sample values are known as statistics. Population values are known as parameters. Its other purpose is that of hypothesis testing.

In inferences of statistics to the parameters we are also interested in knowing how good a statistic is to a parameter. The processing of statistics is based on probability theorem.

LIMITATIONS OF STATISTICS

Statistics are not the solution to all problems. These are just tools to be used in a situation where you have that are agreeable to quantification (e.g. problems reducible to numbers). If you are dealing with problems which are cannot reduced to numbers or qualitative problems then, the use of statistics cannot apply. Statistics can be irrelevant to problems that have qualitative form like those in the form of narrative and not in numerical form.

DESCRIPTIVE STATISTICS

This involves organising and summarising of data as mentioned earlier.

Assume you collect data on UNZA employees trying to study the absenteeism. n=25

Raw Data

37 52 19 48 34 note: this is raw data because it is not organised in order.

41 37 37 23 37

51 31 47 26 42

38 43 36 26 33

41 36 46 38 36

What is supposed to be done to such data is;

First of all, convert the data into some meaningful form/order by placing it in an array. Then, use frequency distribution-ungrouped data or grouped data. Ungrouped simply shows you the number of times a particular observation appears separately. Grouped shows the number of times items appears in groups.

An example of an array is as follows. Start with the least number.

19 23 26 26 31 33 34 36 36 36 37 37 37 37 38 38 41 41 42 43 46 47 48 51 52

FREQUENCY DISTRIBUTION

xi	fi
19	1
23	1
26	2
27	1
31	1
33	1
34	1
36	3
37	3
38	2
41	2
42	1
43	1
46	1
47	1
48	1
51	1
52	1

Therefore ∑fi=25

You have to reduce the number of features by using a form of grouped frequency distribution. If you use class intervals, ensure that classes are exclusive. Also the number of groups created should not too few or too many. Make a judgment to make a reasonable distribution.

GROUPED FREQUENCY DISTRIBUTION

Age Group	fi
15-19	1
20-24	1
25-29	3
30-34	3
35-39	8
40-44	4
45-49	3
50-54	2
	$\sum fi$ = 25

CLASS INTERVALS DETERMINATION

It is vital that the number of intervals is one that guarantees a minimal number of distortions (reduced distortion of information). Also ensure convenience because too many intervals are cumbersome to interpret data (i.e. above we have 8 class intervals).

Decide on the size of the intervals or width of the class interval. What can be done is to use the simple formula, that is, to divide the range by the number of class interval required.

For instance,

Least value = 19

Highest value=52

Difference=33

Therefore 33 divided by 8 =4.125

This means that you will have to have a class interval of four (4).

In demography, we just subtract 15 from 54 and then divide by 8. Then you must round off the value.

THE CONCEPT OF TRUE (REAL) LIMITS AND STATED LIMITS.

Stated Limits, these are limits as given.

True Limits

True limit theoretically is the estimated value that can be assigned to a class interval e.g. 19.999..... And 14.99.....

Simply find the difference between the stated lower limit of an interval (the next interval).

Stated lower of the next interval minus (-) stated upper limit of previous interval. Then divide the answer by 2.

For instance,

15-19

20-24 (20-19) =1. Then ½=0.5

25-29

When this is done subtract the value of 0.5 from all the stated limits and add to the real limits.

True limits extent the boundaries of limits and they bound the stated limits. This construction of true limits has other uses.

True limit remove the uncertainty i.e. people are not exactly 20, 19,12, years old, they have years and months e.g.19.3 years. This will be easily allocated, so that you can get a true picture.

Size of the class interval is the difference between the upper limit and the lower limit, then divide the difference by 2.

REASONS FOR COMPUTINGTRUE LIMITS

- In order to avoid gaps between intervals for continuous data i.e. age, height, weight, etc.
- Avoidance of ambiguity
- Also used in the construction of graphs representing continuous data.
- Also important to ensure that additional accuracy when computing measures like median.

THE CONCEPT OF MID-POINT

The mid-point represents the middle value of a class interval. How to arrive at the mid-point. Add up the lower limit and upper limit then divide by 2.

For example, 15+19 = 34. Then 34/2=17. OR 14.5+19.5=34, $\frac{34}{2}$ =17.

The mid- point is often used in frequency distribution which is grouped. Mid-point is also used to show the difference of occurrences. The mid-point is used in the construction of graphs of frequency polygons.

THE FREQUENCY DISTRIBUTION

Can be thought of in terms of absolute relative intervals which shows actual counts.

Age Group	fi
15-19	1
20-24	1
25-29	3
30-34	8
35-39	4
40-44	3
45-49	2
	25

This information can be presented in relative (%).

Rel % $(^{x}/_{n})$x100, =1/25x100

=3/25x100

=12%

CUMMULATIVE AND DECUMMULATIVE FREQUENCY DISTRIBUTION

Cumulative Frequency Distribution: shows the percentage of the number of observations located below a certain limit. This limit in most cases is invariably true upper limit. Denoted by CF but sometimes it is called the less than distribution. All you have to do is to cumulate the values downwards. For example,

Age Group	ABS	Rel %	CF	%	DCF	%
15-19	1	4	1	4	25	100
20-24	1	4	2	8	24	96
25-29	3	12	5	20	23	92
30-34	3	12	8	32	20	80
35-39	8	32	16	64	12	48

Where,

ABS-Absolute Relative interval

Rel %-Relative percentage

CF- cumulative frequency

In interpreting, you must use true upper limit as the first point of comparison. For instance, if you focus your attention on the give class interval, 16 people are below the age 40 or 39.99.

Decummulative Frequency/ies

Are interested in finding the point of comparison is the true lower limit.

Cumulative values upwards. Decummulative frequencies just like CF but it starts from downwards going up.

GRAPHICAL TECHNIQUES FOR DESCRIBING DATA.

The data that you collect must be organised by presentation. Then you can use any of the following graphical techniques. When you use the graphs ensure that the observations are mutually exclusive data organised in a way that it will categorised in a mutually exclusive manner.

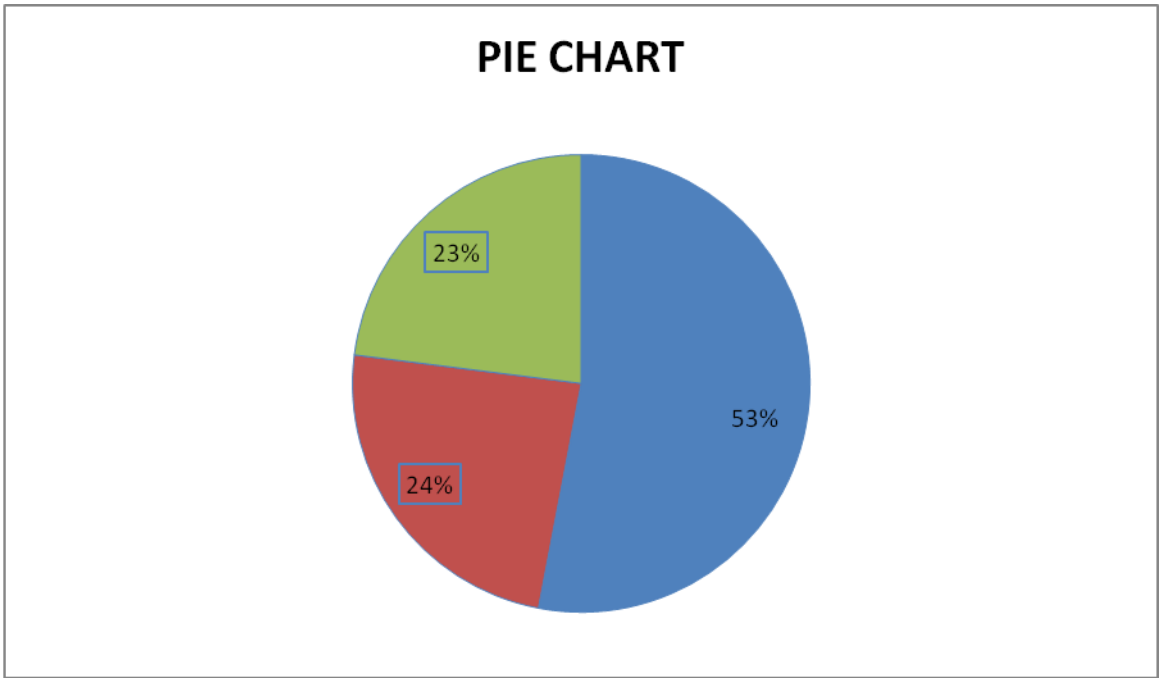
PIE CHART

This provides one of the simplest way of presenting data especially if using qualitative (categorical data) e.g. 1980 census

Town	**Population Size**
Lusaka	538
Kitwe	315
Ndola	282
Total	1,135

The pie chart in most cases display total percentages or numbers of observations falling into each of the categories of the qualitative variables presented in form of a circle, partition into observation/ different categories of the variable. Guideline to be followed when construction of a pie chart.

- Ideally choose a small number of categories advisory use a maximum of six categories.
- Then compute degrees by dividing the number of observation (measurement) in category by the total number of observations (measurement) the multiply by 360^{o}



For instance,

Lusaka 538/1135) x360^{o}=53^{o}

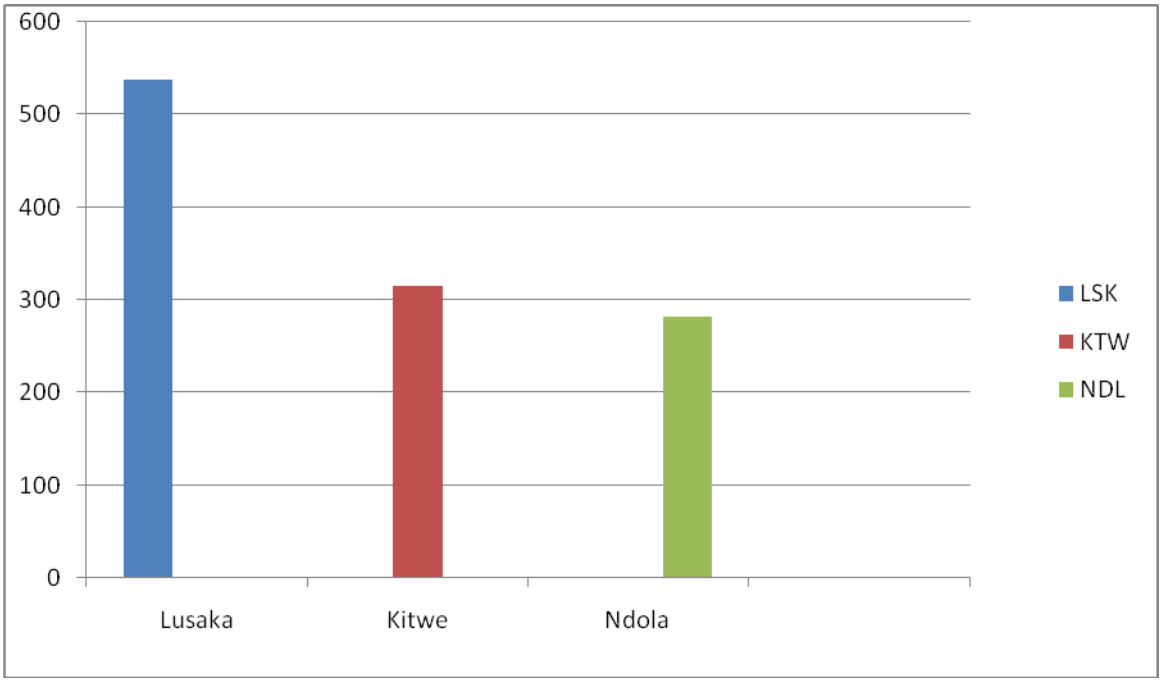
Note; do not use degrees in the pie chart presentation. You have to compute them in percentages.

The Bar Chart

It is also used in organising data in situation where you have quantitative or qualitatative data. It represents data in form of bars.

Construction of Bar Chart

- Label the Frequencies vertically
- Locate the categories of the variable on a horizontal scale.
- Label the frequencies along the vertical axes.
- Use whatever scale is considered appropriate.



Construct a rectangle over each category of the qualitative variable with the height equal to the number of the variable in the category.

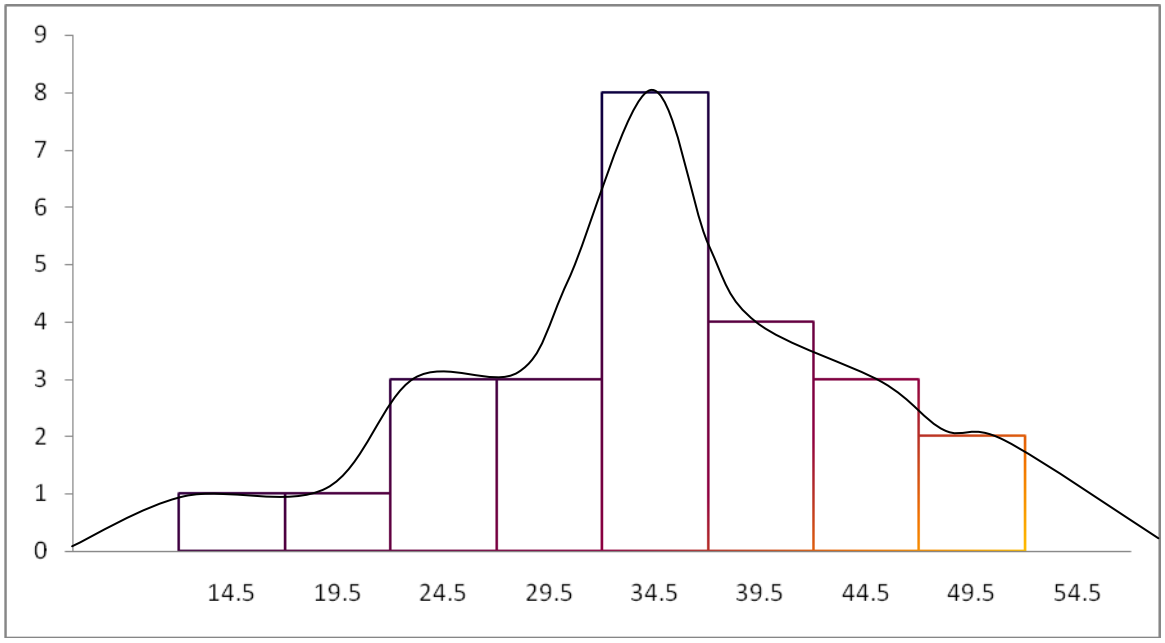
In the construction always leave a space in between each variable to facilitate mutual exclusively of the categories.

FRREQUENCY HISTOGRAM/ POLGON

This is other types of the bars. These types of bars are normally applicable to situations where you are dealing with qualitative and continuous data. Data must be organised before you construct the frequency polygon/ histogram.

Age group	fi	x_2	**%**
15-19	1	17	4
20-24	1	22	4
25-29	3	27	12
30-34	3	32	12
35-39	8	37	32
40-44	4	42	16
45-49	3	47	12
50-54	2	52	8

When constructing the frequency polygon or histogram, you table the frequency along the vertical axis. Then locate the class along the horizontal axis. (Use true limits).



Construct a rectangle over each class interval. Each rectangle's height must be equal to the number of observations.

NUMERICAL METHODS OF DISTRIBING DATA

These methods are used to convey a.....

The numerical methods are also very expatiating or can be used because of expedience. Numerical methods of describing data have an advantage because they use verbal communication to convey a particular picture of a situation.

There are two numerical techniques that are used to;

- Measure of central tendency
- Measure of dispersion or variability

Numerical measures of a population are referred to as parameter where as numerical methods of a sample are referred to as statistics.

N= 598 (pop) 25years average age (parameter)

n= 100 (sample) 24.5 years average age (statistics)

MEASURES OF CENTRAL TENDENCY

Measures of central tendency indicate the main characteristics. These include the mode, mean and median.

MODE

The mode of a central measure is that measurement that occurs most often (with highest frequency).

EXAMPLE

Students who drink beer x times n=25

7	10	8	11	9
9	9	8	9	8
9	9	9	8	9
8	8	9	10	11
10	7	10	9	7

Rearrange the raw data in order of assertion.

7	8	9	9	10
7	8	9	9	10
7	8	9	9	11
8	8	9	9	11
8	9	9	10	11

x_i	f_i
7	3
8	6
9	10
10	4
11	2

MEDIAN

This is simply a middle value when the measurements are arranged in order you can have an even number of observation as odd number. When the measurements are arranged in order, the median is simply the mid value or point.

7 8 9 9 10

7 8 9 9 10

7 8 9 9 10 Median = 9

8 8 9 9 11

8 9 9 10 11

Aptitude reading scores is 10.

Another example is as follows;

95, 86, 78, 90, 62, 73, 81, 82, 84, 76. Arrange in order from the lowest score as follows;

62, 73, 76, 78, 84, 86, 89, 90, 92,95.

Therefore, $\frac{84+86}{2}$ = 85

Mean

The mean is sometimes referred to as the average. The mean is simply the sum of measurements divided by the total number of measurements.

When you talk of population the mean is denoted by μ (greek symbol). While the sample mean is denoted by $\bar{x}$.

The formula for the mean is, $\bar{x} = \frac{\sum_{i-1}^{n} xi}{n}$

Where, $\sum$ = summation of variables / measurements of stating with the lowest value divide by total number of students. For example, 7+10+8+.....= $\frac{221}{25}$ = 8.84 = 9

Weighted Mean

This is used to find the degree of use attached to each of the observations. Also used in comparisons, for instance, the performance of students. See the example below.

SS 241, distribution of marks (wi).	John scored (xi)	Jane scored (xi)
Test 20	70	90

Research	30	90	85
Exam	50	85	70

The weighted mean is given by; $\bar{x} = \frac{\sum wixi}{\sum wi}$

John,

xi	$wixi$
70	1400
90	2700
85	4250

$\sum wixi$ = 8350

Jane

xi	$wixi$
90	1800
85	2550
70	3500

$\sum wixi$ = 7850

Substituting in the formula we get

For John, $\bar{x} = \frac{8350}{100}$

=83.5%

For Jane, $\bar{x} = \frac{7850}{100}$

= 78.5 %

As shown above, 100 is the summation of $\sum wi$ which comprises of the test, research and Exam. It is therefore, worth to conclude that John did better than Jane.

Measure Of Central Tendency

Age Group	fi	xi	$fixi$
15-19	1	17	17
20-24	1	22	22
25-29	3	27	81
30-34	3	32	96
35-39	8	37	296
40-44	4	42	168
45-49	3	47	141
50-54	2	52	164

$\sum fixi$ = 925

The formula of the mean for grouped data is; $\bar{x} = \frac{\sum fixi}{n}$

Substituting in the above formula we get;

$\bar{x} = \frac{925}{25}$

= 37 years.

Median

When dealing with the grouped data the median is given by; $Md = L + \frac{(\frac{n}{2} - F)i}{f}$

Where, L is the true lower limit of the class interval in which median value is located. In this case the true lower limit is 34.5, n = 25, F is the cumulative frequency corresponding to the class interval preceding the one that contains the median item, f is the frequency of distribution class interval.

Now substituting in the our formula, we get;

Md = 34.5+ $\frac{(12.5-8)5}{8}$

= 37.5

Mode

This measure has two different approaches.

- The crude mode; simply involves picking out the mid points of the highest interval.
- Using the median of interpolation. In this respect the mode is given by;

Mode = $L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) i$

Where,

$\Delta_1 = fm_0 - f_1$,

$\Delta_{2=} fm_0 - f_2$,

fm_o = frequency of the model class,

f_1= frequency in a class interval preceding the modal class.

f_2= frequency in a class interval after the modal class. Therefore, Mode = 34.5+ $\left(\frac{5}{5+4}\right)5$

=34.5 +$\left(\frac{5}{9}\right)5$

=37.2years

THE EMPERICAL MODE

Require that you have the mean and median.

Mean-3 (mean-median)

=37-3 (37-37)

=37years.

This relies on the already computed mean and median. Interpretation is the same; the majority of these people are below 37 years.

Choice of an appropriate measure of central tendency

This will be dealt with by looking at the strengths and weakness attitudes of these measures. These measures include; mode, mean and median.

Mode; has the following weaknesses;

- Does not use all the values in the distribution. It is difficult to use in further computation.
- Some people find it difficult to interpret the mode.
- The mode by its nature makes it possible that you can have more than one mode. It is problematic when it comes to chose.

Strengths: it can be very useful in circumstances like planning and decision making. Eg (manufacturing) production of shoes, one can know which shoe size is mostly worn by consumers.

Mean

The mean is more commonly used measure of central tendency. Its advantages are as follows;

- it is easier to understand and used in everyday life.
- It takes into account all values in distribution.

Disadvantages of a mean

This has a disadvantage in that it can be affected by the presence of extreme values in a distribution. Hence, the mean may not be very reliable measure. For example,

Per capita income $=\frac{national\ income}{population}$, then it was $350, for Zambia. It is not reliable in the sense that it's not every Zambian who earns this amount in a year. Some have more while others have less.

Median

The median has advantages over the mean. Because;

- it is more stable than the mean.
- It is not affected by extreme values and because of this; the median would be a better tool to use in determining the above.
- It is good at showing the relative position of the people (measure).

Disadvantages

- It is mostly confused with the mean.
- It does not take into consideration other values into consideration in comparison to the mean.eg patterns of consumption of alcohol

xi	xi	xi
89	89	144
83	83	83
77	77	77
20	75	75

Md = 81 Md =81 Md=81

$\bar{x}$ =70 $\bar{x}$ =81 $\bar{x}$ =92

MEASURES OF DISPERSION

Measures variability / spread of measurement to see how they differ from each other or from the central value. Refer to the extent to which values in a distribution vary from the centre.

THE RANGE most base measure of dispersion is one of the set of the simplest measure of dispersion-difference between the largest and smallest value in the observation given when

dealing with grouped data, you can have the range which is crude does not/ gives little information about variability or dispersion of the measurement(about variability about the mean).

DEVIATION FROM THE MEAN

A measure in the form $xi - \bar{x}$ for each observation you get the measure of deviation from the mean.

The mean deviation is simply by the formula; $MD = \frac{\Sigma(xi-\bar{x}}{(n-1)n}$

ABSOLUTE MEAN DEVIATION

Is an attempt to improve upon the mean deviation it therefore deals with absolute figures. The absolute mean deviation is given by;

$$MD = 1/2\left[\frac{\sum|xi - \bar{x}|}{n}\right]$$

It's also difficult to use it because it not easy to interpret because in some cases it gives large values so it's hardly used in statistics.

VARIANCE

Is a measure better for variability or dispersion the variance of any observation/ measurement from to the mean is the sum of square deviation s from the mean divided by n-1.

Formula for the variance is;

$$S^2 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{n-1}$$

For a population variance is; $\sigma^2 = \frac{\sum_{i=1}^{n}(xi-\mu)^2}{n}$

If you use n and n-1 you may be leaving room for of the sample n may be useful if the sample size is large and close enough to the population. The variance is not an end but used to get to the standard deviation.

STANDARD DEVIATION FOR UNGROUPED DATA

Is the square root of the variance is much better measure of dispersion E.g. data on cigarette smoking.

xi	$xi - \bar{x}$	$(xi - \bar{x})^2$

5	1.8	3.24	
4	0.8	0.64	
3	-0.2	0.04	
1	-2.2	4.84	
3	-0.2	0.04	$\sum xi$ =8.80, n =5

Substituting the above in the formula for the variance,

We have; $\frac{8.8}{4}$ =2.2

Therefore, $S = \sqrt{S^2}$

$S = \sqrt{2.2}$

= 1.48

This means that, 50% of these people take 1.48 and the other 50% are below the mean.

COMPUTATION

$$S^2 = \frac{n\sum_{i=1}^{n} xi^2 - (\sum xi)^2}{n(n-1)}$$

xi	xi^2
5	25
4	16
1	1
3	9

$\sum xi^2 = 60$

$$S^2 = \frac{5(60) - (16)^2}{5(5-1)}$$

$$= \frac{300 - 256}{20}$$

= 2.2

Standard deviation (S)= $\sqrt{S^2}$, $S = 1.48$

MEASURES OF DISPERSION FOR GROUPED DATA

Age Group	fi	xi	$xi - \bar{x}$	$(xi - \bar{x})^2$	$fi(xi - \bar{x})^2$
15-19	1	17	-20	400	400
20-24	1	22	-15	225	225
25-29	3	27	-10	100	300
30-34	3	32	-5	25	75
35-39	8	37	0	0	0
40-44	4	42	5	25	100
45-49	3	47	10	100	300
50-54	2	52	15	225	450

$\sum fi(xi - \bar{x})^2$= 1850

Formula for grouped data;

$$S^2 = \frac{\sum fi\,(xi-\bar{x})^2}{n-1}$$

$$= \frac{1850}{25-1}$$

=77.08

$=\sqrt{77.08}$

=8.8

COMPUTATIONAL FORMULA

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

QUARTILE DEVIATION

This is a measure of the dispersion about the median and is based on the inter – quartile range. The inter – quartile range that is, the distance between the first quartile and the third quartile. The first quartile or Q_1 or 25% is that point in the distance which has frequencies about it. The third quartile Q_3 for has ¾ or 75% of observation frequencies below it or ¾ larger above it in the distribution.

Age Group	f	F
20-24	3	3
25-29	4	7
30-34	5	12
35-39	6	16

40-44	5	23
45-49	4	27
50-54	3	30

$$Q_1 = L + \frac{\left(\frac{1}{4}n - F\right)i}{f}$$

$Q_1 = 29.5 + \frac{(7.5-7)5}{5}$ =30

$$Q_3 = L + \frac{\left(\frac{3}{4}n - F\right)i}{f}$$

$Q_3 = 39.5 + \frac{(225-18)5}{5} = 44$, The Quartile deviation (QD)

$$QD = \frac{Q_3 - Q_1}{2}$$

$QD = \frac{44-30}{2}$

= 7

The smaller the number the greater the tendency in terms of concentration towards the median.

One weakness is that, the quartile deviation does not take into account of the values between the first and third quartile.

COEFFICIENT OF SKEWNESS

Other measures of dispersion do not talk about the direction of the dispersion. This is what coefficient of skewness does better. It allows the computation of a measure of skewness which shows the direction of dispersion around the centre. It is a much more supperio measure of dispersion than other measure of dispersion.

It also shows whether or not that there symmetry or lack of symmetry in a dispersion. This is important in knowing whether the dispersion is normal or not.

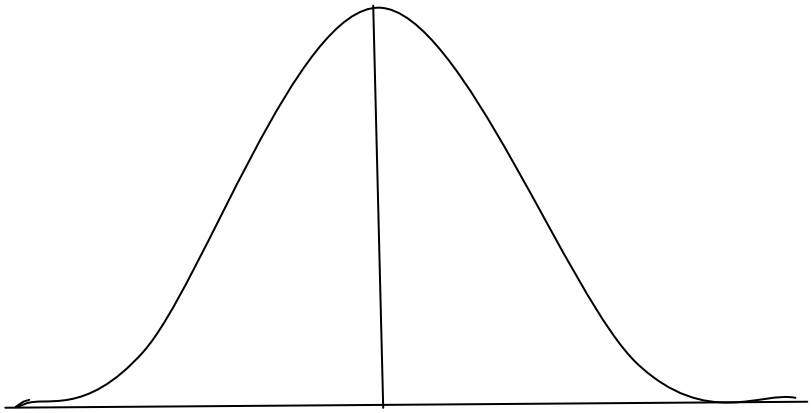
The coefficient of skewnnes can be computed if there is a mean, mode, median and standard deviation. The mode is normally left out because of some of its weakness that is, there can more than one mode.

Skewness = $3\left(\frac{\bar{x}-median}{S}\right)$

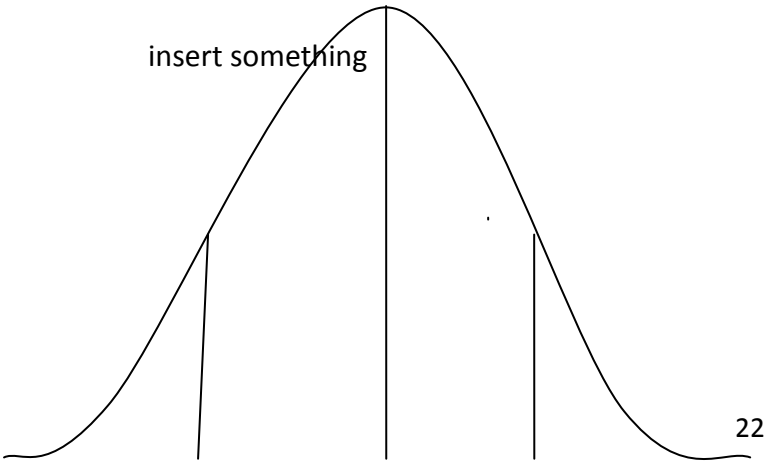
THE RELATIONSHIP BETWEEN SKEWNESS AND RELATIVE POSITION OF THE MEAN,MODE AND MEDIAN.

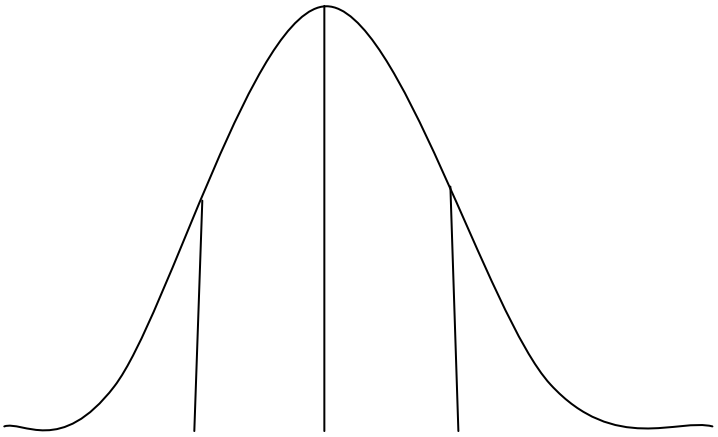
Depending on the given values, the skewness can either be positive or negative or asymmetrical. Suppose there is $S = 5,\ Md = 25,\ \bar{x} = 25, Mode = 25$

Insert



insert something





$\bar{x}$ Md $Mode$

If there is $S = 5$

$Md = 25$

$\bar{x} = 20$

$Mode = 30$

Skewness= $3\left(\frac{20-25}{5}\right) = -3$

$Mode$ Md $\bar{x}$

if there is $S = 5$

$Md = 25$

$\bar{x} = 30$

$Mode = 20$

Skewness= $3\left(\frac{30-25}{5}\right) = 3$

THE COEFFICIENT OF VARIATION

This is measure that is use to compare variables which are measured in variables that cannot be compared directly because they are expressed in different unit. It is normally used in relation to groups e.g. the can be two of students who are assessed on the basis of numeracy and literacy. Numeracy will be measure in different units and so will be literacy. There will be a mean or different means for each different SD.

THE CONCEPT AND DEFINATION OF PROBABILITY

Probability refers to likelihood of occurrence of an event. In probability theory normally numerically evaluate this likelihood of occurrence of an event. What is a probability that a randomly selected student can be male or female? That is the probability will be one or zero.

VOCABURALY ASSOCIATED WITH PROBABILITY EXPERIMENT

An experiment is a situation with defined set of outcomes. If you flip a coin there will be two outcomes that is head or tail. Selection of a student randomly will either be male or female.

EVENTS AND SAMPLE POINTS

Possible outcomes of an experiment are referred to as an event or sample points or possible outcomes. If a coin is flipped and the head comes up, then the head is the sample point.

SAMPLE SPACE

This refers to a list of all possible output of an experiment, so that if you want to predict the birth of a baby, the list of all possible outcomes will be male or female.

COMPOUND EVENTS

Are those events that can be decomposed into two or more events e.g. car accidents. Possible events;

Car accident; die or injured. No car accident; no death or no injury.

MUTUALLY EXCLUSIVE EVENTS

If you flip a coin head and tail are mutually exclusive events. This means that the occurrence of the events affects the other e.g. you cannot be dead and alive at the same time.

INDEPENDENT EVENT

These are events where the occurrence of one event does not affect the occurrence of another event. You can have two events occurring at the same time e.g. people being very slim, being very fat, being dull or intelligent. If being very fat influences being very intelligent, it is referred to as dependent events.

Dependant Events

The occurrence or non occurrence of an event has a bearing on the likelihood of occurrence of another event. For example, studying, success or failure in exams. That is, reading hard leads to success in exams.

Types of probability

- A prior
- Experimental
- Subjective

A Prior probability refers to 'before the event'. All possible events are known before and have equal likelihood of occurrence. For example, flipping a coin, the likelihood of getting a head or a tail is a prior. The probability of this outcome is 50 to 50, that is, its ether a head or a tail.

$$P(head) = \frac{heads}{heads + tails}$$

$$= \frac{H}{H+T}$$

$$= \frac{1}{2} \text{ or } 0.5 \text{ or } 5\%$$

On the basis of this it follows that an event which is represented by (*E*) can occur in A-different ways and can happen A'- different. Then, it follow of that, an event E is going to happen is simply;

$P(E')=\frac{A}{A+A'}$ Same as, $P(H)=\frac{H}{H+T}$

If these events are the probability that an event *E* will occur, then the probability that it an event *E* will not occur is;

$$P(E') = \frac{A'}{A + A'}$$

The total number of event is equal to the sum of the sample space, which is equal to one;

Total number of events= *A+A'=n*

If this is the case then it follows that the probability of *E* will occur $P(E) = \frac{A}{n}$ and the probability E will not occur, $P(E') = \frac{A'}{n}$ therefore, $P(E) + P(E') = 1$

AXIOMS OF PROBABILITY

One of the axioms is that no probability can be greater than one and no probability can be less than zero. Any probability lies between 1 and 0. All the probability will fall between zero and one. That is, $0 \leq P(E) \leq 1$

The probability that an event will not occur is simply; 1-$P(E), which\ is\ the\ same\ as\ (one - probability\ of\ not\ occuring).$

The sum of mutually exclusive event is equal to one. $P(E_1) + P(E_2) + \cdots \ldots . P(E_n) = 1$

Probability that an event will not occur is; $P(E') = 1 - P(E)$, For example, the probability of dying; $P(D) = 0.75$, then the probability of surviving; $P(S) = 1 - 0.75$, The axioms are derived from a prior probability.

EXPERIMENTAL PROBABILITY

Is that probability based on actual observation and empirical evidence? In most cases, it is based on limited number of observation derived from random sampling experiment. In most cases it is expressed in form of relative frequencies$\left(\frac{x}{n}\right)$, e.g. $P(E) = \frac{x}{n} = \frac{A}{n}$

SUBJECTIVE PROBABILITY

This is not based on actual observations. It relies on intuition and personal conviction that an event will take place.

THE MATHEMATICS OF PROBABILITY

This is based on four rules were three are apply to the addition of probability and one to multiplication.

Addition of Probability

The general rule of addition of probability is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. It applies to any two events that are mutually exclusive. If the events are mutually exclusive then the formula is changed to the following;

$P(A \cup B)$=$P(A).P(B)$

Multiplication of Probability

The probability that two independent events will occur as the product of the probability of the separate event (independent event); $P(A \cap B) = P(A).P(B)$, but if events are not independent; $P(A \cap B) = P(A).P(B/A)$

A related probability to these probabilities is referred to as conditional probability $P(A/B) = \frac{P(A).P(B)}{P(B)}$ or $\frac{P(A).P(B/A)}{P(B)}$

Special Case of Independence

When *A* and *B* are independent events, it follows that $P(A) = P(A/B)$ or $P(B) = P(B/A)$ In such situation you have the following kind of scenario;

$P(A \cap B) = P(A).P(B),$ $P(A \cap B) = P(A),$ $P(B) + P(A).P(B/A)$

ADDITIONAL PROBABILITY

Radio Listen ship

Radio Station	Female E_4	Male E_5	Total
Christian Voice E_1	450	500	950
Phoenix E_2	300	800	1100
ZNBC E_3	100	350	450
Total	850	1650	2500

Addition rule for any two events which are not necessarily mutually exclusive.

What is the probability of listening to radio Christian Voice and being female? Label the events as;

Probability of listening to Christian Voice = P(E_1)

Probability of being female = P(E_4)

Therefore the probability of listening to Christian voice and being female is given by;

$$P(E_1 \text{ or } E_4) = P(E_1) + P(E_4) - P(E_1 \text{ and } E_4)$$

Computation; Probability of listening to Christian voice $=\frac{950}{2500} + \frac{850}{2500} - \frac{450}{2500}$

=0.38+0.34-0.18

= 0.54

Note: subtract the joint occurrence to avoid double counting.

ADDITION OF MUTUAL EXCLUSIVE EVENT

The probability of being female or male;

$P(E_5 \text{ or } E_4) = P(E_5) + P(E_4)$

$= \frac{1650}{2500} + \frac{850}{2500} = 1$

$P(E_4) = 1 - 0.34$

$P(E_5) = 1 - 0.66$

MULTIPLICATOIN RULE FOR ANY TWO EVENTS

Probability of any two terms is the product of the probability of the individual events. For example, find the joint probability of being male as well as listener to ZNBC;

$P(E_5) \text{ and } E_3) = P(E_5).P(E_3/E_5)$

Where;

Probability of being Male $= \frac{1650}{2500}$

$Probability\ of\ listening\ to\ ZNBC\ given\ that\ one\ is\ male = P(E_3/E_5) * P(E_5)$

$$P(E_5 \text{ and } E_3) = \frac{350}{1650} * \frac{1650}{2500}$$

$$= 0.14$$

CONDITIONAL PROBABILITY

What is the probability of listening to Christian voice given that one is female.

$$P(E_1/E_4) = \frac{P(E_1 n E_4)}{P(E_4)} \; OR \; P(E_1) * \frac{P(E_4/E_1)}{P(E_4)}$$

Where;

$$P(E_1) = \frac{950}{2500}$$

$$P(E_1/E_4) = \frac{450}{950}$$

$$P(E_4) = \frac{850}{2500}$$

Now substituting in the above expression, we get;

$$P(E_5 \; and \; E_3) = \frac{950}{2500} * \frac{\frac{450}{950}}{\frac{850}{2500}} = 0.53$$

MULTIPLICATION RULE FOR INDEPENDENT EVENTS

	$FEMALE^{E4}$	$MALE^{E5}$	**TOTAL**
$christian\ voive^{E1}$	470	380	950
$phoenix^{E2}$	650	440	1100
$ZNBC^{E3}$	270	180	450
TOTAL	1500	1000	2500

Example

Find joint probability of listening to ZNBC and being male.

P (E_3 andE_5)

=**P** (E_3), **P** (E_3/E_5) **P** (E_3) = $\frac{450}{2500}$, **P** (E_3/E_5) = $\frac{180}{450}$

P (E_3 andE_5) = $\frac{450}{2500}$, $\frac{180}{450}$

= **P** (E_3), **P** (E_5) = 0.07 **OR** **P** (E_3), **P** (E_5) = $\frac{1000}{2500}$, $\frac{450}{2500}$

Meaning; the probability of being male does not influence listening to ZNBC and vice-versa. It proves the independence of listenership from sex.

TYPE OF DISTRIBUTION

Observed or Empirical distribution is the kind which relies entirely on observed values (actual observation). An example of these is the frequency distribution.

Probability Distribution

This is simply a theoretical distribution of all possible events and probabilities of occurrence of each event.

Expected Distribution is simply the product of the probabilities or the occurrence of each event and the total number of events.

Expected Distribution

This is simply the product of the probabilities or the occurrence of each event and the total number of events.

Event	**Probability distribution**	**Expected distribution**	**Observed / empirical distribution**
Heads	0.5	10	12
Tails	o.5	10	8
Total	1.0	20	20

INTRODUCTION TO STATISTICAL INFERENCE

When you talk about statistical inferences, we talk of a situation where you draw conclusions about a whole population on the basis of a sample.

Population - μ parameter.

Sample -$\bar{x}$ statistics.

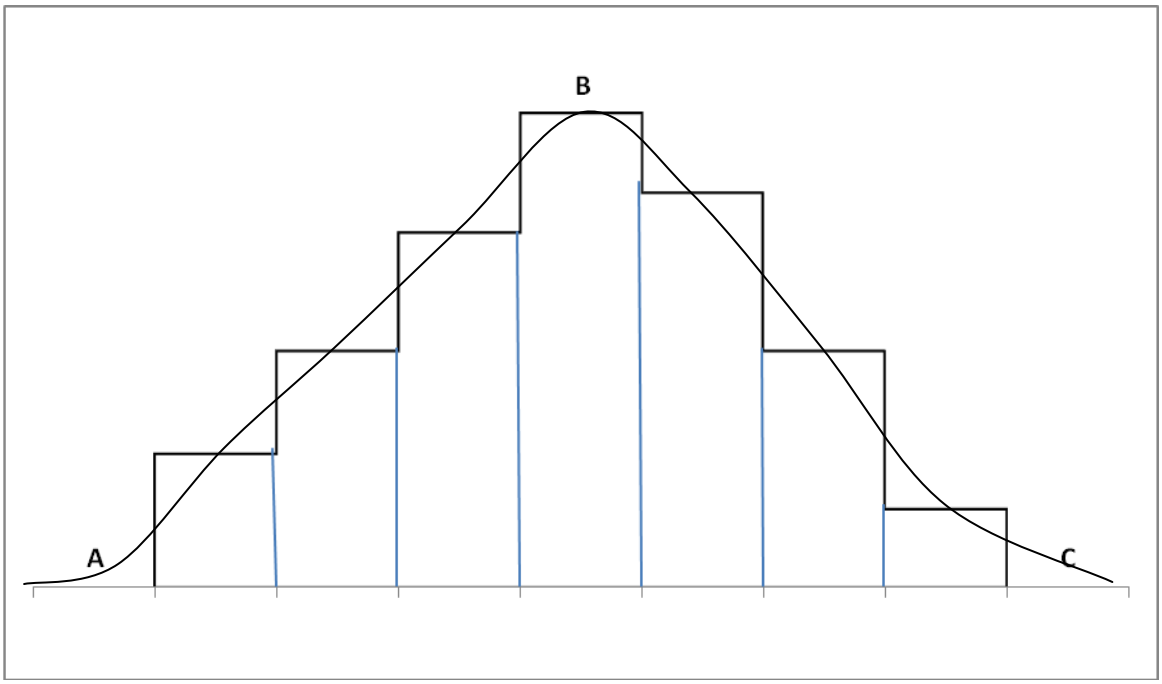
The sample has to be representative of the population and has to be randomly selected giving each element of the population an equal chance of being selected. One must be mindful of the fact that the inferences we are making may be an error or wrong (there is a chance that the inference may not actually reflect the parameter). To estimate the error we need to have an understanding of a theoretical probability distribution which reveals the normal curve or normal distribution.

Normal Curve

This is a very important tool in probability distribution because a large number of distributions conform to be approximately normal. It is also important because of its significance in statistical inferences. Empirical distribution tends to be unique from each other and can vary depending on the data. Its possible to conceive of a theoretical

satisfaction in which a smooth curve can be drawn were you can actually have a smooth especially where there is an infinite sample. If the sample size has been increased, the curve becomes smoother and smoother.

The normal curve for normal distribution is a theoretical presentation of a manner where most variables tend to distribute themselves randomly. A normal curve is theoretical representation of the manner in which most attributes or traits of variables which occure at random tend to distribute themselves naturally. Eg if we are talking about height, there is a tendency for variables like height to cluster up around the centre and those values that are extreme will tend to be far from the centre , these will therefore be the extremes of the curve. Other examples are weight, age, performance in

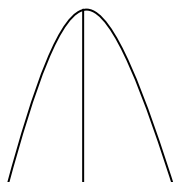


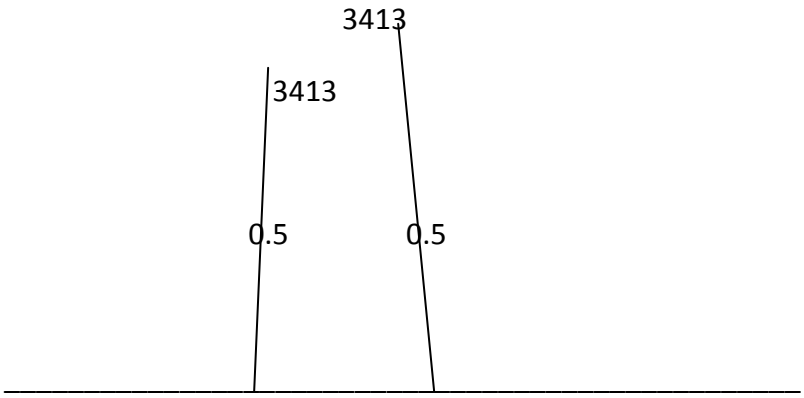
A *is the extremities,* **B** *is where most random variables are connected(at the Centre),****C*** *is the extremities*

This is an example of a normal curve. Knowledge of the normal curve is very important in statistics. If you draw a sample from a population and you have variables of the sample, the distribution in both the sample the population will tend to be the same. This assumption made in a sample tends to be the same as those in a population. This is therefore the importance of a normal curve which is the basis of statistical inference.

PROPERTIES OF THE NORMAL CURVE

1. It is unimodal
- It has an identical mean, median and mode.
- The mean, median and mode coincide at one point.





2. It is symmetrical
- Both sides of the baseline, you have 0.5 or 50 of the observations.
- The area under the curve adds up to unit(one) ie 0.5+0.5=1.

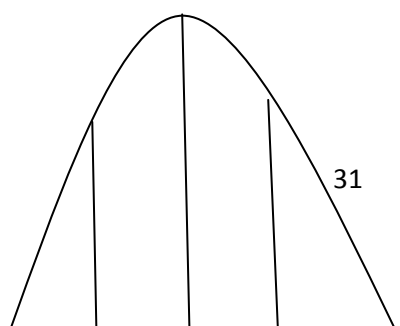
3. It is asymptotic
- It shows that the area between the curve and the baseline extends to infinite on both sides of the curve ie from positive infinite to positive infinite.
- The proportion not under the curve is very small.
- In theory it covers almost all the instances.

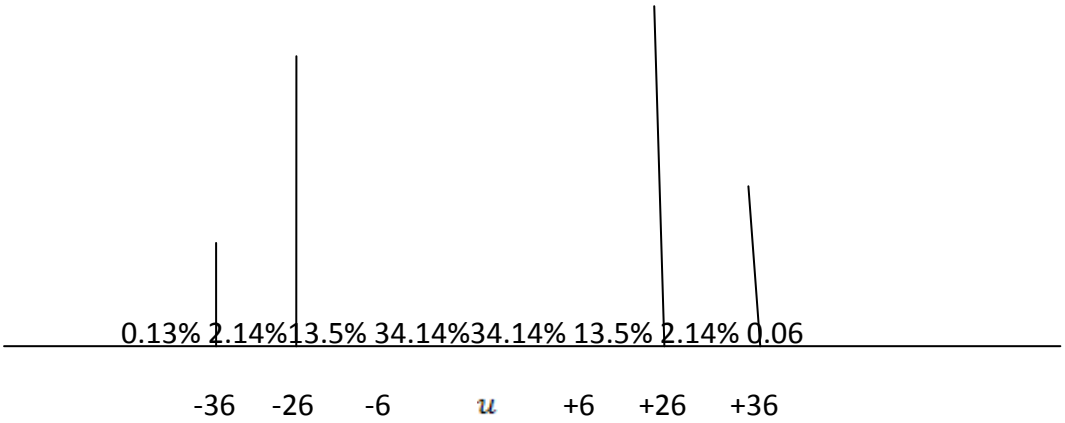
4. It is continous
- The NC deals with variables which are continuous, ie that take numerical variables eg speed, height, age.
- It rarely deals with the measures on a norminal scale.

5. It has a standard deviation.

Standard deviation marks (establishes) the distance on the baseline from the centre where the mean is located. These are established in such a way that the area between the curve and the baseline can be expressed in proportion, percentages, or even probabilities. These partitioning is done in such a way as to show that about 68% of the area around the curve lies within one standard deviation of the mean. About 95% of the area under the curve lies within 2 standard deviation of either side of the mean. About 99.7% of the curve lies within 3 standard deviation of the curve or ***

In most instances, most variables conform to this pattern. In most cases, the coefficient of **** may not necessarily be zero because of the issue of errors.





THE STANDARD NORMAL (DISTRIBUTION) CURVE AND ORDINARY NORMAL CURVE AND STARDARD SCORED

On the basis of the mean and the standard deviation of the randomly distributed data it is possible to construct a standard normal distribution. The standard normal distribution resembles an ordinary normal curve.

There is only one standard normal curve distribution unlike ordinary normal curve which can be several.

In order to go round these problems it comes necessary to standardise the ordinary normal curve. This standard normal curve which has the same mean and standard deviation has the mean of 0 and a standard deviation of 1, at all times.

Standardisation uses a very simple formula. The formula is referred to as standard score or standard normal deviation or z-score.

Formula; $Z = \frac{X-\bar{X}}{S}$ $or = \frac{X-\mu}{\sigma}$ *for the population.*

By using this formula you can standardise any ordinary normal curve.

Let's suppose you have a random sample of 10 students.

$n = 10$, *students had scored the following.*

Xi	X^2	$Xi - \bar{X}$	$Xi - \bar{X})^2$
20	400	-30	900
40	1600	-10	100
50	2500	0	0

50	2500	0	0
90	8100	-40	1600
70	4900	-20	400
30	900	-20	400
60	3600	10	100
10	100	-40	1600
80	6400	30	900

$\bar{X} = 50, \quad S = 1, \quad$ *using the above figure to substitute in the formula above,*

starting with the first student who got 10,*we get* ; -1.16, -0.39, 0.00, 0.00, 1.55, 0.77, -0.77, 0.39, -1.16 respectively as our *Z-scores.*

What the responses mean is that once you have n+z-score it means that particular score lies above the move. But when you have a+z-score it means that particular score lies below the mean.

While negatives are located to the left of the mean or the left of zero, the positives are located to the right of the mean. Each z-score is associated with proportion of area under the standard normal curve.

COMPUTATION

If you have two student in SS242 group

John Banda (JB) scores 40%

Jane Phiri (JP) scores 80%

Can you establish the percentage of students who scored between the 2.What is the probability that a student scored between John and Jane? Given that the mean is 50 and SD is 25.82

	$Xi - \bar{X}$	Z	$Proportion$
JB =40	-10	-0.39	0.1517
JP =80	30	1.16	0 .3770
			0.5287

When you are dealing with opposite sides of the curve add the corresponding properties.

Then we also have Peter Mwanza (PM) 60% and David Chanda (DC) 90%.

	$X - \bar{X}$	Z	$Proportion$
PM=60	10	0.39	0.1517
DC=90	40	1.55	0.4314
			0.2877

When you have 2 scores on the same side of the curve, you subtract. (Subtract small variables from the larger ones).

IMPORTANT USES OF THE NORMAL CURVE

There are important in the estimation of parameters on the basis of sample statistics. Using the standard score (z-score), you can determine the distance between the parameter and the statistics. You can determine the distance of how far away statistics is from the parameter.

To estimate this distance requires an understanding of 3 theoretical principles namely:

- Sampling distribution
- Law of large numbers.
- Central Limit theorem.

LAW OF LARGE NUMBERS

One basic principle about parameters is that if one draws a sample randomly, the distribution of the statistics from the large sample will be similar to distribution of the parameter from the population.

When sample size increases, sample statistics becomes a more accurate estimator of parameter (population).

CENTRAL LIMIT THEOREM

It states that if you draw a sufficiently large sample size or number of sufficiently large sample size, you compute for each one of the samples.

A sample statistics mean you end up with a sample distribution the means.

N=5000

n=100

40 samples

For each sample compute a mean and eventually end up with sampling distribution of means. The mean of the sampling distribution of means will be equal to the population mean. The importance of the sampling distribution is that

- Estimates parameters and
- Significance tests (or hypothesis testing).

The sampling distribution of means is very important because it is often used in statistical inference. For any given score you can know the probability of locating in a normal curve. Any sample you pick will fall within two deviations.

The sampling distribution of means constitutes the normal distribution of a curve. This follows that all characteristics of a normal curve apply in the sampling distribution of means i.e. unimodal, symmetrical, asymptotic and continuous.

	Mean	Standard Deviation
Population	μ	σ
Sample	$\bar{X}$	S
Sampling Distribution	μ	$\frac{\sigma}{\sqrt{n}}$

THE STANDARD ERROR OF THE MEAN

The standard deviation of the sampling distribution is called the standard error of the mean. Knowledge of the standard error helps in estimating how accurately a sample mean estimates a population mean. In other words it is the measure of precision of sample estimates.

A sample mean has 68% chance of being a standard mean of the population mean.

$\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}}$ $S\bar{x} = \frac{S\bar{x}}{\sqrt{n}}$

Note: the smaller the standard error the more the precise the estimate is, similarly the larger the SE, the less precise the estimate will be.

ESTIMATION OF PARAMETERS

There are two types of parameter estimates that are possible.

- Point estimate

- Interval estimate.

POINT ESTIMATE

What is done is simply use a sample statistic to estimate the population mean. You can only do so if the basis of computing sample statistics is done randomly and it becomes a representation of the population. If this is not the case, you are likely to make an error in estimation.

$$\mu = \bar{X}$$

$$\bar{X} = \mu$$

If this is the case, there is a probability of error and the error of estimation is given by this;

$$|\bar{X} - \mu|$$

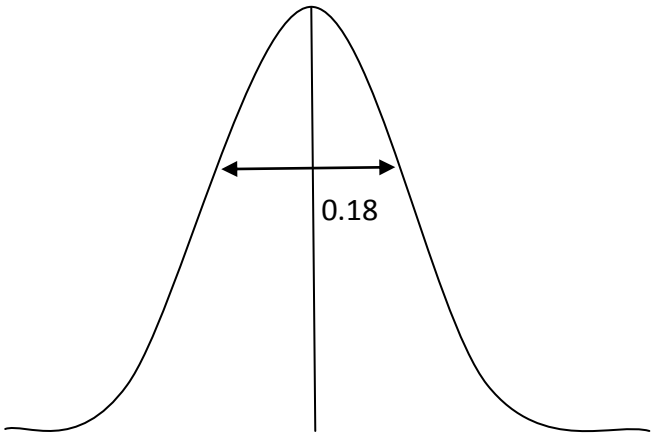
The *E* of *E* is simply the appearance between what is you think the parameter value is and what is actually is. You can on the basis of this go a step further. You can compute the bound on the error of estimation for the point estimate.

This bound on the error of estimation is the measure of how good our inferences of the estimate are. The smaller the bound, on the *E* of *E,* the better the inference. The bound on the E of E is given by the formula;

$$ZS\bar{X} = \frac{ZS}{\sqrt{n}}$$

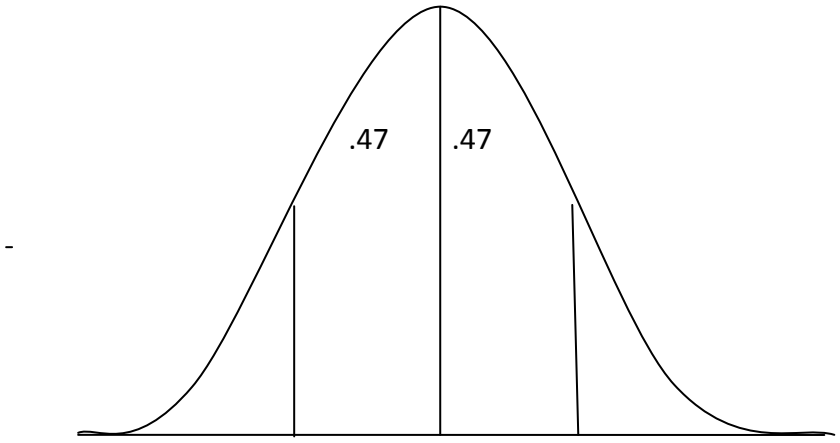
The basis for using this is that we confine ourselves to two standard deviations. Suppose you have a random sample of 150 UNZA students and interested in estimating the average number of years students spend on campus. Take the random sample and ask students what years they are in or how long they have been on campus. You will find that the mean of years students are suppose to spend on campus$\bar{X} = 3.2\ years$. Therefore, $\mu = \bar{X} = 3.2\ years.$ when you use standard deviation of; SD =1.1years. Pick up the sample mean to represent the population mean.

$\frac{2*(1.1)}{\sqrt{150}} = 0.18$



3.2

There is a 90% chance that any sample mean lies within two standard deviations. It gives confidence that the average years spent on campus is just a small distance of 0.18 from the actual mean. This is a purely a good estimation.



WHAT DETERMINES THE GOODNESS OF AN ESTIMATE?

There are a number of factors;

1. Unbiasiness: an estimate is unbiased if the mean of the sampling distribution equals the population mean or population parameter. For a situation where you have sufficiently large sample size the mean the sampling distribution will be equal to the population mean.
2. Consistence: an estimate is consistent when it gets closer and to the parameter as the sample size becomes larger and larger.
3. Efficiency: an estimate is efficient when the standard error for the sampling distribution is small. The smaller the standard error, the more efficient the estimate.

INTERVAL ESTIMATE

It's very possible when using a point estimate it may not accurately estimate the population mean. As such it's better an interval estimate; it is simply a range of continuous values of the statistic within which a true parameter is located with a known degree.

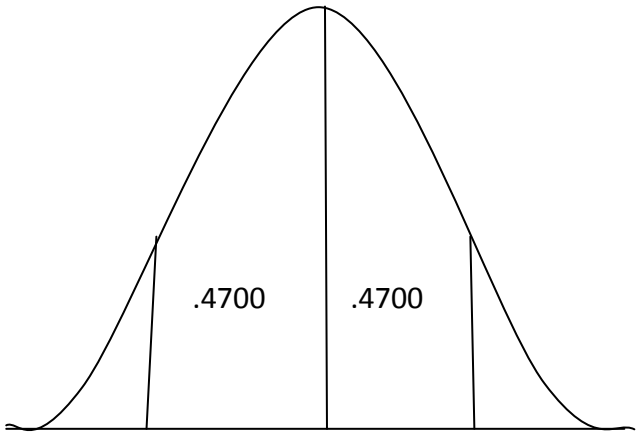
95% Confidence Coefficient

$\bar{x} \pm z.s\bar{x}$

$u \pm 6 = 68\%$

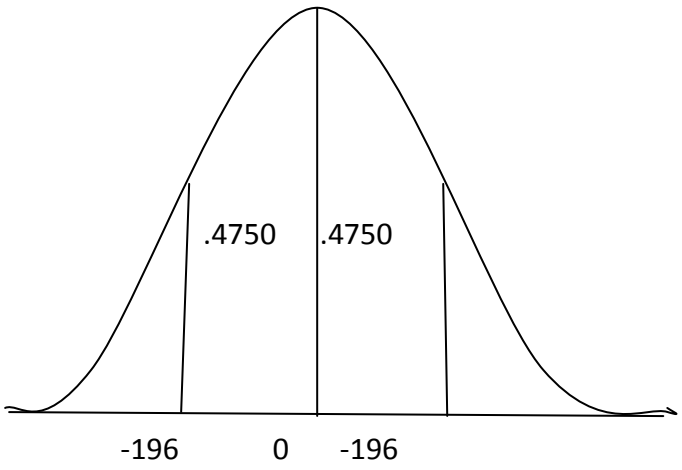
$u \pm 26 = 95\%$

$u \pm 30 = 99\%$



When computing sample mean, you have to compute a confidence interval. The confidence intervals vary dependence on certain factors. When you have that, you can evaluate the goodness of an interval estimate. Evaluate the probability that an interval will encompass the parameter. This probability is a confidence coefficient. This is a probability that a given interval will encompass the parameter to be estimated.

You should be 99% percent sure that the interval you are constructing will contain the mean (this is the confidence coefficient and it varies). 95% =0.95 (z-score). If you divide 95% or 0.95 by 2 you get 0.4750.



Take a random sample of n=36. Get questionnaires.

When you combine these variables you find that the mean amount of money is x=k3792 and the standard deviation is S= k124.

HYPOTHESIS TESTING

Procedures in Hypothesis Testing.

Hypothesis testing actually involves confirming whether comparison has to be made between the mean based on a sample and the mean of a sampling distribution (population mean).

When the mean from the sample divides substantially the mean from the sampling distribution, then you can reject the hypothesis that state the sample mean and the population mean are equal. Then you can accept that the, two are different as stated by hypotheses.

PROCEDURE IN HYPOTHESIS TESTING

1. Formulation of the Theoretical or Theoretical or Research Hypothesis (H).

This is that hypothesis which can be deduced from existing theory or based on experience on an observation or other sources that are available. E.g. using income as an example, the research hypothesis could be H_1 average amount got by UNZA employees is above K570,000.

$H_1: \mu > K570{,}000$

$H_1: \mu_1 > \mu_2$

2. Formation of the null or statistical hypothesis (H_0).

The null hypothesis that is directly testable. All the time contradicts the research hypothesis.

$H_0: \mu = K570{,}000$

$H_0: \mu_1 = \mu_2$

$\mu_1 = \mu_2 = 0$

E.g. the null hypothesis could state that the average amount got by UNZA employees is above K570, 000.

Rejection of null hypothesis increases the problem that the research or theoretical hypothesis could very well be a correct hypothesis.

Acceptation of the null hypothesis implies that research hypothesis may be in error. After formulating the hypothesis you worry about making assumptions.

3. Making assumptions.

Assumptions are mainly concerned with the distribution of the parameter. If the sample is sufficiently large, then you can use an assumption of normally of the distribution of the parameter. It's normally assumed that the parameter is normally distributed.

An assumption of a random sampling is also made (always assume that random sampling has been used). An assumption concerning the level or scale of measurement is made. It's important because it helps in the choice of statistical test to use. An assumption of non parametric such as ordinal and nominal is also made.

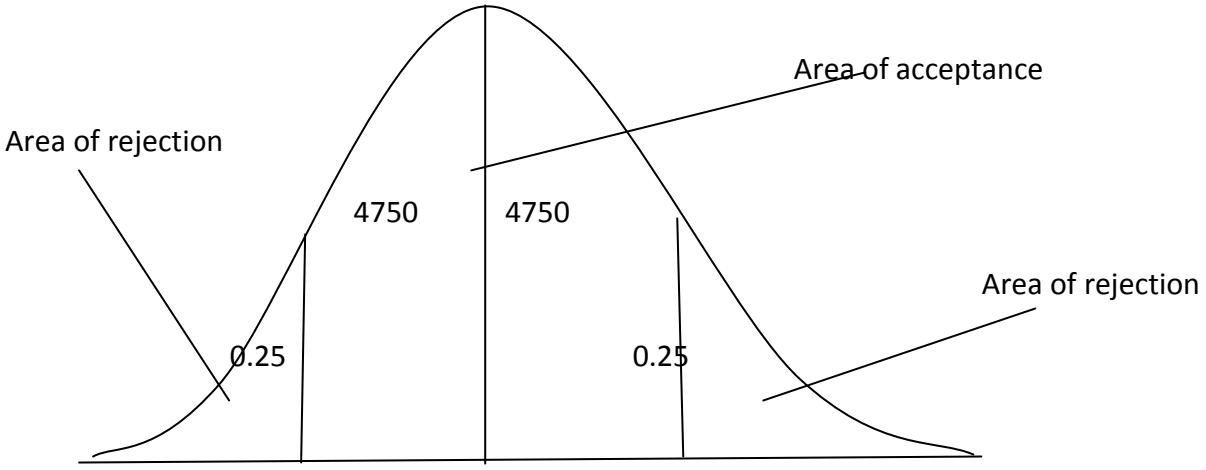
4. Obtaining appropriate sampling distribution.

Fairly straight forward when using random distribution. If you assume normally then you will use normal or Z distribution. The choice of critical values which are like cut off points the acceptance or rejection of the statistical analysis or hypothesis.

5. Choosing the significance level and determining the critical region.

In choosing and determining the above you can make two types of errors.

- Type one error; refers to an error of rejecting the null hypothesis when it is in fact true. The problem by committing type one error is known as the significance level of signifance level denoted α(alpha). The most frequently used levels of significance are .05 and .01. The levels of significance correspond to some Z score of 1.96. Once the level of significance has been shown, you can determine the critical value (they determine where the critical regions are located). That is cut off values or points for the rejection or acceptance of the null hypothesis.



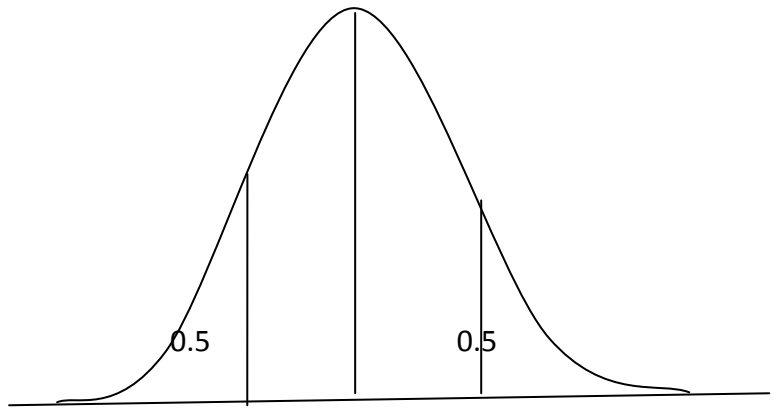
-1.96 +1.97

They represent the cut off points, therefore the area between – and + are area of acceptance. That is between -1.96 and 1.96. Area of acceptance is the area used in hypothesis testing. If a computed value falls in that area, then the null hypothesis is accepted.

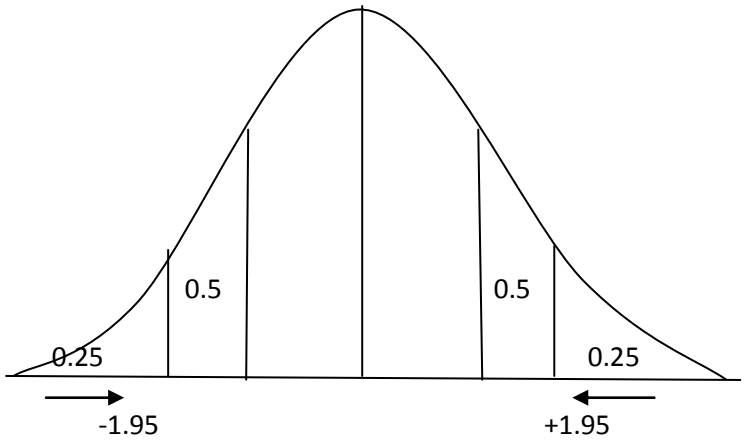
The area of rejection fall outside the area mentioned above, that is -1.96 and +1.96 (critical values). If the observed value falls in this area (outside the critical region), then reject the null hypothesis.

- Type two Error

Arises from the mistake of accepting the null hypothesis when it is actually false and when research hypothesis true. The research hypothesis is denoted by β (beta).



Type one and two errors are inversely related. For example if you change the rejection area to increase type one error then type two error decreases and vice-versa. Type one error can be increased by shrinking the area of acceptance. E.g. moving -1.96 and 1.96 to some new points, in order to reduce the area.



Expanding the area of acceptance, results into an increase in the probability of accepting the null hypothesis.

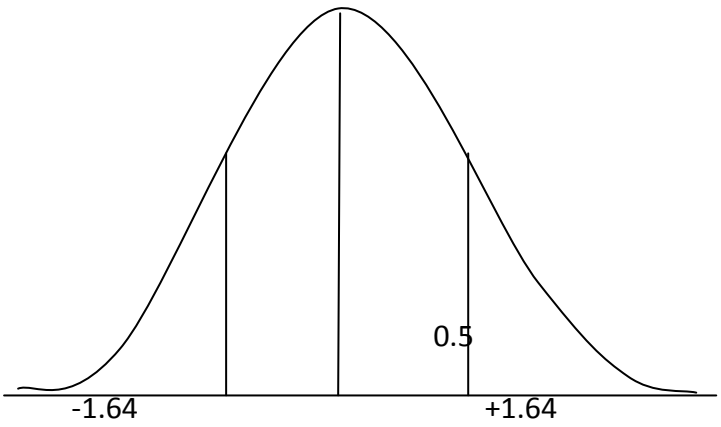
CHOICE OF THE LEVEL OF SIGNIFICANCE

A prevailing situation dictates the choice of the level of significance. In certain situations where it might be preferable to commit type one error, it might be better to impose strigent tests to make sure that the product does not harm human beings. In manufacturing for instance, the ideal standard can be 75. i.e. μ = 75 then it is perfect to be put on the market. If $\mu < 75$ retest

$H_o : \mu = 75$ reject

$H_1 : \mu < 75$ retest

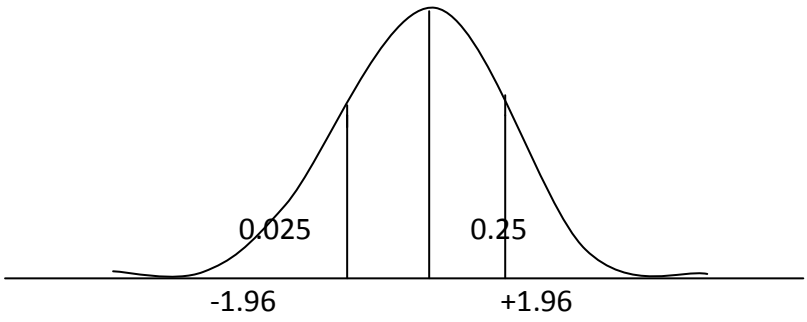
In such a situation, it might be better to increase the probability of type one error so as to increase the probability of rejecting the null hypothesis. So it means increasing the percentage to about 10% that is making sure the area of acceptance contracts and increasing the area of rejecting.



There are cases where it is possible to commit type two error. Assume there a statistically minded judge who says if someone has committed five crimes. Then it is a death sentence. There is also the possibility of someone committing more than five crimes. $H_0: \mu = 5$ Retrial or acquit. In this case it is preferable to commit type two error. That is, accept the null hypothesis. In this case, you broaden the area of acceptance and increase the probability of accepting the null hypothesis.

TYPES OF TESTS

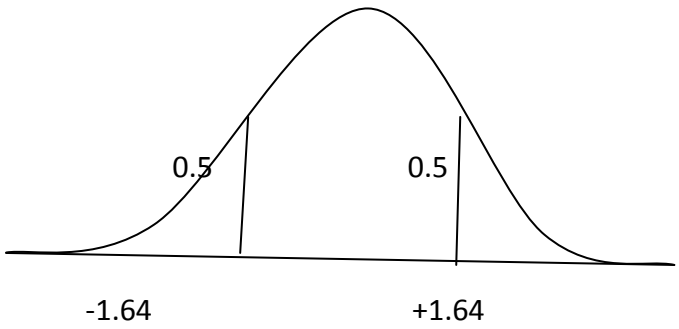
One Tailed and two Tailed tests/ Directional and Non-directional tests.



$u = 75$

$u \neq 75$

The above is an example of a two tailed test because both sides of the tail are used. This is also known as non directional test. For a two tail place the critical values on both sides of the tail.



Greater than in the H_1 if it is less than locates the critical value on the left.

CRITICAL VALUES OF ONE TAILED TESTS

	SIGNIFICANCE	LEVEL
	0.05	0.01
One tailed test	1.645	2.33
Two tailed test	1.96	2.58

6. COMPUTING THE TEST STATISTIC

At this stage, look at the sample data and calculate the test statistic known as the observed value. This will normally come in the form of Z-score or a t-score depending on the sample size. Once it has been computed, look at the data and make a decision based on which side the Z-score is placed on the curve.

7. DECISION; this depends on the area of rejection or acceptance.
8. CONCLUSION; on the basis of the decision get to the conclusion.

TESTING A HYPOTHESIS ABOUT A PROPORTION

This is a scenario where you may want to make the inferences about a proportion. The formula is;

$$z = \frac{p_s - p_u}{\sqrt{p_u \, {}^{q_u}/_n}}$$

For example, you are a consultant of an organisation and asked to evaluate a programme or rehabilitate some criminals. Then you are given a thousand of files. Take a random sample out of the thousand of files. Say $n = 125$ cases. When you take two sample out of this sample you find the percentage of successful cases is 55% or$p_s = 55\%$, then manager tell you that their standard is 60% or $p_u = 60\%$.

$$z = \frac{p_s - p_u}{\sqrt{p_u * \frac{q\mu}{n}}}$$

Then you get to be asked if the job they are doing is below to standard that is 60%.

PROCEDURE

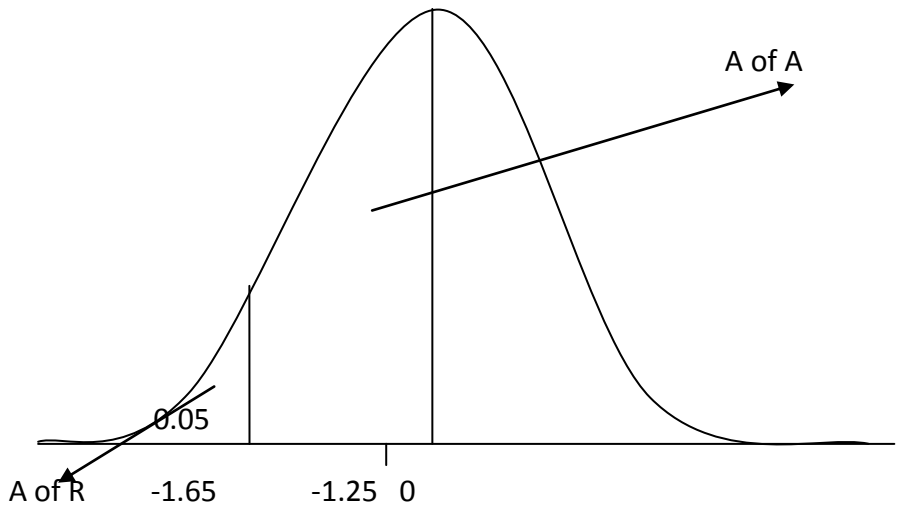
$H_0: P\mu = 0.60$

$H_1: P\mu < 0.60$

Assumption: Random sampling, normality of distribution of the parameter and level of measurement.

Decision Rules: Given $\propto = 0.05$ or 5% level of significance, one tailed test determines critical values based on the information given. See where to place it on the table whether left or right.

$$Z = \frac{Ps - P\mu}{\sqrt{\frac{P\mu q\mu}{n}}}$$



If

$If\ Zobs > -1.65\ fall\ in\ the\ area\ of\ acceptance(A\ of\ A),$
$If\ Zobs \leq -1.65\ falls\ in\ the\ area\ the\ rejection\ (A\ of R)$

COMPUTATION

$$Z = \frac{Ps - P\mu}{\sqrt{\frac{P\mu q\mu}{n}}} \qquad P\mu = 0.60,\ q\mu = 1 - 0.60 = 0.40, n = Sample \text{ size}$$

$$Z = \frac{0.55 - 0.60}{\sqrt{\frac{0.6*0.4}{125}}}$$

$$Zobs = 1.14$$

Decision: Accept H_0

Conclusion: Its highly probable that the agency has maintain the same standard.

INFERENCES ABOUT TWO MEANS-(Difference between the means)

You may want to compare two means reflecting two samples or you may compare two means of two populations and then establish of there is a significant difference between in the mean of female student and male student.

For example, you are a researcher interested to have an understanding of social perception of kids who are at nursery school going and those who are not.

N_1=77 $N_2 = 64$

$\overline{x_1}$=15.3 $\overline{x_2}$ = 15.3

$s_1 = 3.4$ $s_2 = 4.6$

INFERENCE CONCERNING DIFFERENCES BETWEEN TWO MEANS LARGE SAMPLE

This happens in this situation we are interested in comparing two parameters from two different populations. E.g. the mean value of male population and the female population. This comparison requires the tests of the differences between two mean.

STATEMENT OF HYPOTHESIS

Suppose a test is current out on male and female students' performance in an aptitude. The population is stratification into male and female.

Group	n	$\overline{x}$	s
Male	133	25.34	5.05
Female	162	24.94	5.44

Can you establish whether there is a statistical difference in the performance of male and females in relation to the aptitude test?

$H_0: \mu_1 = \mu_2$ (two tailed non directional test)

$H_1: \mu_1 = \mu_2$

Assumption

The assumption here differ from those where there is only one mean.

- The subjects that are included in this situation are independently and randomly selected.
- The groups are independent. The population variances are homogenous or equal. This assumption is susceptible to variation. To avoid this, compute a group variance. The population on the meant is normal, that is ($\overline{x_1} - \overline{x_2}$). The level or scale measurement is interval.

Decision Rule

If -1.96<$Z\ obs$<+1.96, accept H_0

$Z\ obs \leq -1.96\ or\ Z\ obs \geq +1.96$ Reject H_0

<u>Computation</u>

Formula $Z = \frac{u_1 - u_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1 - n_2}}}$ population information

$= \frac{\overline{x_1 - x_2}}{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}}$ sample information

$$Z\ obs = \frac{25.34 - 24.44}{\frac{5.05}{133} + \frac{5.44^2}{162}}$$

= 0.654

Decision; Accept the null hypothesis.

Conclusion; there is not any difference statistically in the performance of male and females.

DIFFERENCE BETWEEN PROPORTIONS LARGE SAMPLE RESULTS

In such a situation, hypothesis using P instead of H. E.g. conduct two survey in Lusaka, Kabwe to ascertain viewer habits towards ZNBC. During a survey the following Data is obtained;

Kitwe $\frac{680}{1000}$ $p_1 = 0.68$

$q_1 = 0.32$

Lusaka $\frac{444}{600}$ $p_2 = 0.74$

$q_2 = 0.26$

Question: is there any statistical difference between the residents of the two towns. That is, viewership of Kitwe greater than that of Lusaka using 5% level of significant?

$Ho: p_{u_1} = p_{u_2}$

$Hi: p_{u_1} > p_{u_2}$

ASSUMPTIONS: same as with difference between means.

DECISON RULES

If $Z\ obs \leq +1.641\ accept\ HO$

If $Z\ obs \leq +1.641\ reject\ Ho$

COMPUTATION

$$Z\ obs = \frac{ps_1 - ps_2}{\sqrt{\frac{ps_1\%}{n_1} + \frac{ps_2\%}{n_2}}}$$

$= \frac{0.68-074}{\sqrt{o}}$ **insert something**

DECISION: Accept the null hypothesis

CONCLUSION: it is not true that there is greater viewership in Kitwe than in Lusaka.

INFERENCE ABOUT A MEAN:

Small Sample Result $n \leq 30$

T-Distribution

That distribution was a distribution that a statistician W.S Consset came up with. When he was constantly taking rejecting hypothesis for small samples using Z-distribution formula for t-distribution is *tobs* $= \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

It's referred to as the student distribution Gauze came up with the normal distribution Z, also called the Gauze distribution.

PROPERTIES

- The t-distribution like the z- distribution is also symmetrical about the mean.
- Different from the standard normal distribution because the t- distribution is more variable. Because of this there are many different t-distributions. For each distribution there is an associated degree of freedom which is normally $(n-1)$
- Each distribution is specified by a parameter called degree of freedom.

Sample size is an equivalent to the distribution of t- approximates, the standard normal distribution, as the sample size becomes larger. (Degree of freedom also increases as sample size becomes larger). If you focus your attention on properly on the graph you notice that. EG, take small sample size of n=2, the degree of freedom is n-1 which is 2-1.

You are carrying out a research on a small population of one class and carries about how much money students on newspapers. You suspect the expenditure may differ from what has been established in the past. What do you do? Take a sample of 10 i.e. $n = 10$. You compute the amount of money on newspapers k4100, that is,

$n = 10$

$\bar{x} = K4{,}100$

$s = K359$

This moves you to hypothesis (a test for a single mean).

DECISION RULES

Given $\alpha = 0.5, non\ directional\ (2\ failed) test\ with\ df2 = n - 1 = 10 - 1 = 9$

If -2.262 $< tobs < 2.262, accept\ H_0$. If tobs $\leq -2.262\ or\ tobs \geq +2.262, reject\ H_0$

Computation

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

$\frac{4100-4200}{\frac{359}{\sqrt{10}}}$

INFERENCE CONCERNING DIFFERENCE BETWEEN MEANS OF SMALL SAMPLE

For instance, suppose someone says they observed male students to be heavier than female because they eat better (male).

Males	Females
$n_1 = 10$	$n_2 = 10$
$\overline{x_1} = 47\ kg$	$\overline{x_2} = 45\ kg$
$S_1 = 1.21$	$S_2 = 0.92$

Assumptions

The male and female students are independent of each other.

HYPOTHESIS

$H_0 = \mu_1 = \mu_2$

$H_1 = \mu_1 > \mu_2$ (One tailed directional)

DECISION RULES

Given$\propto = 0.05, directional, (t - tailed)\ and\ n_1 + n_2 - 2 = 18$

If tobs$< 1.734, accepted\ H_0$. If tobs $\geq 1.734, rejected\ H_0$

COMPUTATION

tobs = $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)S^2+(n_2-1)S_2^2}{n_1+n_2-2}}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$

$$tobs=\frac{47-45}{\sqrt{\frac{9(1.21)+9(0.90)}{18}\left[\frac{1}{10}+\frac{1}{10}\right]}}$$

=4.01

Decision: Reject H_0

CONCLUSION: it's highly that male students are much heavier than female students.

SIMPLE ANALYSIS OF VARIANCE

(ANOVA) dealing with more than 2 means, you want to compare means for different crime rates, EG how, high, median. Instead of working directly with means, ANOVA involves working directly variances hence 2 independent estimates of a common variance are required. One of these estimates of common variance is based upon the variability between groups and is often referred to as the between group variance

E.g. you may have $\{ |low\ |medium|high| \}$

The other independent estimate independence variance is based upon the variability within groups hence is called within group variance.

THE UNDERLYING LOGIC OF ANOVA

This determines the differences among the group means are significant by comparing them to the variations within groups. This takes the form of the comparison is a static called the F-ratio.

$$F-ratio=\frac{VBG}{VWG}$$

Once you have this F-ration, the F-ration is, in relation to the F critical the more significant are the differences among the means or among the category of sample means. When you compute you have to compare with the critical value. If it is larger or greater then there must be some significant level. ANOVA unlike other distribution uses the F- distribution (which is closer to the normal distribution) so that when testing the hypothesis, you compare the F-ration with the F-critical, in any case or instance EG given crime rate; low , medium, high.

To get F- critical:

1. Must specify or have degrees of freedom.
2. Level of significance
3. Must have two estimates of degrees of freedom.

For variance BG you must have estimates of degrees of freedom. Take the number of observations and subtract by one. For instance, BG= J-1=3-1=$2dfb$

WG=$n - J = 12 - 3 = 9dfw$ here as shown the number of observations should subtract the number of groups.$\propto = 0.05$ Degrees of freedom $\frac{DFB}{DFW}$

For variance BG you must have estimates of degrees of freedom. Take the number of observations and subtract by one.

TUTOR(1)	$(xij - \bar{\bar{x}})^2$	**TUTOR (2)**	$(xij - \bar{\bar{x}})^2$	**TUTOR(3)**	$(xij - \bar{\bar{x}})^2$
80	16	70	36	63	169
92	256	81	25	76	0
87	121	78	4	70	36
83	49	74	4	58	324
$\bar{x}$=85.5		$\bar{x}$=75.75		$\bar{x}$=66.75	

HYPOTHESIS

$H_0: \mu_1 = \mu_2 = \mu_3$

$H_1: \mu_1 \neq \mu_2 \neq \mu_3$

After this move on to assumptions

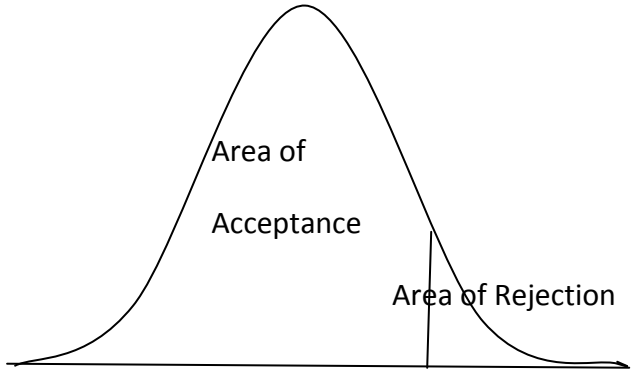
1. Subjects are independently and randomly selected
2. Groups are independent from each other
3. Population variances are homogenous
4. The population distribution from which the samples have been drawn is normal
5. The scale of measurements is interval

Decision Rules:

Given 0.05 level of significant

$dfb = J - 1$ =3-1= 2

$dfw = n - 1 = 12 - 3 = 9$



If F ratio $<$ 4.26 accept H_1 and if F ratio $\geq$4.26, reject H_0

Computations of the grand mean involve summing total observations and divide by the total number of observations. Given by,

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

Summing the totals observations across rows (i) and columns (j)

Next: computation of the sum or rather the total sum of squares involves finding the sum of square deviations, involves finding of the square of deviations of the observations of the grand mean which is 76.

Eventually you sum up that is, add up all square deviations.

TUTOR(1)	$(xij - \bar{\bar{x}})^2$	**TUTOR (2)**	$(xij - \bar{\bar{x}})^2$	**TUTOR(3)**	$(xij - \bar{\bar{x}})^2$
80	16	70	36	63	169
92	256	81	25	76	0
87	121	78	4	70	36
83	49	74	4	58	324
$\bar{x}$=85.5	442	$\bar{x}$=75.75	69	$\bar{x}$=66.75	529

$\bar{\bar{x}} = \frac{\sum\sum xij}{n}$ =76

COMPUTATION OF THE TOTAL SUM OF SQUARES

$\sum(xij - \bar{\bar{x}})^2$ Refer to the above table, TSS=422+69+529=1040

COMPUTATION OF THE SUM SQUARES BETWEEN GROUPS

$(xij - \bar{\bar{x}})^2$, multiply each square deviation by the size of sample as follows, $nj(xij - \bar{\bar{x}})^2$ *and size of the category is used as a weighing factor*.

COMPUTATION OF THE VARIANCE BETWEEN GROUPS

You simply devide the total sum of squares between G by the degrees of freedom

$$variance\ between = \frac{total\ sum\ of\ squares\ between}{DFB}$$

$$VB = \frac{SS}{DFB} = \frac{703.2}{2} = 351.75$$

COMPUTATION OF THE SUM OF THE SQUARE WITHIN GRUOPS

Involves the subtraction of the total sum of square between G from total sum of the squares.

SSW = TSS-SSB

= 1040-703.5

= 336.5

COMPUTATION OF THE VARIANCE WITHIN GRUOPS

Involves the dividing of the total sum of squares within groups by the degrees within groups.

VW=$\frac{SSW}{DFW}$ =$\frac{336.5}{9}$ =37.39

COMPUTATION OF THE F-RATIO

Simply divide the variance between by variance within.

F-RATIO=$\frac{VARIANCE\ BETWEEN}{VARIANCE\ WITHIN}$

=$\frac{35175}{37.39} = 9.41$

Decision Rule: Reject the H_0

CONCLUSION: it is highly likely that there is a significant difference among the tutors and performance of students.

REGRESSION ANALYSIS

It is a descriptive tool by which is a researcher you try and determine the linear dependence of one variable by another.EG you might want to determine the linear dependence between income and education. That is, Linear dependence=Y

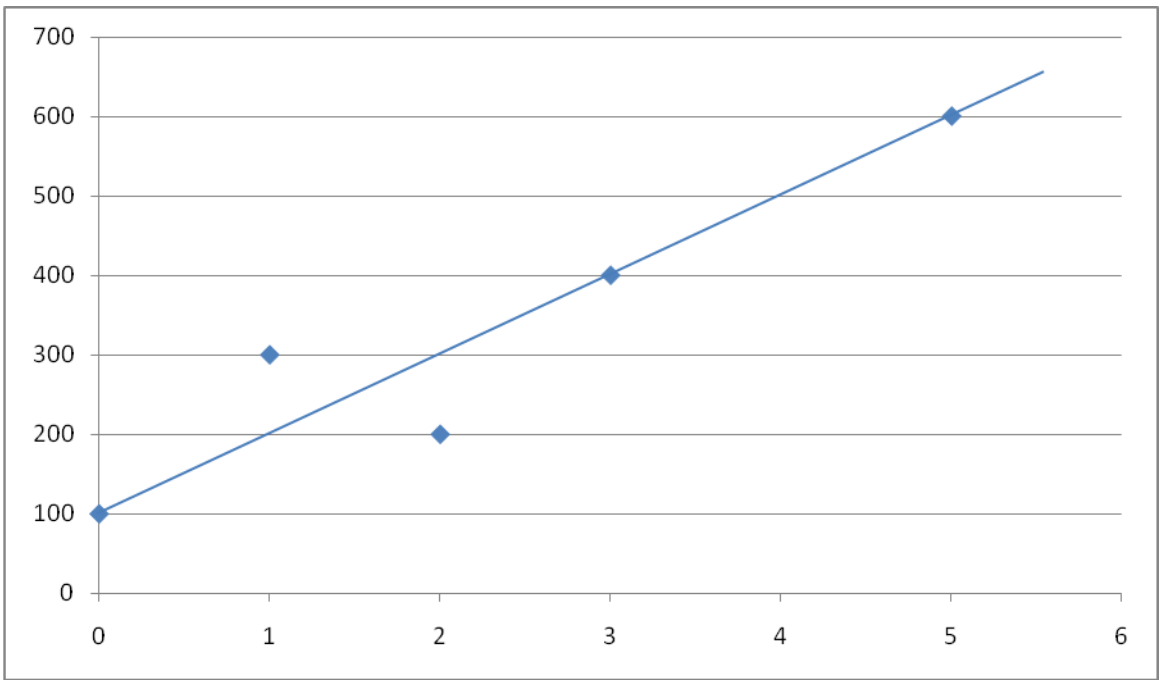
X= linear independent.

You can determine the linear dependence of one variable to several variable. In that case you use only the best linear prediction equation in evaluating the prediction accuracy of the regression equation. Sample Rejection equation is, Y=$A + Bx \; or \; \alpha + \beta x$

Multiple Regressions equation is, Y=$A + Bix_1 + B_2$

THE EYEBALL FITTING TECHNIQUE

You rely on the scatta graph,



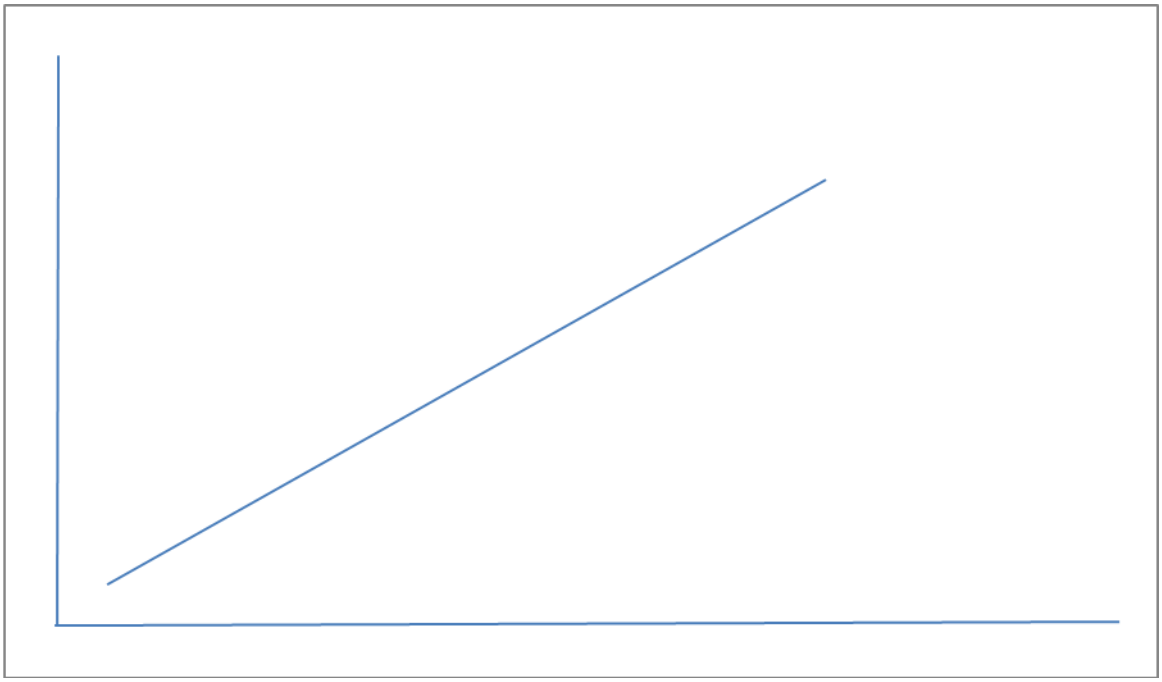
You put observations on this graph dependent variations are the vertical axis while independent on the horizontal.

THE METHOD OF LEAST SQUARES

This is the better method of predicting values of Y on the basis of X. It involves minimisation of the sum of the squared residuals along the regression line.

$Y' = A + Bx$

Y' is the predictable value of X then the error of prediction is represented by $Residual = Y - Y'$. This means the difference between the actual value of Y and the predicted value of X is the error of prediction. The method of least squares attempts to minimise the sum of squared residuals along the regression line for all sample points. All the predicted values for Y are supposed to lie along the regression line or at least the square line. The vertical distances from the regression line represent the residuals or regression.



$Y = A + Bx \; where \; A \; is \; Y - intercept$

On the basis of this predicted equation we can derive some very important constants.

Constant A is also referred to as the Y intercept. At represents the point at which the regression line crosses the Y axis or value of Y when X=0.

For example, Y-axis= Earnings, X=Education

Constant B (regression Coefficient) represents the slope of the regression line. Coefficient also indicates the expected change in Y for a unit change in X.

Y-axis (earnings) X-axis (education) represents the expected change in income resulting from an extra year in education.

S	$Post\ Sec\ Education(yrs)$	XY	$Income(Y)$	X^2	Y^2
1	5	300	600	25	3600
2	7	522	76	49	5776
3	15	1440	96	225	9216
4	12	1200	100	144	1000
5	8	648	81	64	6561
6	7	525	75	49	225

7 10 800 80 100 we can predict one's earning on basis of their education.

$$B = \frac{N(\sum xy)-(\sum x)(\sum y)}{N(\sum x^2)-(\sum x)^2}$$

$$= \frac{7(5445)-(64)(568)}{7(656)-(64)^2}$$

=3.55

Interpretation; for any increased number of years in education one is expected to get K3.55 more.

$$A = \bar{Y} - B\bar{x}$$

$$= 81.14 - 3.55(9.14)$$

=48.69 therefore, $Y = 48.69 + 3.55x$

E.g. predict the earnings of an employee in an institution who has 6 years' post secondary school education (PSSE).

$$Y = 48.69 + 3.55(6)$$

=69.99

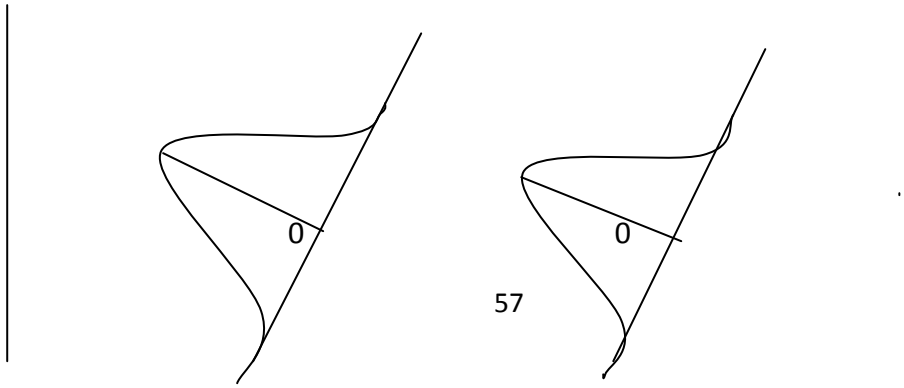
Some earnings 107.75 estimate the number of years he /she has in PSSE.

107.75= 48.69+3.55x

$$x = 16.64yrs.$$

ASSUMPTIONS OF REGRESSIONAL ANALYSIS

1. Assumption of linearity, they $\bar{x}$ and $\bar{y}$ are assumed to be on the regression line. Relation between X and Y is such that as X changes results in a systemic change in Y and these changes move in a straight line. In case of curvilinear you have to carry out some transformation using logarithms.
2. Normality- for ant fixed value the independent variable (X) the distance of the dependent variable (Y) is normal.



3. Equality of Variance; (Homoscedatiaty) means you make the assumption that the average size of residual along the regression line is constant.

4. Independence; suggests that the observation of Y must be statistically independent of each other e.g. you cannot have members of the same family.
5. Random Sampling; randomly and independently selected. Scale of measurement is usually interval.

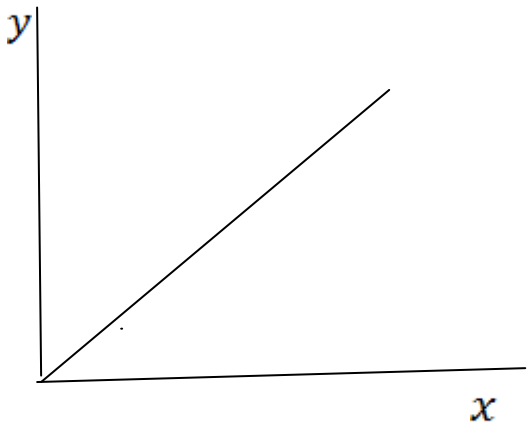
CORRELATION ANALYSIS

Correlation refers to the existence of a relation between two variables by changng the variables moving in some direction you have positive coleration and vice versa.

The persion product moment coleration coefficient is used to measure the strength of the relationship between the two is (X and Y) can also be used to measure how well the data fits in a straight line.

PROPERTIES OF COLERRELATION COEFFICIENT

The coefficient has between -1 and 1values. The values of +1 indicate the positive correlation.



-1 represents a weak negative correlation, value of zero indicates non existence of relationship. A strong positive correlation does not automatically indicate causality. Assumptions

1. Linearity
2. Normality
3. Round sampling
4. Interval scale of measurement
5. Independence
6. Quality of variance =∑ of squared computation

Corr- coefficient $(\gamma) = \frac{N\sum xY - (\sum X)(\sum Y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum Y - (\sum Y)^2]}}$

Introduction to Non parameter Test

Are those tests don't make any assumptions about parameter (e.g. about the mean). They do not make any assumptions about the normality of the distribution of the population from where a sample was drawn. NPT are also designed mostly for ordinal or normal scales. They deal with interval in terms of measurements.

ADVANTAGES

1. NPT are much easier to understand than PT.
2. NPT can be used with very small sample.
3. Relaxation of the assumption concerning normality.

TYPES

The rank order correlations sometimes referred to as **Spearman's** rank order correlation coefficient.

$$\gamma_{s=1-\left[\frac{6\sum D_i^2}{N^3-N}\right]}$$

Can be used if you want to find out if there is a correlation in the performance of the structure in M162 and SS242

D_i^2	D_i	M162	SS242	Rank M162	Rank SS242
1	1	80	75	10	9
4	2	70	60	7	5
36	-6	60	77	4	10
0	0	50	45	1	1
25	-5	57	72	3	8

25	-5	55	70	2	7
36	6	72	50	8	3
1	1	62	55	5	4
36	6	75	47	9	2
0	0	65	65	6	6

Firstly you can rank the values of the two variables i.e. performance in M162 and SS242 starting from the lowest value to the highest.

Then you compute the difference between the ranks. (Refer back to the table above)

Following the formula you have:

$$1-\frac{6(166)}{10(10^2-1)}=\frac{996}{990}$$

=

CONTIGENCY TABLES

CHI-SQUARE TEST (TEST OF INDEPENCE)

If we have more than 1 more than variable and measurements for two or more variable i.e. BIVARIABLE

MULTIVARIABLE

You can arrange them in a two way table as follows

	Female	Male
MMD		

UPND

UNIP

ZRP

An arrangement like this is used to determine two variables are related or independent of each other or whether it is possible to predict one variable on the basis of the other variable.

These two variables are sometimes referred as contingency tables because the test for hypothesis always that there's a contingent relationship between them.

A test for independence is used

$$\sum_i \sum_j \frac{(Oij - Eij)^2}{Eij}$$

E.g. where you try to establish the extent to which religious affiliation affects attitude towards abortion.

n=456

Attitude torwards.....	**Protestant**	**Catholics**	**Total**
For	126	99	225
Against	71	162	233
Total	179	261	458

Hypothesis

Ho: There's no relation between R A & attitude towards abortion.

Ho: There's no relationship

Hi: There's a relationship

Assumptions

1. Random samples
2. Groups are independent
3. Each observation must fall in one and only one category or group
4. The sample must be large enough or fairly large that no expected frequency less than 5 when the number of rows and columns is greater than 2 i.e. r and c >2. If the number of rows and columns is equal to 2 i.e. r and c = 2

5. Scale measurement should be.... normal or.....

Next: Then move on to the decision Rules: you have to compute the degrees of freedom given $\alpha = 0.5$

Formula: $df = (r-1)(c-1)$

$= (2-1)(2-1) = 1df$

How to use the table for non-directional test if $x^2 < 3.84$ accept H_o

If $x^2 \geq 3.84$ Reject H_0

Formula for expected frequency is;

$$Eij = \frac{row\ total\ X\ column\ total}{N}$$

$$= \frac{225(197)}{458}$$

$$= 96.8$$

O ij	***E ij***	***(O ij- E ij)***	$(O\ ij - E\ ij)^2$	$\frac{(O\ ij - E\ ij)^2}{E\ ij}$
126	96.78	29.22	853.84	8.82
99	128.22	-29.22	853.84	6.66
71	150.22	-29.22	853.84	8.52
162	132.78	29.22	853.84	6.43

Conclusion

It's highly probable that there's a relationship between the attitude and religious affiliation.

PRINCIPLE OF TABLE READING

Decide the independent and dependant variables. Compute the percentages in terms of categories/ independent variables. RA can be taken to be indepent variable. Do the analysis by comparing the ppercntages of.....

CHI-SQUARE AND THE GOODNESS OF FIT TEST

The goodness of fit test is that system used when you are trying to determine to what extent your observed data matches the expected data e.g.....

One test and 2 failed tests Directional and non directional tests .....