Statistics

Statistics is the science of collecting, analyzing and making inference from data.
It is the selection, collection and organization of basic facts into meaningful data.
It is the totality of scientific methods used in collecting, presenting, summarizing and analysis of data as well as drawing valid inferences/conclusions.

Importance of Statistics

-It helps in presenting laborious or big information in summarized and precise form.
-It is used in projecting trends of events which are highly required for planning.
-Enables us to design, collect relevant and accurate data for solving problems objectively.
-Allows us to analyze and interpret numerical information in a scientific manner so that we derive logical and reliable conclusions
-Aids in estimation, forecasting and quality control
-It is used by researchers in many fields to organize, analyze and summarize data.
-It is also often used to communicate research findings and to support hypotheses, give credibility to research methodology and conclusions.
-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.

The information in the steps listed below can be used when planning and conducting your research. It can also be used to evaluate the research of others when completing your literature review.

Steps in the Data Analysis process

Step 1	Pose a question
Step 2	What to measure and how
Step 3	Collecting data
Step 4	Summarizing and displaying data (measures of central tendency indicate what is typical of the average subject)
Step 5	Analyzing data and interpreting results (Hypothesis testing; we look to reject or fail to reject the null hypothesis)

Descriptive &inferential statistics

- Descriptive statistics is used to summarize a set of data into just a few numbers that represent the entire data set.
- Summarizes may be tabular, graphical and numerical.
- Most frequently used are measures of central tendency (mean, median, mode) and variability (variance and standard deviation).
- Research involves sampling in which conclusions about the population are drawn from observations of samples
- Statistical methods which are used for making inferences/conclusions about the population called **inferential statistics.**

Terminologies

Data

Data are defined as a series of observations, measurements or facts. Data are divided into two types;

1) Qualitative (categorical) data

-Values that can be placed in non-numerical categories

-Examples gender, socio economic status, religious preferences

Qualitative data can be classified further as either nominal or ordinal.

2) Quantitative data

-Consist of values representing numerical counts or measurements

-Can be classified as either continuous or discrete

i) Continuous data

- Data that can take on any value in a given interval
- Can be a fraction or a whole number
- Examples
  - Weight; someone can weigh 150kgs
  - Time; the time can be part of an hour (12:48pm)
  - Distance; length can be measured in parts (2.36km)

ii) Discrete data

- Data that only takes whole number form
- Examples
  - Number of students (whole numbers only)
  - Number of times a student took their driver's test (whole numbers only)

Data can further be divided into;

Primary data

Data collected, presented and analysed by the investigator (researcher). It is the original data got from experiments or collecting information. Primary data can also be got from: -Enquiry surveys which involves interviewing the subjects of the study

-Use of questionnaires

-Direct observation

Advantages of primary data

-It is timely (NOT OUTDATED)

-It is relevant to the area you need

-It is accurate

Disadvantages of primary data

-It is expensive to collect

-It is time consuming

-At times it is difficult or impossible to collect

Secondary data

These are data published by someone else other than the person who collected, analysed and presented them. They are available from previous studies, compilations, investigations, researches etc. Sources of secondary data include textbooks, academic researches, magazines, newspapers, academic journals etc.

Advantages of secondary data

-It is cheap to collect

-It saves time

Disadvantages of secondary data

-Sometimes it is outdated

-Sometimes it is irrelevant

-It may be innacurate

Observation

This is a value that a variable has been found to have on a particular individual unit.

Population denoted by N

This is the group of individuals or items or objects or variables with the same characteristics about which we want to obtain information (data) e.g. Students of KIU. In studying the population, we use what we call-sample because;

-Population is very large and use of samples saves time, money and labour -It might be impossible to enumerate all the members of the population -To avoid destroying all members of the population -Accurate results are obtained from studying a sample

Sample Denoted by n

This is a group of individuals or objects drawn from a population for in depth study.

Advantages

-Cost is low

-Time saving

-Information from a sample can be used to draw valid conclusions about a population

-Complete enumeration of the population might be impossible

Parameters

Numerical summary measures used to describe a statistical population (such as the population average).

Sample statistics

Numerical summary measures used to describe a sample (such as the sample average) whose values vary according to the random sample collected.

Distributions

The distribution of the variable tells us what values the variable takes and how often each value occurs.

The frequency distribution of a (discrete) variable is the set of possible values of the variable together with its associated frequencies.

Sampling survey methods

This is the selection of elements from a population for an in-depth study. These elements could be items, objects, people etc……… the results from the sample are then used to generalize about the population.

Sampling frame

This is the list of elements in the population considered for selection into the sample when carrying out probability sampling

Types of sampling methods

i)Non-random (ii) Random

Non-random samples are those drawn without any selection processes that depend upon probabilities. Two main types may be identified

- Accessibility methods
- Purposive sampling methods

These are used for reasons of convenience, rather than for reasons of accuracy. They are relatively cheap and quick, but may give unrepresentative results

i) Non random sampling methods

Accessibility sampling (convenient sampling)

In this sampling scheme only the most easily accessible sampling units are selected.

Purposive sampling (judgmental sampling).

This is an improvement on accessibility sampling

The sample is chosen to be representative of the whole population by using personal judgment

Judgment is generally made by a field or enumerator, that is the person who will ask questions of the sampled people.

ii) Random sampling methods

Random or probability sampling methods depend upon some random process such as a random number table;

Every member of the parent population has known non-zero probability of being chosen for the sample.

There are a number of probability sampling methods, whichever one is best for any particular project depends upon the resources available and the degree of error that is acceptable. Probability sampling is preferable because it gives more reliable results for a given fixed sample size.

Simple random sampling

A simple random sampling every possible sample of size and from the population of size N has an equal chance of selection.

It may be shown that, as a consequence of this, every member of the member of the population has an equal chance of being included in the sample.

This is the type of sampling methods are based on the simple ransom sampling (s.r.s) approach. To actually draw a sample of size n we could take N identical strips of paper, one for each member of the population. Then ensures random selection, but can be tedious, especially with large populations! An alternative is to use random numbers.

Stratified random sampling

Sometimes the sampling units may be sub-divided into various mutually exclusive types known as strata. It may then be possible to draw a simple random sample (s.r.s) separately from the population of each type. The resulting sample is called a stratified sample.

Advantages

Stratified sampling has two major advantages

- With a sensible choice of stratification factor the stratified sample is expected to be more reliable than as.r.s of the same size.
- Estimators to known precision may be calculated for each individual stratum

Stratification spreads the sample more evenly across the population than the s.r.s could be expected to do.

Stratification is, however, only of use is a sensible stratification factor is used. This means choosing something, which is;

- Already known, easily found, for every member of the population
- Directly associated with (at least one of) the variables to be measured in the survey.

When sampling human populations, age and sex are frequently sensible stratification factors since they are often recorded on administrative registers and are often related to variables of interest, such as income, health and lifestyle.

Proportional stratification

One decision the must be taken when stratified sampling, is how to spread the sample across the strata. The simplest approach is to take a constant proportion form each stratum that is to ensure that;

$$\frac{n_1}{N_1} = \frac{n_2}{N_3}$$

This is called proportional stratification

Proportional stratification is not always possible. Proportional stratification is no necessarily the best way of spreading in the sample. Cost consideration may also determine the makeup of the stratified sample. If one stratum is very expensive sample we may have to restrict the size of its particular sample. Hence rural area may be deliberately underrepresented in a stratified sample. Disproportionate sampling is no problem, provided a large enough sample is recruited from each stratum is recruited from each stratum to provide a reliable estimate for the stratum. The weighted average technique used to produce overall population estimates, takes accounts of possible disproportionate sampling

Systematic random sampling method

Random sampling from a large population is cumbersome. An alternative procedure is to list the population in some order, for example alphabetically and choose kth member from the list after obtaining a random starting point. For example, if we choose every 10th member on the list (list of how many) we should form a 10 sample, and if we chose every 20th member we should form 1 5% sample. Such procedure is called systematic sampling

Cluster sampling

Sometimes, instead of sampling the individuals from who we wish to take responses, it is more convenient to sample groups, or cluster, of these individuals. The sample of clusters in chosen by some probability-sampling scheme, such as simple random sampling.

This approach is called cluster sampling. It is general used for one or both of the following reasons;

- To reduce field costs. For example a simple random sample of forest officers from some stations might well result in a sample spread over all the stations in the country. To visit all the stations would be very expensive. However, if we treated stations as clusters and then say, sampled five clusters, we would then only need to visit five places.
- Clustering is thus usually done for convenience, rather than to improve precision of the results.
- Estimates from cluster samples are most precise when cluster members are different in relation to the subject of interest and yet the average make-up of each in much the same. In this situation each cluster will give a reasonable representation, in miniature, of the whole.
- This is the exact opposite to the situation desired for strata, and it is important that the two are not confused- although both are groups of responding individuals. Clusters are sampling units but strata are not (we sample from all strata, but only take a selection of the clusters).
- Unfortunately, convenient clusters for sampling purpose are often internally similar for example-forest officers usually have much the same opinions and lifestyle and members of a particular station experience the same problems.
- In most surveys a cluster sample will actually decrease the precision compared with as.r.s of the same size.
- Often the lost precision is more than compensated by the reduced cost.

Multi-stage sample

- As in cluster sampling, multi-stage sampling considers the individual sampling units at more than one level.
- Instead of sampling the whole set of units in any one group, we randomly sample a sub-set of each higher-level group.
- What were called clusters" in previous sections are now called primary sampling units.
- The individuals within the primary sampling units are called the secondary sampling units.

For example we might sample districts at the first level of sampling, the countries within the districts at the second level, stations within counties at the third level and pupils within stations at the fourth and last level. This gives a four-stage sampling scheme.

The most appropriate number of level to use will vary with the problem at hand using more levels is likely to improve the accuracy (the sample is better spread) but increase cost (for the same reason). One practical limitation is that, at each stage, there must be asset of sampling frames available from which to make selection at the following stage.

Comparison of sampling methods

Simple random sampling, stratified sampling, cluster sampling and multi stage sampling can be compared in two ways

- Cost
- Precision

In general the better spread the sample is across the whole population, the more expensive but the more precise can the survey be expected to be. We can expect the following rank orderings from best to worst.

Cost saving	**Precision**
1. Cluster	1. Stratified
2. Multi-stage	2. Simple –random
3. Simple random	3. Multi-stage
4. Stratified	4. Cluster

The list go in completely opposite orders

If precision is the only consideration, stratified sampling should be used provided a sensible stratification factor is known

If cost saving is the paramount issue, cluster sampling may be the best.cluster and multi stage give savings because of reduction in travelling. In postal surveys this is usually not an issue, although it may still be convenient to sample postal sectors rather than individual addresses (at least at the first stage). Stratified sampling is usually only slightly more expensive than simple random sampling; the extra costs are mainly administrative.

Variable

A variable is a characteristic of a population or sample that can take more than one value such as household income can be denoted by X_1, X_2, X_3…….X_n. A variable can be either qualitative or quantitative.

Qualitative variable

A variable which describes characteristics which may be categorized rather than measured e.g. colour of hair/eyes, gender (male or female), blood group, marital status….Qualitative variables are non-numeric and non-measurable.

Quantitative variable

Observations are numeric and measurable. They are divided into discrete and continuous variables.

Discrete (discontinuous) variables are numerical variables but the values take only whole number forms e.g. number of goals scored, number of cars passing a given point.

Continuous variables are also numerical variables but the values are not restricted to specific values but can take any value within an interval e.g. height, weight, temperature.

Scales of measurement

- Typically all variables are written down using numbers. In the case of qualitative, number coding is done. E.g.
- Sex of patient: 1=female, 2=male
- Size of a manufacturing company; 1='small, 2=medium, 3=large
- Employment status 0=employed, 1=unemployed, 2 = not seeking work
- For qualitative variables the numbers used are simply labels
- Whilst the number two is clearly twice the number one and 40 students is twice 20 students we can not say a male is twice a female etc.
- Not all measurements carry the same sort of information
- What we can do with our data will depend on what information the numbers carry, i.e. what sort of measurement scale the variables are recorded on.
- Not all measurements carry the same sort of information the numbers carry, i.e. what sort of measurement scale the variables are recorded on.

Types of measurement scales

There are four types of scales; the nominal and ordinal scales, which measure qualitative data and the interval and ratio, which measure quantitative data.

Nominal scale

This is least powerful level of measurement. Nominal scale is a system of assigning of number symbols to events in order to identify them. We cannot average these numbers or compare assigned numbers of one group with numbers assigned to another group i.e. numbers have no quantitative value and cannot be ranked or ordered.

The counting of members in each group is the only possible arithmetic operation when a nominal scale is employed. Hence we are restricted to use mode as the measure of central tendency.

No measure of dispersion for nominal scale.

Chi-square test) is the most common test of statistical significance that can be used i.e. nominal data is counted data.

Nominal scales are widely used in survey research when data are being classified by major sub-groups of the population.

Ordinal scale

- Lowest level of the ordered scale
- Places the events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule.
- Rank order represent ordinal scales and are frequently used in research relating to qualitative phenomena
- A student's position in class is an example of ordinal scale

- Permits ranking of items from highest to lowest
- No absolute values, and the real differences between adjacent ranks may appropriate measure of central tendency is the median
- A percentile or quartile measure is used for measuring dispersion
- Correlations are restricted to rank correlation
- Measures of statistical significance are restricted to the non-parametric methods (not discussed in this course

Interval scale

- Intervals are adjusted in terms of some rule that has been established as a basis for making the units equal.
- Have an arbitrary zero i.e. lack a true zero (absolute zero)
- Temperature (e.g. Fahrenheit scale) is an example of an interval scale
- Provide more powerful measurement than ordinal scales for intervals scale. Also incorporates the concept of equality of intervals.
- More powerful statistical measures can be used with intervals scales.
- Arithmetic mean is an appropriate measure of dispersion
- Product moment correlation techniques are appropriate and the generally used tests for statistical significance are the ‘t’ and ‘F’ test.

Ratio scale

- Most powerful (precise) scale of measurement e.g. length, time weight, height, distance.
- All statistical techniques are usable with ratio scales and all manipulations that one can carry out with ratio scale values e.g. multiplication and divisions
- Arithmetic mean can be used as a measure of central tendency and coefficient of variation may also be calculated.

N.B:

- Researchers should use the scale that provides the most precise description
- Researchers in physical and biological sciences have the advantage to describe variables in ratio scale form but behavioral sciences are generally limited to describe variables in interval scale from, a less precise type of measurement.

Data

Discrete data

Table 1.1 shows the numbers of trees infected by pathogens in 30 randomly chosen home gardens

Table 1.1

1	3	2	3	2	1	3
2	5	3	1	2	2	2
4	2	2	3	0	2	3
0	2	0	2	3	3	2
2	3	1	3	4	1	1

This is an example of discrete data. The data is raw because it has not been ordered in any way. Discrete data take only exact values.

Arrangement of data

Count the number of times each value occurs and form a frequency distribution shown in table 1.2

Table 1.2. Frequency distribution

No. of trees	tally	Frequency (count)
0		3
1		5
2		11
3		8
4		2
5		1
Total		30

Ungrouped discrete data in the form of a frequency distribution can be represented diagrammatically by a vertical line graph in which the height of each line represents the frequency.

Vertical line graph to show the numbers of trees infected in 30 home gardens.

From the graph it can be seen that the mode is 2 trees per home garden are infected.

Continuous data

Table 1.3 shows the heights of 20 seedlings in nursery. The heights have been measured correct to the nearest cm

Table 1.3: Heights of 20 trees

133	136	120	138	133
131	127	141	127	143
130	131	125	144	128
134	135	137	133	129

This is an example of continuous data continuous data cannot take exact values, but can be given only within a certain range or measured to a certain degree of accuracy. E.g. 144cm (correct to the nearest cm) could have a risen from any value in the interval $143.5\text{cm} \leq h < 144.5\text{cm}$. To form a frequency distribution for the heights of the 20 seedlings we usually group the information into'classes' or 'intervals' for example (table 1.4).

Table 1.4 Classes

Two ways of writing classes	
119.5-124	120-124
124.5-129	125-129
129.5-134	120-134
134.5-139	135-139
139.5-144	140-144

The values 119.5, 124.5, 129.5………………..144.5 are called the class boundaries. The upper boundary (u.c.b)

Of one interval is the lower class boundary (l.c.b) of the next interval.

Interval width

The width of an interval = u.c.b-1.c.b

Therefore the width of the first interval = 124.5-119.5=5cm. indeed the width for each interval is 5.

Tallying can be done to group the heights. The final frequency should read as shown in table 1.5.

Table 1.5: Frequency distribution with tallies

Height (cm)	Tally	Frequency
119.5-124.5		1
124.5-129.5		5
129.5-134.5		7
134.5-139.5		4
139.5-144.5		3
Total		20

Disadvantage

When the data are presented in a frequency distribution, the original information is lost. For instance, we do not know the value of the one item in the first interval only that it lays between 119.5 and 124.5cm.

Calculating mean: example 1

3 methods

1. Input in the calculator and then read off the mean
2. Calculating mean when data are grouped using a frequency table
3. Using computer programs e.g. Excel

20	28	26	20
20	28	23	22
25	20	21	25
25	24	29	23
20	29	20	25

Exercise 1a:

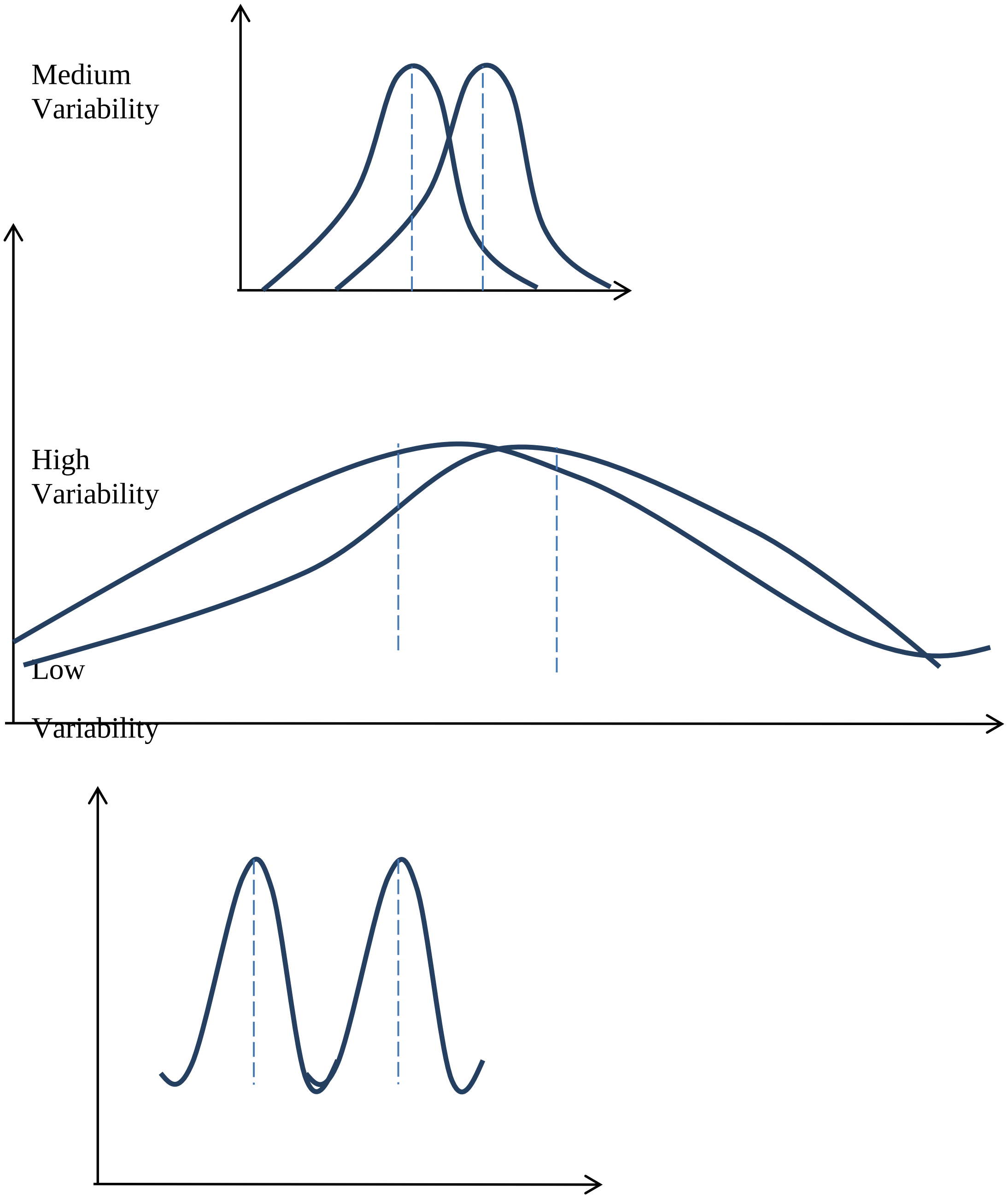
Calculate the mean using frequency table

25	38	36	30	33	36
26	38	33	32	31	37
27	30	31	35	30	38
28	34	29	33	37	39
29	29	30	35	31	40

Standard deviation

- Measure of how spread data are from the mean
- σ(sigma)-is SD of population
- S=SD for sample, s^2=variance
- Variance is used to compare data sets with normal distribution
- SD is the most widely used measure of variability

Larger and smaller SD



Standard deviation for ungrouped data

Standard deviation of a population:$\sigma_x = \sqrt{\frac{\sum(X-\mu_x)^2}{N}}$

Standard deviation of a population:$S_x = \sqrt{\frac{\sum(X-\bar{X})^2}{N-1}}$

Standard deviation for grouped data

$$\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{\sum f}}$$

Calculating SD: Example 2

Ungrouped data

Grouped using a frequency table

Use data of example 1

Exercise 1b: Calculate the SD of data in Ex. 1a

Null & Alternative hypothesis

A hypothesis in a proposal or a report presupposes a statistical test

Null: a statement that there is no relationship between variables (H_o or H_N)

Alternative: a statement that suggests a potential outcome one may expect (H_1 or H_A)

- H_o “There is no difference between……….”
- H_1 the mean weight of males is higher than that of females
- $H_{o:}$ There is no significant difference in the anxiety level of children of High IQ and those of low IQ
-
- H_1: The anxiety level of children of High IQ is lower than those of low IQ

The final conclusion of the study will either accept or reject the hypothesis being tested

Parametric and non-parametric tests

Parametric and non-parametric tests		
Measure	Parametric	Non-parametric
	Mean	Median
Variance	Homogeneous	Any (F-test)
Distribution	Normal	
Typical data	Ratio/interval	Ordinal
2 paired samples	Paired t-test	Wilcoxon test
Independent, 2 samples	t-test	Mann-Whitney U
2 sample means	z-test	
Correlation	Pearson	Spearman’s
2 Independent samples	One-way Anova	Kruskal-Wallis test
2 Independent matched samples	Two-way Anova	Friedman test
Frequency data		Chi-Square test

The normal curve

Properties

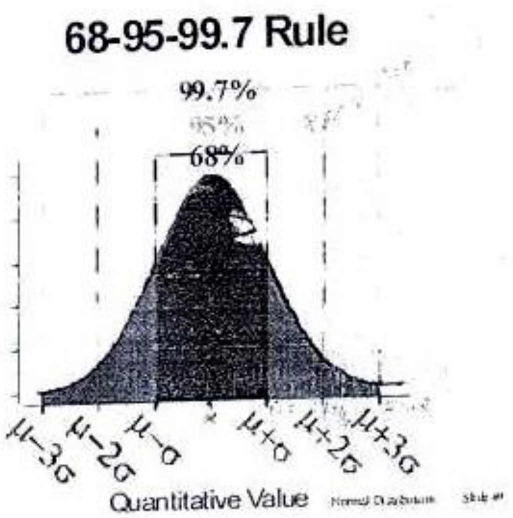
1. Bell shaped
2. Symmetrical around mean
3. Mean=mode=median
4. Obeys the 68-95-99 rule
    - 68% obs. Are 1SD from $\mu(\mu \pm 1SD)$
    - 95%-2SD away from $\mu(\mu \pm 2SD)$
    - 99.7%-3SD away from $\mu(\mu \pm 3SD)$

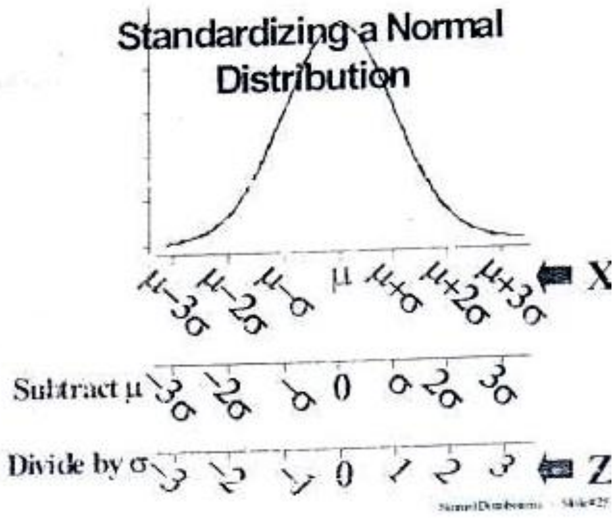
The standard normal curve: the Z-score

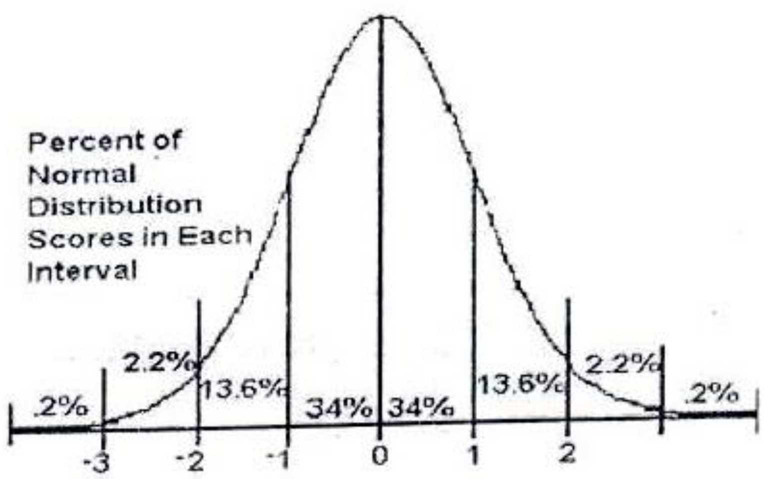
- The No. of SD units the observation is waya from
- If n>30, n is large, μ
- Z=(x- $\mu)/\sigma, Z = (x-)/s$

Properties of a standardized normal curve

1. Bell shaped
2. Mean=mode=median=0
3. SD=1
4. It obeys the 68-95-99 rule
    - 68% obs. lie within ± 1 from mean
    - 95% ± 2away from mean
    - 99.7% ± 3 way a from mean







It is calculated that

95% observations lie within ±1.96 (instead of ±2):

P(Obs outside 95%)<0.05

If Z(cal) is <1.96, means of 2 populations being compared are similar

No significant difference (P=0.05); Accept Ho

If Z(cal) is >1.96, means of 2 populations are different

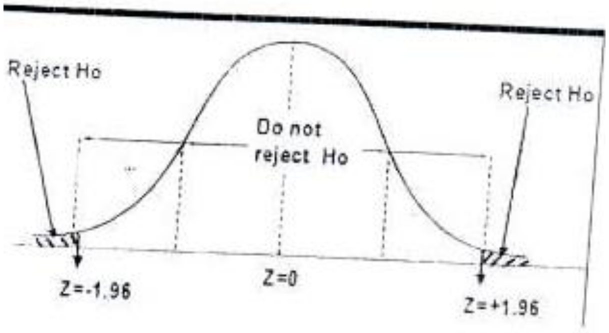
Significant at p=0.05; Reject Ho

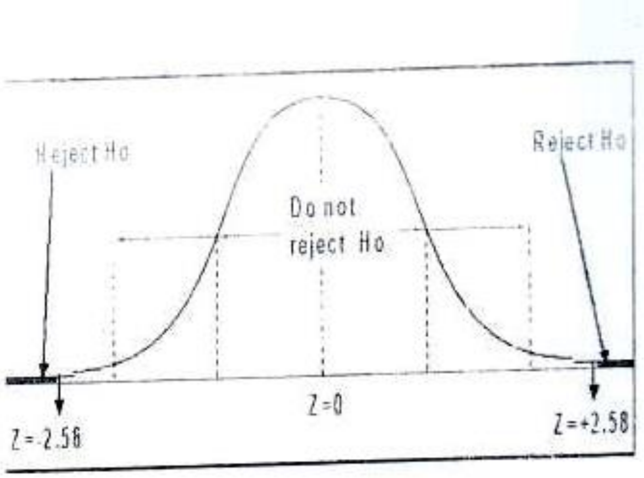
99% Obs lie within ±2.58 (instead of ±3

- P(Obs outside 99%)<0.01
- If Z(cal) is <2.58, means of pop I & II being compared are similar

- No significant difference at P=0.01; Accept Ho
- If Z(cal) is >2.58, means of pop I & II being compared are different
- Significantdifference at P=0.01; Reject Ho

This is called a one sample z-test





Examples of one sample z-test

Used to compare whether an individual found belongs to a large population of known mean and SD

Example

1. On the basis of large samples, the mean length of a population of seeds is estimated to be 5.8mm and SD is 0.17mm. Is it likely that randomly selected seeds of lengthy 4.3 mm belong to this population (Use P=0.05).
2. Show whether a migratory fish of length 20.5cm could have come from a large population of mean length 25.5cm and standard deviation 0.26cm
3. How true is it that a migratory bird of length 32.5cm came from a population of 45 individuals whose mean length and SD are 25.7am and 0.28cm respectively?

Z-test for compare means of 2 large samples (n>30)

- 2 sample test
- Used to compare means of 2 large populations
- Z-test is preceded by the F-test

Formula

- Z_{cal}>1.96, 2.58 reject Ho at P=0.05 and 0.01 respectively
- Z_{cal}<1.96, 2.58 accept Ho at P=0.05 and 0.01 respectively

F-Test

- F-test test the similarity of variance
- Requirement for a parametric test
- Formula: larger variance/smaller variance
- $H_o=V_1=_{V2,}$df (N_1-1, n_2-1) appendix 8
- $F_{cal}<F_{tab}$, accept H_o (similar variances)
- Proceed with the parametric test
- If $F_{cal}>F_{tab}$, we reject H_o (different variances)
    - Non parametric analysis applies or data are transformed

Example : 2 sample Z-test

John and joseph were found preparing to sell fresh immature Tilapia zilli that were measured and found with the following parameters below. Test statistically whether the two men could have obtained the fish from the same source

	John	Joseph
No. Indiv	60	48
Mean length	56.6	37.4
SD	0.7	0.8