Unit 3

Methods of Data Collection
Sampling Design
Data Analysis

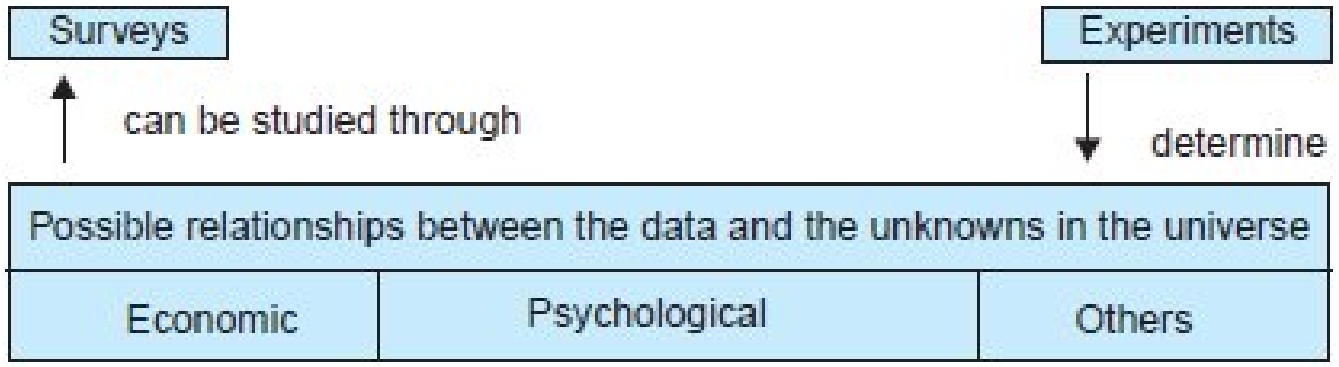
Methods of Data Collection

Data Collection

- The task of data collection begins after a research problem has been defined and research design/plan chalked out
- Two types of data viz., primary and secondary.
- The *primary data* are those which are collected afresh and for the first time, and thus happen to be original in character.
- The *secondary data* are those which have already been collected by someone else and which have already been passed through the statistical process
- The methods of collecting primary and secondary data differ

Collecting Primary Data

- During the course of doing experiments in an experimental research
- In descriptive research - perform surveys, whether sample surveys or census surveys,
  - through observation
  - through direct communication with respondents
  - through personal interviews.



Methods of collecting primary data

- (i) observation method
- (ii) interview method
- (iii) questionnaires
- (iv) schedules

(i) Observation

- Observation becomes a scientific tool and the method of data collection for the researcher,
- serves a formulated research purpose
- is systematically planned and recorded
- is subjected to checks
- controls on validity and reliability
- the information is sought by way of investigator’s own direct observation without asking from the respondent

(i) Observation

- Advantages
  - subjective bias is eliminated
  - Information relates to what is currently happening
  - independent of respondents' willingness to respond or cooperate
  - deal with respondents who are not capable of giving verbal reports of their feelings
- Limitations
  - expensive method
  - Limited information
  - unforeseen factors may interfere during observation

(ii) interview method

- involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses
- Through personal interviews or telephone interviews
- *Personal interviews* - Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person
- Telephone interviews – conducting the interview over the phone
- *structured interviews* - use of a set of predetermined questions and of highly standardized techniques of recording. The interviewer in a structured interview follows a rigid procedure laid down, asking questions in a form and order prescribed

(ii) interview method

- *unstructured interviews* – flexibility of approach to questioning.
- Unstructured interviews do not follow a system of pre-determined questions and standardized techniques of recording information.
- The interviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at times he may omit certain questions if the situation so requires.
- change the sequence of questions.
- this sort of flexibility results in lack of comparability of one interview with another and the analysis of unstructured responses becomes much more difficult and time-consuming

Advantages of interview method

- (i) More information and that too in greater depth can be obtained.
- (ii) Interviewer by his own skill can overcome the resistance, if any, of the respondents;
- (iii) There is greater flexibility
- (iv) applied to recording verbal answers to various questions.
- (v) Personal information can as well be obtained easily under this method.
- (vi) Samples can be controlled more effectively
- (vii) The interviewer can usually control which person(s) will answer the questions.
- (viii) The interviewer may catch the informant off-guard and thus may secure the most spontaneous
- reactions
- (ix) The language of the interview can be adopted to the ability or educational level of the person interviewed
- (x) The interviewer can collect supplementary information about the respondent's personal characteristics

Limitations of interview method

- (i) It is a very expensive method, specially when large and widely spread geographical sample is taken.
- (ii) There remains the possibility of the bias of interviewer as well as that of the respondent;
- (iii) Certain types of respondents such as important officials may not be easily approachable
- (iv) This method is relatively more-time-consuming
- (v) The presence of the interviewer on the spot may over-stimulate the respondent
- (vi) selecting, training and supervising the field-staff is more complex
- (vii) Interviewing at times may also introduce systematic errors.
- (viii) Effective interview presupposes proper rapport with respondents

(iii) questionnaires

- Popular in case of big enquiries
- Adopted by private individuals, research workers, private and public organizations and even by governments
- A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms.
- The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply
- The respondents have to answer the questions on their own.

Advantages of questionnaires

- 1. There is low cost
- 2. It is free from the bias of the interviewer; answers are in respondents' own words.
- 3. Respondents have adequate time to give well thought out answers.
- 4. Respondents, who are not easily approachable, can also be reached conveniently.
- 5 Large samples can be made use of and thus the results can be made more dependable and reliable

Limitations of questionnaires

- 1. Low rate of return of the duly filled in questionnaires
- 2. It can be used only when respondents are educated and cooperating.
- 3. The control over questionnaire may be lost once it is sent.
- 4. Difficulty of amending the approach once questionnaires have been despatched.
- 5. Possibility of ambiguous replies or omission of replies
- 6. difficult to know whether willing respondents are truly representative.
- 7. slow method

Main aspects of questionnaires

- General form: Structured or unstructured questionnaire – Closed / open type questions
- Question sequence: Proper sequence of questions reduces the possibility of questions being misunderstood. Avoid questions that strain memory, questions of personal character and questions related to personal wealth at first. Determine question sequence using pilot survey. Question sequence from general to more specific.
- Question formulation and wording: Question should be simple, understandable, concrete and conform to respondent's way of thinking. MCQ's and closed questions.

Main aspects of questionnaires

- Question formulation and wording: Open ended questions are difficult to handle, raising problems of interpretation, comparability and interviewer bias. No good questionnaire will have one type of questions. Words with ambiguous meanings to be avoided.
- Essentials: Appropriate number of questions – Logical sequence – No technical terms or vague expressions – Questions may be dichotomous, multiple choice or open-ended. Questions affecting sentiment of the respondents should be avoided. Appearance of the questionnaire is also significant.

(iv) schedules

- like the collection of data through questionnaire
- schedules (proforma containing a set of questions) are being filled in by the enumerators who are specially appointed for the purpose.
- Enumerators put to them the questions from the proforma in the order the questions are listed
- record the replies in the space meant for the same in the proforma.
- schedules may be handed over to respondents and enumerators may help them in recording their responses
- Enumerators explain the aims and objects of the investigation and also remove the difficulties which any respondent may feel in understanding

(iv) schedules

- requires the careful selection of enumerators for filling up schedules
- The enumerators should be trained to perform their job well
- the nature and scope of the investigation should be explained to them thoroughly
- Enumerators should be intelligent and must possess the capacity of cross examination
- they should be honest, sincere, hardworking and should have patience and perseverance.
- useful in extensive enquiries and can lead to fairly reliable results.

- very expensive
- usually adopted in investigations conducted by governmental agencies or by some big organizations.
- Population census all over the world is conducted through this method.

Questionnaire vs Schedules

- Questionnaire answered by respondent and schedule by enumerators
- Data collection through schedule is relatively more expensive as one has to appoint enumerator and train.
- Non-response and bias is more for questionnaire
- Identity of respondent known in data collection through schedule
- Questionnaire data collection is very slow and no personal contact
- Questionnaire data collection possible only with literate respondent
- Difficulty in sending more enumerators over a relatively wide area
- Success of questionnaire depends on quality and physical appearance of questionnaire

Collecting Secondary Data

- Secondary data must possess following characteristics:

  1. Reliability of data: The reliability can be tested by finding out such things about the said data: (a) Who collected the data? (b) What were the sources of data? (c) Were they collected by using proper methods (d) At what time were they collected?(e) Was there any bias of the compiler? (t) What level of accuracy was desired? Was it achieved ?

  2. Suitability of data: The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. Hence, if the available data are found to be unsuitable, they should not be used by the researcher

  3. Adequacy of data: If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher.

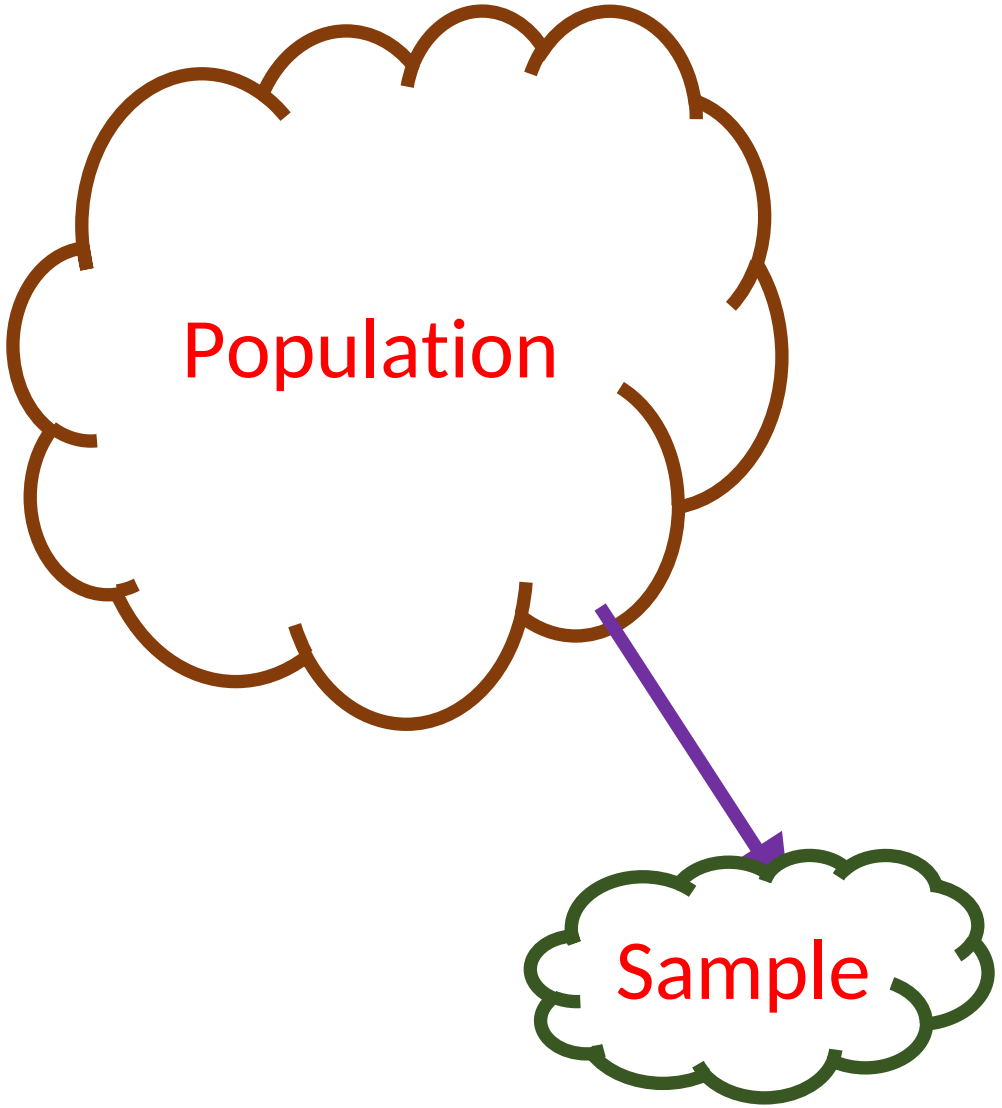
Collecting Secondary Data

- Secondary data may either be published data or unpublished data
- published data are available in: (a) various publications of the central, state are local governments; (b) various publications of foreign governments or of international bodies and their subsidiary organisations; (c) technical and trade journals; (d) books, magazines and newspapers; (e) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.; (f) reports prepared by research scholars, universities, economists, etc. in different fields; and (g) public records and statistics, historical documents, and other sources of published information.
- The sources of unpublished data - diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labour bureaus and other public/private individuals and organisations.

Sampling Design

Sampling Fundamentals

- A complete enumeration of all items in the ‘population’ is known as a census inquiry.
- Sampling is the selection of some part of an aggregate or totality on the basis of which a judgement or inference about the aggregate or totality
- it is the process of obtaining information about an entire population by examining only a part of it.
- To draw inferences based on samples about the parameters of population from which the samples are taken.



NEED FOR SAMPLING

1. Sampling can save time and money.
2. Sampling may enable more accurate measurements when conducted by trained and experienced investigators.
3. Sampling remains the only way when population contains infinitely many members.
4. Sampling remains the only choice for destructive testing.
5. Sampling usually enables to estimate the sampling errors

Important Definitions

1. *Universe/Population:* 'Universe' refers to the total of the items or units in any field of inquiry; 'population' refers to the total of items about which information is desired.

2. *Sampling frame:* The elementary units or the group or cluster of such units, called as sampling units. A list containing all such sampling units is known as sampling frame.

Thus sampling frame consists of a list of items from which the sample is to be drawn. If the population is finite and the time frame is in the present or past, then it is possible for the frame to be identical with the population.

In most cases they are not identical because it is often impossible to draw a sample directly from population. As such this frame is either constructed by a researcher for the purpose of his study or may consist of some existing list of the population.

For instance, one can use telephone directory as a frame for conducting opinion survey in a city. Whatever the frame may be, it should be a good representative of the population.

Important Definitions

3. *Sampling design:* A sample design is a definite plan for obtaining a sample from the sampling frame.

Technique or the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn

4. *Statistic(s) and parameter(s):* A statistic is a characteristic of a sample, whereas a parameter is a characteristic of a population.

Measures such as mean, median, mode of the samples are called statistic(s)

Measures such as mean, median, mode of the population are called as parameter(s).

Important Definitions

5. *Sampling error:* Sample surveys do imply the study of a small portion of the population and hence, there would naturally be a certain amount of inaccuracy in the information collected.

- This inaccuracy may be termed as sampling error or error variance.

6. *Precision:* Precision is the range within which the population average (or other parameter) will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate ± or as a numerical quantity.

- if the estimate is Rs 4000 and the precision desired is ± 4%, then the true value will be in the range(Rs 3840 to Rs 4160)

Important Definitions

7. *Confidence level and significance level:*

The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits.

Example- a confidence level of 95%, then we mean that there are 95 chances in 100 that the sample results represent the true condition of the population.

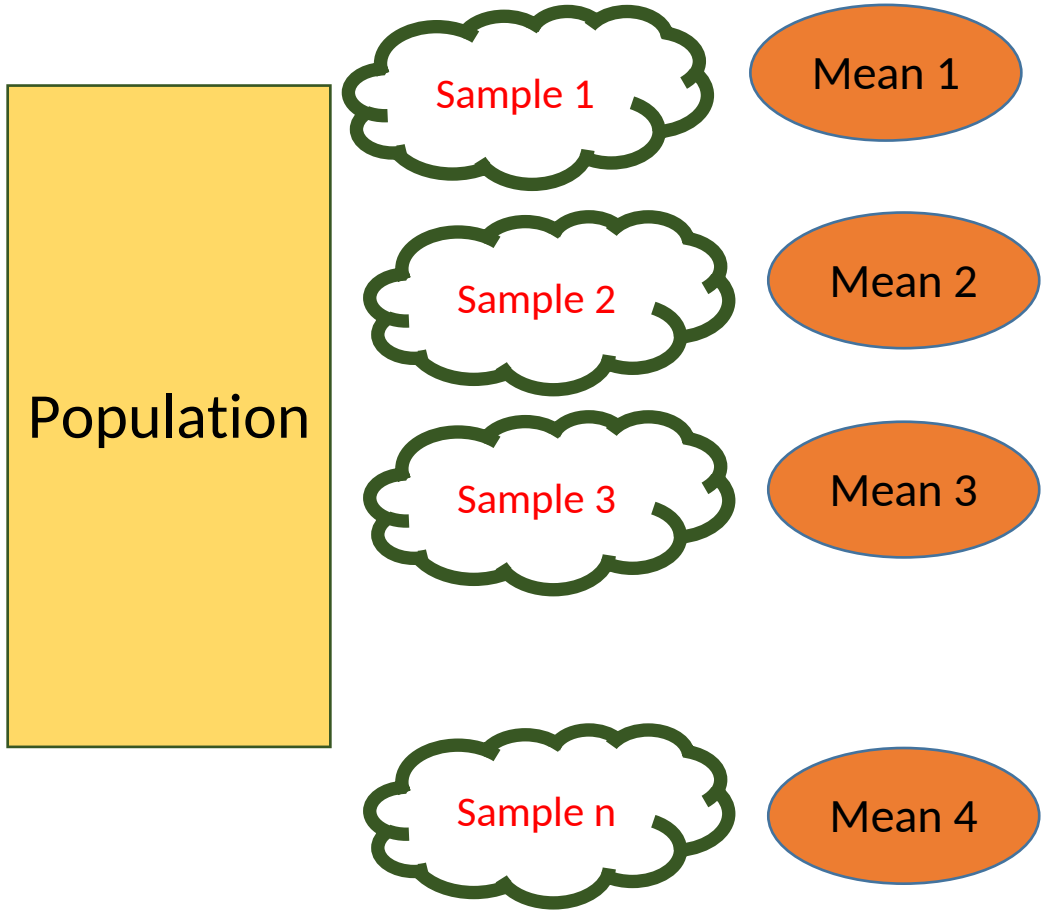
The significance level indicates the likelihood that the result will fall outside that range.

If the confidence level is 95%, then the significance level will be (100 – 95) i.e., 5%;

Important Definitions

8. *Sampling distribution:*

- If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may have its own value for mean
- All these sample means, together with their relative frequencies, will constitute the sampling distribution of mean
- Similarly, sampling distribution of standard deviation or the sampling distribution of any other statistical measure.
- The sampling distribution tends quite closer to the normal distribution if the number of samples is large.
- The significance of sampling distribution - the mean of a sampling distribution is the same as the mean of the universe.



STEPS IN SAMPLE DESIGN

- (i) **Type of universe:** The universe can be finite or infinite.
- (ii) **Sampling unit:**Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual.
- (iii) **Source list:** It is also known as ‘sampling frame’ from which sample is to be drawn
- (iv) **Size of sample:** This refers to the number of items to be selected from the universe to constitute a sample.
- (v) **Parameters of interest:** In determining the sample design, one must consider the question of the specific population parameters which are of interest. Example – average height of people
- (vi) **Budgetary constraint**
- (vii) **Sampling procedure**

CRITERIA FOR SELECTING A SAMPLING PROCEDURE

- **1. Inappropriate sampling frame:** If the sampling frame is inappropriate i.e., a biased representation of the universe, it will result in a systematic bias.
- **2. Defective measuring device:** If the measuring device is constantly in error, it will result in systematic bias.
- **3. Non-respondents:** If we are unable to sample all the individuals initially included in the sample, there may arise a systematic bias.
- **4. Indeterminancy principle:** Sometimes we find that individuals act differently when kept under observation than what they do when kept in non-observed situations.
- **5. Natural bias in the reporting of data:** Natural bias of respondents in the reporting of data is often the cause of a systematic bias

Sampling errors

- *Sampling errors* are the random variations in the sample estimates around the true population parameters.
- *Sampling error* can be measured for a given sample design and size.
- The measurement of sampling error is usually called the ‘precision of the sampling plan’.
- If we increase the sample size, the precision can be improved.
- But increasing the size of the sample has its own limitations viz., a large sized sample increases the cost of collecting data and also enhances the systematic bias.
- The effective way to increase precision is usually to select a better sampling design which has a smaller sampling error for a given sample size at a given cost.

IMPORTANT SAMPLING DISTRIBUTIONS

(1) sampling distribution of mean;

(2) sampling distribution of proportion;

(3) student's '*t*' distribution;

(4) *F* distribution; and

(5) Chi-square distribution

(1) sampling distribution of mean

- Measurable data, where mean can be computed.
- Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population
- If samples are taken from a normal population, the sampling distribution of mean would also be normal
- If sampling is from a population which is not normal even then the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large

(2) sampling distribution of proportion

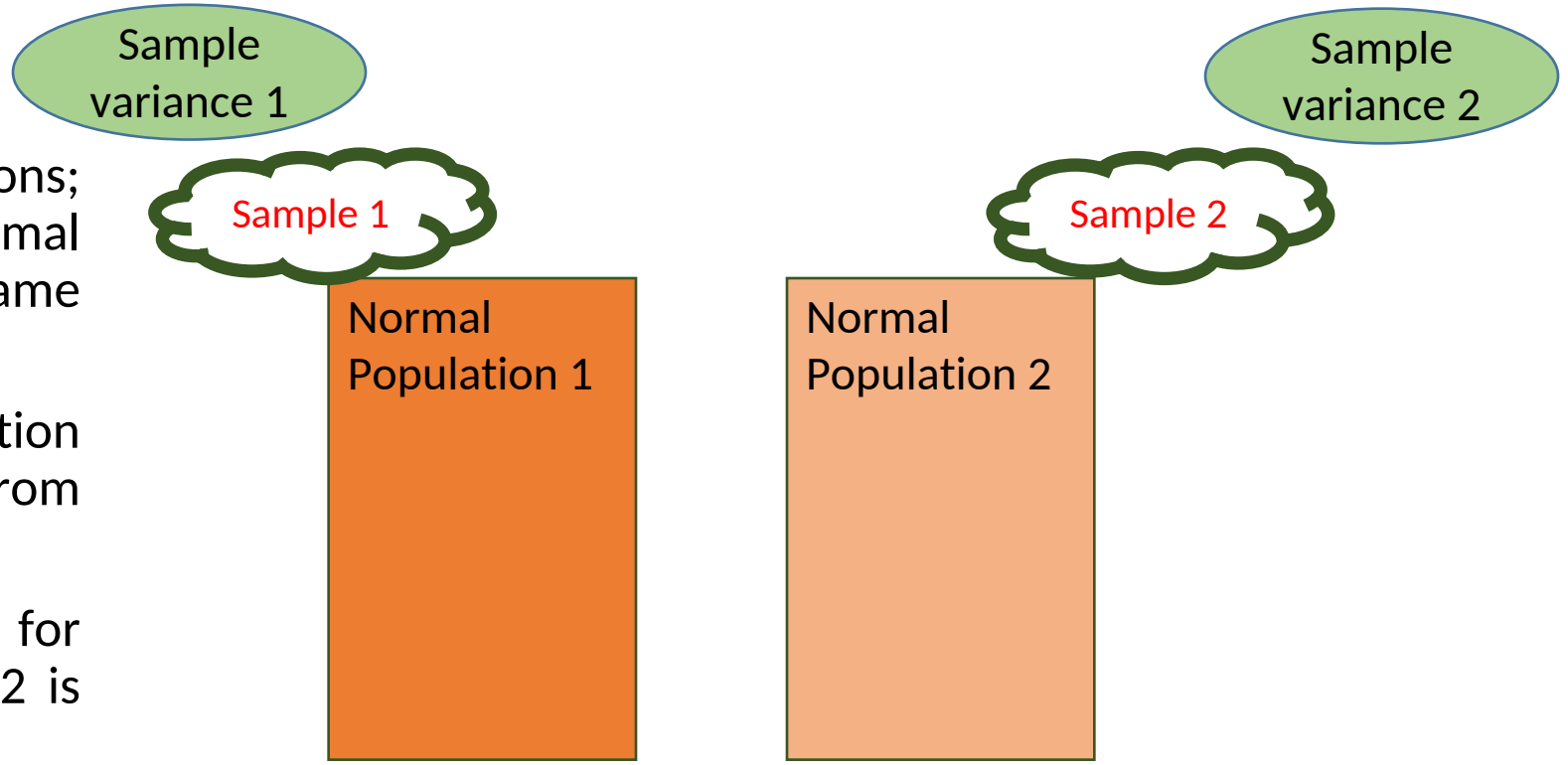
- In case of attribute data (good / bad, go/ no-go, defective/non-defective data)
- Example - proportion of defective parts in large number of samples that have been taken from an infinite population
- The probability distribution of these proportions known as the sampling distribution of the proportions
- For attribute data, the sampling distribution is Binomial distribution. But when the sample size (n) becomes larger and larger, the sampling distribution tends to become normal distribution

(3) student's '*t*' distribution

- When population standard deviation is not known and the sample size is small size ($n < 30$), we use *t* distribution for the sampling distribution of mean
- t-distribution is symmetrical
- As sample size (n) gets larger, the shape of the *t* distribution becomes approximately equal to the normal distribution.

(4) *F* distribution

- Consider 2 different populations; these populations follow normal distribution and have same variance.
- Sample 1 is drawn from population 1 and sample 2 is drawn from population 2.
- Sample variance 1 is computed for sample 1 and sample variance 2 is computed for sample 2.
- To analyse the differences between these sample variances, F-distribution is used.



These 2 populations have same variance

(5) Chi-square distribution

- Chi-square distribution is not symmetrical
- all the values are positive
- variances of samples require us to add a collection of squared quantities and thus have distributions related to chi-square distribution.
- This distribution is used for judging the significance of difference between observed and expected frequencies
- To test the goodness of fit (how well the assumed probability distribution fits the data)

CENTRAL LIMIT THEOREM

- When sampling is from a normal population, the means of samples drawn from such a population are themselves normally distributed.
- But when sampling is not from a normal population, the size of the sample plays a critical role.
- Whether the population follows normal distribution or not, the means of the samples drawn from that population follow normal distribution. (for sample size, n>30)

STANDARD ERROR

- The standard deviation of sampling distribution of a statistic is known as its standard error (S.E)
- helps in testing whether the difference between observed and expected frequencies could arise due to chance (inherent) causes in the process.
- gives an idea about the reliability and precision of a sample
- enables us to specify the limits within which the parameters of the population are expected to lie

CHARACTERISTICS OF A GOOD SAMPLE DESIGN

(a) Sample design must result in a truly representative sample.

(b) Sample design must be such which results in a small sampling error.

(c) Sample design must be viable in the context of funds available for the research study.

(d) Sample design must be such so that systematic bias can be controlled in a better way.

(e) Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

Types of Sample Design

- There are different types of sample designs based on two factors
  - -the representation basis
  - - the element selection technique

- Representation basis - probability sampling or non-probability sampling
  - Probability sampling is based on the concept of random selection
  - non-probability sampling is 'non-random' sampling

- Element selection basis, the sample may be either unrestricted or restricted.
  - 'unrestricted sample' - each sample element is drawn individually from the population at large
  - 'restricted sampling' – all other forms of sampling

CHART SHOWING BASIC SAMPLING DESIGNS

Element selection technique ↓ / Representation basis →	Probability sampling	Non-probability sampling
Unrestricted sampling	Simple random sampling	Haphazard sampling or convenience sampling
Restricted sampling	Complex random sampling (such as cluster sampling, systematic sampling, stratified sampling etc.)	Purposive sampling (such as quota sampling, judgement sampling)

Non-probability sampling

- It is the sampling procedure which does not have basis for estimating the probability that each item in the population has of being included in the sample.
- It is also known as deliberate sampling, purposive sampling and judgement sampling.
- Items for the sample are selected deliberately by the researcher
- His choice concerning the items remains supreme.
- the organizers of the inquiry purposively choose the particular units of the universe in the sample.
- It will be typical or representative of the whole population.
- the judgement of the organizers of the study plays an important part in this sampling design.
- Example - if economic conditions of people living in a state are to be studied, a few towns and villages may be purposively selected for intensive study on the principle that they can be representative of the entire state.

Non-probability sampling

- Personal element (organizer's bias) has a great chance of entering into the selection of the sample.
- The investigator may select a sample which shall yield results favorable to his point of view
- there is always the danger of bias entering into this type of sampling technique.
- If the investigators are impartial, work without bias and have the necessary experience so as to take sound judgement, the results obtained may be tolerably reliable.
- Sampling error cannot be estimated; the element of bias, great or small, is always there.
- In small inquiries and researches by individuals, this design may be adopted because of the relative advantage of time and money
- **Quota sampling** is also an example of non-probability sampling. Here, the interviewers are simply given quotas to be filled from the different strata, with some restrictions on how they are to be filled.
- The actual selection of the items for the sample is left to the interviewer's discretion.
- This type of sampling is very convenient and is relatively inexpensive.
- The samples so selected certainly do not possess the characteristic of random samples.
- Quota samples are essentially judgement samples and inferences drawn on their basis are not amenable to statistical treatment in a formal way.

Probability sampling

- It is also known as ‘random sampling’ or ‘chance sampling’.
- every item of the universe has an equal chance of inclusion in the sample.
- It is, so to say, a lottery method in which individual units are picked up from the whole group not deliberately but by some mechanical process.
- It is blind chance alone that determines whether one item or the other is selected.
- The results obtained from probability or random sampling can be assured in terms of probability i.e., we can measure the errors of estimation or the significance of results obtained from a random sample. Hence, the **superiority of random sampling design over the deliberate sampling** design.
- It ensures the law of Statistical Regularity - if on an average the sample chosen is a random one, the sample will have the same composition and characteristics as the universe.
- This is the reason why random sampling is considered as the best technique of selecting a representative sample.

Probability sampling

- A method of sample selection which gives each possible sample combination an equal probability of being picked up and each item in the entire population to have an equal chance of being included in the sample.
- This applies to **sampling without replacement** i.e., once an item is selected for the sample, it cannot appear in another sample again
- Implications of probability or random sampling are:
  - (a) It gives each element in the population an equal probability of getting into the sample; and all choices are independent of one another.
  - (b) It gives each possible sample combination an equal probability of being chosen.

- Note: Sampling with replacement – the item selected for the sample is returned to the population before the next item is selected for the sample. The same item could appear twice in the same sample).

Measurement and Scaling Techniques

MEASUREMENT IN RESEARCH

- We measure physical objects as well as abstract concepts.
- By measurement we mean the process of assigning numbers to objects or observations
- measurement is a process of mapping aspects of a domain onto other aspects of a range according to some rule

Levels of statistical measurement

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale

Nominal Scale

- Also called the categorical variable scale, is defined as a scale that labels variables into distinct classifications and doesn't involve a quantitative value or order.

Nominal Scale Examples

- Gender
- Political preferences
- Place of residence

What is your Gender?	What is your Political preference?	Where do you live?
• M- Male • F- Female	• 1- Independent • 2- Democrat • 3- Republican	• 1- Suburbs • 2- City • 3- Town

Ordinal Scale

- A variable measurement scale used to simply depict the order of variables and not the difference between each variable.
- the order of variables is of prime importance
- Analyzing results based on the order along with the name becomes a convenient process for the researcher
- Example – Likert Scale as shown

How satisfied are you with our services?

- Very Unsatisfied – 1
- Unsatisfied – 2
- Neutral – 3
- Satisfied – 4
- Very Satisfied – 5

Interval Scale

- a numerical scale where the variables' order is known and the difference between these variables.
- Variables that have familiar, constant, and computable differences are classified using the Interval scale.
- Example: Eighty degrees is always higher than 50 degrees, and the difference between these two temperatures is the same as the difference between 70 degrees and 40 degrees

Ratio Scale

- variable measurement scale that not only produces the order of variables but also makes the difference between variables known, along with information on the value of true zero.

What is your daughter's current height?

- Less than 5 feet.
- 5 feet 1 inch – 5 feet 5 inches
- 5 feet 6 inches- 6 feet
- More than 6 feet

What is your weight in kilograms?

- Less than 50 kilograms
- 51- 70 kilograms
- 71- 90 kilograms
- 91-110 kilograms
- More than 110 kilograms

LEVELS OF MEASUREMENT

Nominal
Named variables

Ordinal
Named
+
Ordered variables

Interval
Named
+
Ordered
+
Propotionate interval between variables

Ratio
Named
+
Ordered
+
Propotionate interval between variables
+
can accommodate absolute zero

Types of Measurement Scales

Nominal scale

It's used to label variables in different classifications and does not imply a quantitative value or order.

Ordinal Scale

It's used to represent non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc.

Interval Scale

It's defined as a numerical scale where the order of the variables as well as the difference between these variables is known.

Ratio Scale

It's a variable measurement scale that not only produces the order of the variables, but also makes the difference between the known variables along with information about the value of the true zero.

Sources of Error in Measurement

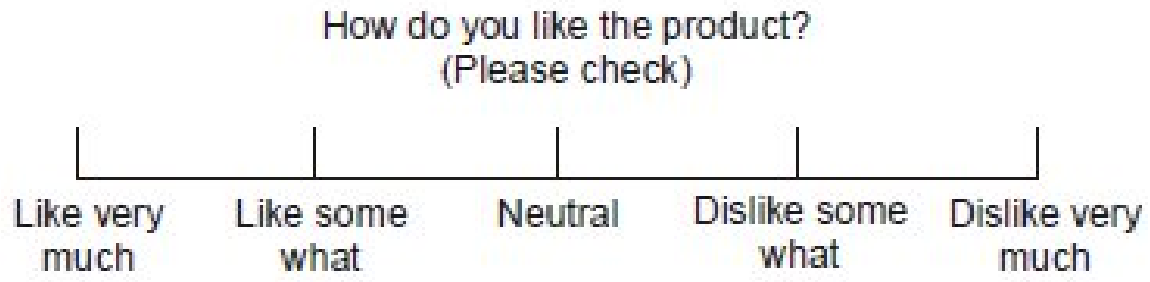
- **(a) Respondent:** At times the respondent may be reluctant to express strong negative feelings or it is just possible that he may have very little knowledge but may not admit his ignorance
- **(b) Situation:** Any condition which places a strain on interview can have serious effects on the interviewer-respondent rapport. Example - if someone else is present
- **(c) Measurer:** The interviewer can distort responses by rewording or reordering questions
- **(d) Instrument:** Error may arise because of the defective measuring instrument

Meaning of Scaling

- Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude and other concepts.
- This can be done in two ways viz., (i) making a judgement about some characteristic of an individual (ii) constructing questionnaires

Scaling Technique - Rating scale

- **Rating scale :** involves qualitative description of a limited number of aspects of a thing or of traits of a person. When we use rating scales (or categorical scales), we judge an object in absolute terms against some specified criteria
- The results obtained from their use, compare favorably with alternative methods.
- They require less time, are interesting to use and have a wide range of applications.
- They may be used with a large number of properties or variables.



Data Analysis

Testing of Hypotheses

- Hypothesis - a mere assumption or some supposition to be proved or disproved
- For a researcher, hypothesis is a formal question that he intends to resolve
- Hypothesis may be defined as a proposition or a set of propositions put forth as an explanation for the occurrence of some phenomena
- Examples –
  - "Students who receive counselling will show a greater increase in creativity than students not receiving counselling"
  - "the automobile *A* is performing as well as automobile *B*."

Null hypothesis and alternative hypothesis

- **Null Hypothesis (H_0)**
  - This can be thought of as the implied hypothesis.
  - “Null” meaning “nothing.”
  - This hypothesis states that there is no difference between groups or no relationship between variables.
  - The null hypothesis is a presumption of status quo or no change.
  - The researcher wishes to disprove null hypothesis
- **Alternative Hypothesis (H_a)**
  - This is also known as the claim
  - This hypothesis should state what you expect the data to show
  - This is your answer to your research question

Null hypothesis and alternative hypothesis Example

test the hypothesis that the population mean (μ) is equal to the hypothesised mean $(\mu_{H_0}) = 100$.

$$H_0 : \mu = \mu_{H_0} = 100$$

Alternative hypothesis	*To be read as follows*
$H_a : \mu \neq \mu_{H_0}$	(The alternative hypothesis is that the population mean is not equal to 100 i.e., it may be more or less than 100)
$H_a : \mu > \mu_{H_0}$	(The alternative hypothesis is that the population mean is greater than 100)
$H_a : \mu < \mu_{H_0}$	(The alternative hypothesis is that the population mean is less than 100)

Select the relevant alternative hypothesis for the research

level of significance

- The significance level is the maximum value of the probability of rejecting *H*0 when it is true

- In case we take the significance level at 5 per cent, then this implies that *H*0 will be rejected when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if *H*0 is true.

- It is usually determined in advance before testing the hypothesis.

Type I and Type II Errors

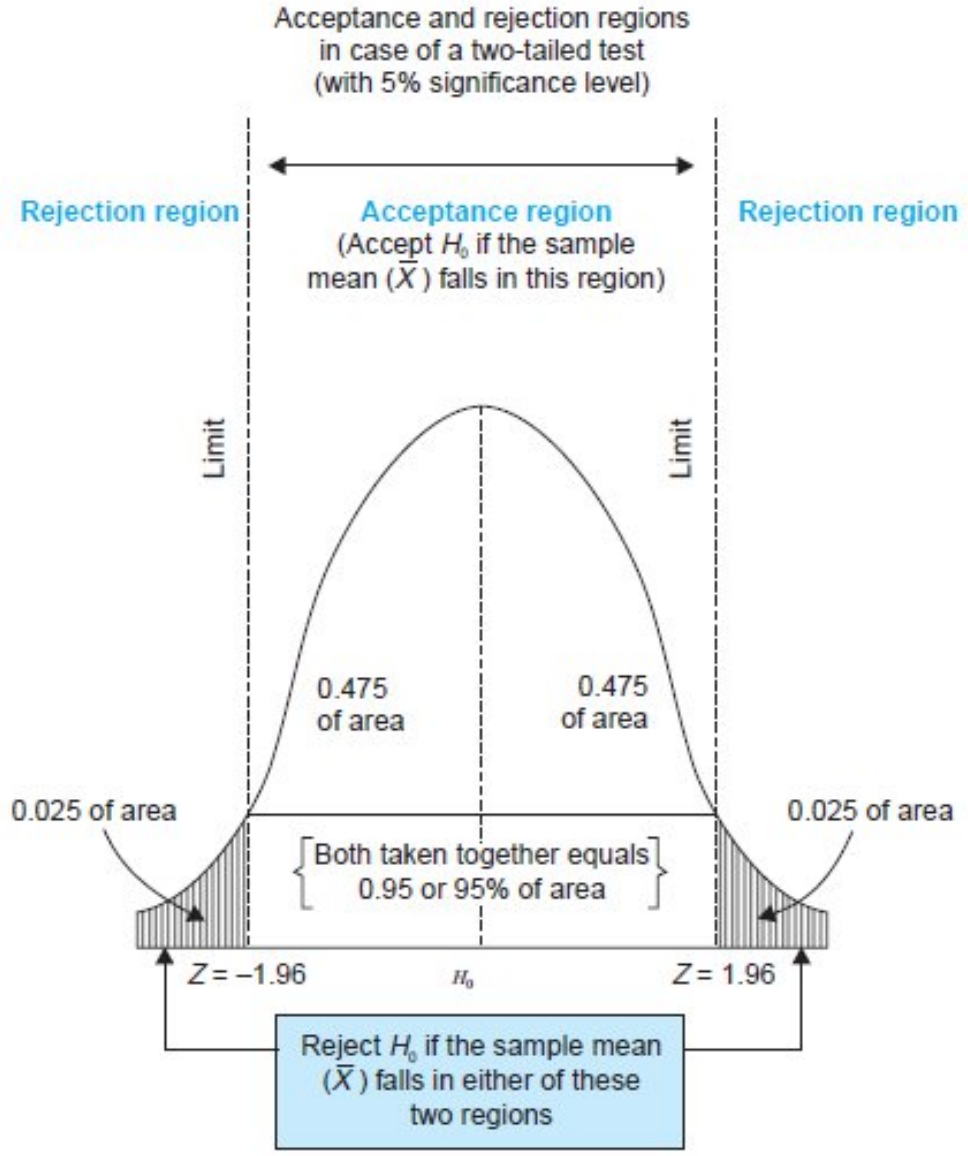
- Type I error: We may reject *H*0 when *H*0 is true
- Type II error: We may accept *H*0 when in fact *H*0 is not true

	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error (α error)
H_0 (false)	Type II error (β error)	Correct decision

Two-tailed and One-tailed tests

A **two-tailed test** rejects the null hypothesis if the sample mean is significantly higher or lower than the hypothesized value

$$H_0: \mu = \mu_{H_0} \text{ and } H_a: \mu \neq \mu_{H_0}$$



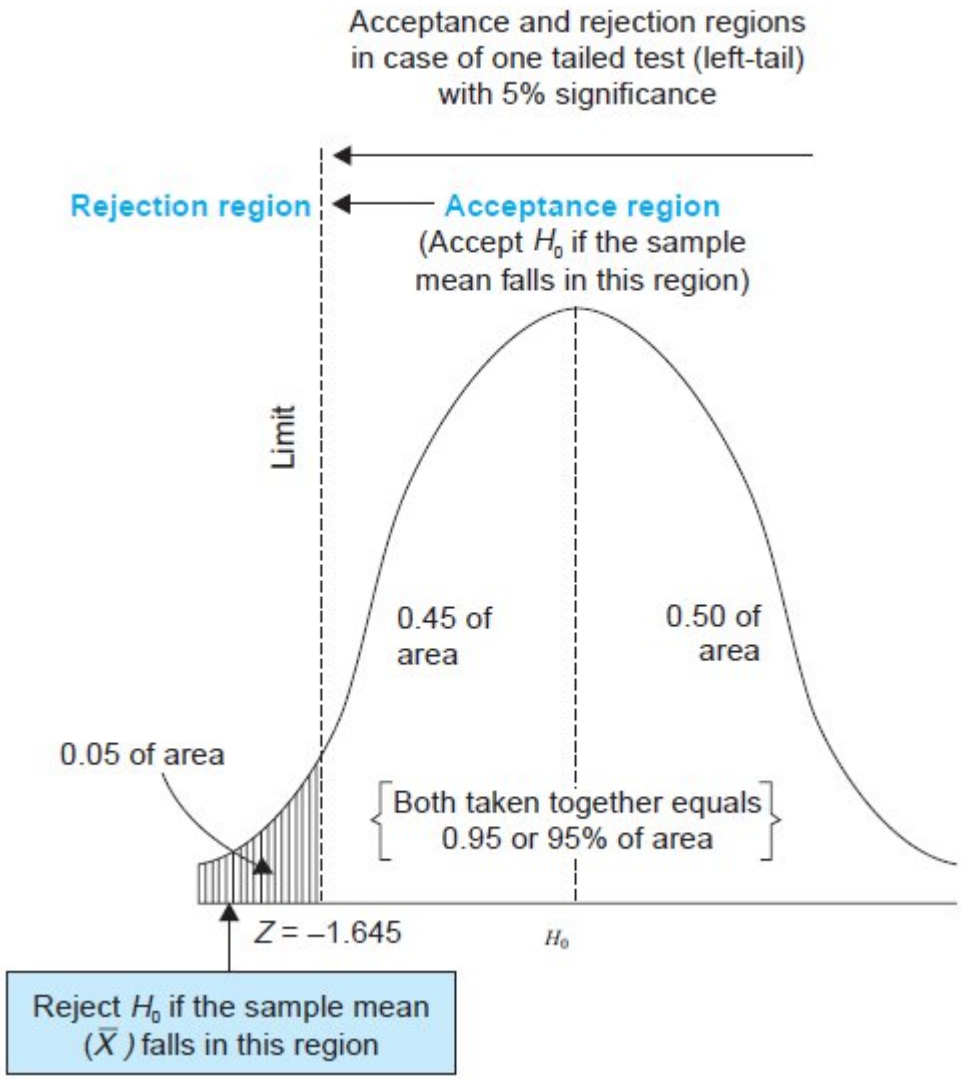
Two-tailed and One-tailed tests

A **one-tailed test** would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesised value.

It can be left tailed or right tailed test.

Example of left tailed test

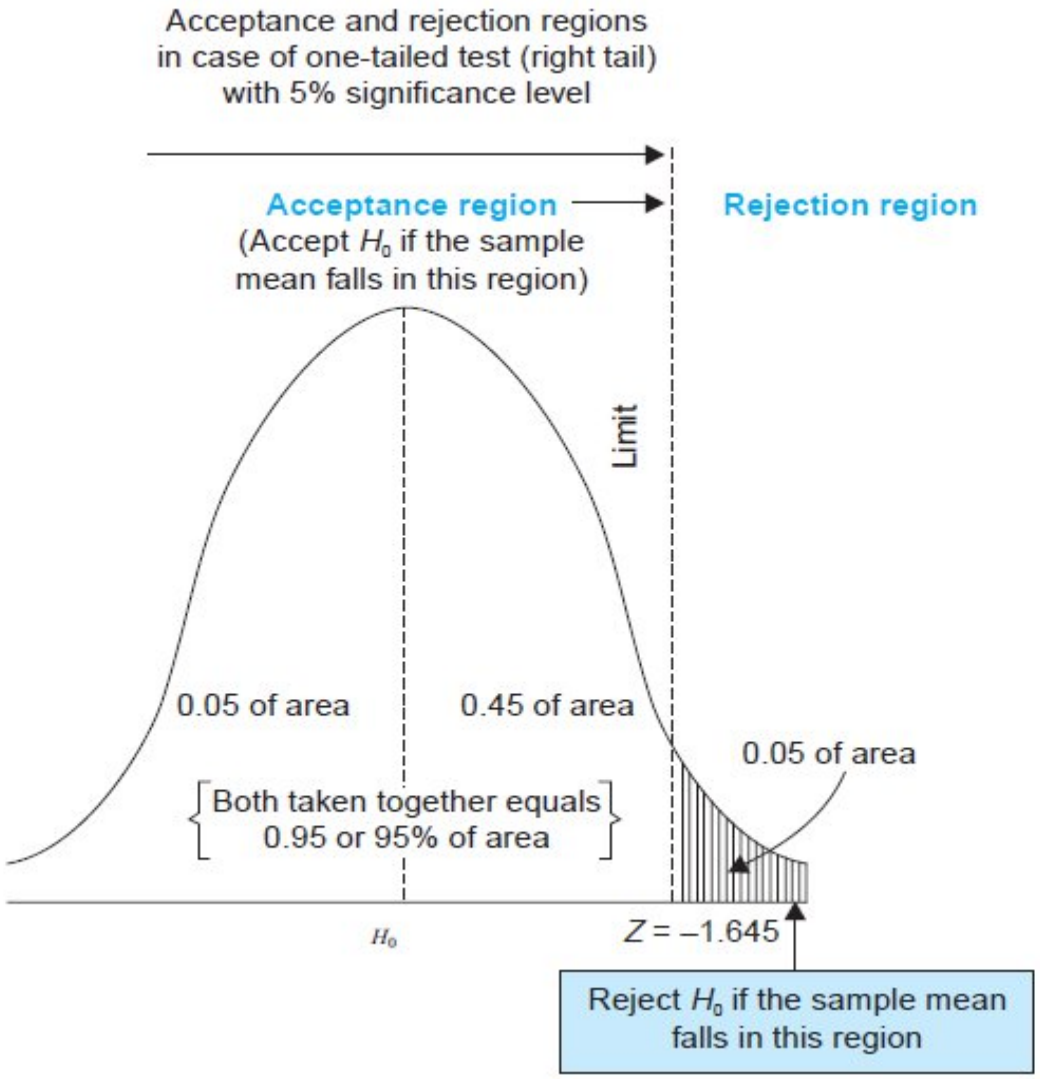
$$H_0 : \mu = \mu_{H_0} \text{ and } H_a : \mu < \mu_{H_0}$$



Two-tailed and One-tailed tests

Example of right tailed test

$$H_0: \mu = \mu_{H_0} \text{ and } H_a: \mu > \mu_{H_0}$$



Procedure For Hypothesis Testing

(i) *Making a formal statement:* The step consists in making a formal statement of the null hypothesis (*H*0) and also of the alternative hypothesis (*Ha*).

Null hypothesis $H_0 : \mu = 10$ tons

Alternative Hypothesis $H_a : \mu > 10$ tons

(ii) Selecting a significance level: The hypotheses are tested on a pre-determined level of significance and as such the same should be specified.

Generally, in practice, either 5% level or 1% level is adopted for the purpose.

Procedure For Hypothesis Testing

(iii) *Deciding the distribution to use:* to determine the appropriate sampling distribution.

The choice generally remains between normal distribution and the *t*-distribution.

(iv) *Selecting a random sample and computing an appropriate value:* to select a random sample(s) and compute an appropriate value from the sample data concerning the test statistic

(v) *Calculation of the probability:* to calculate the probability that the sample result would diverge as widely as it has from expectations, if the null hypothesis were in fact true.

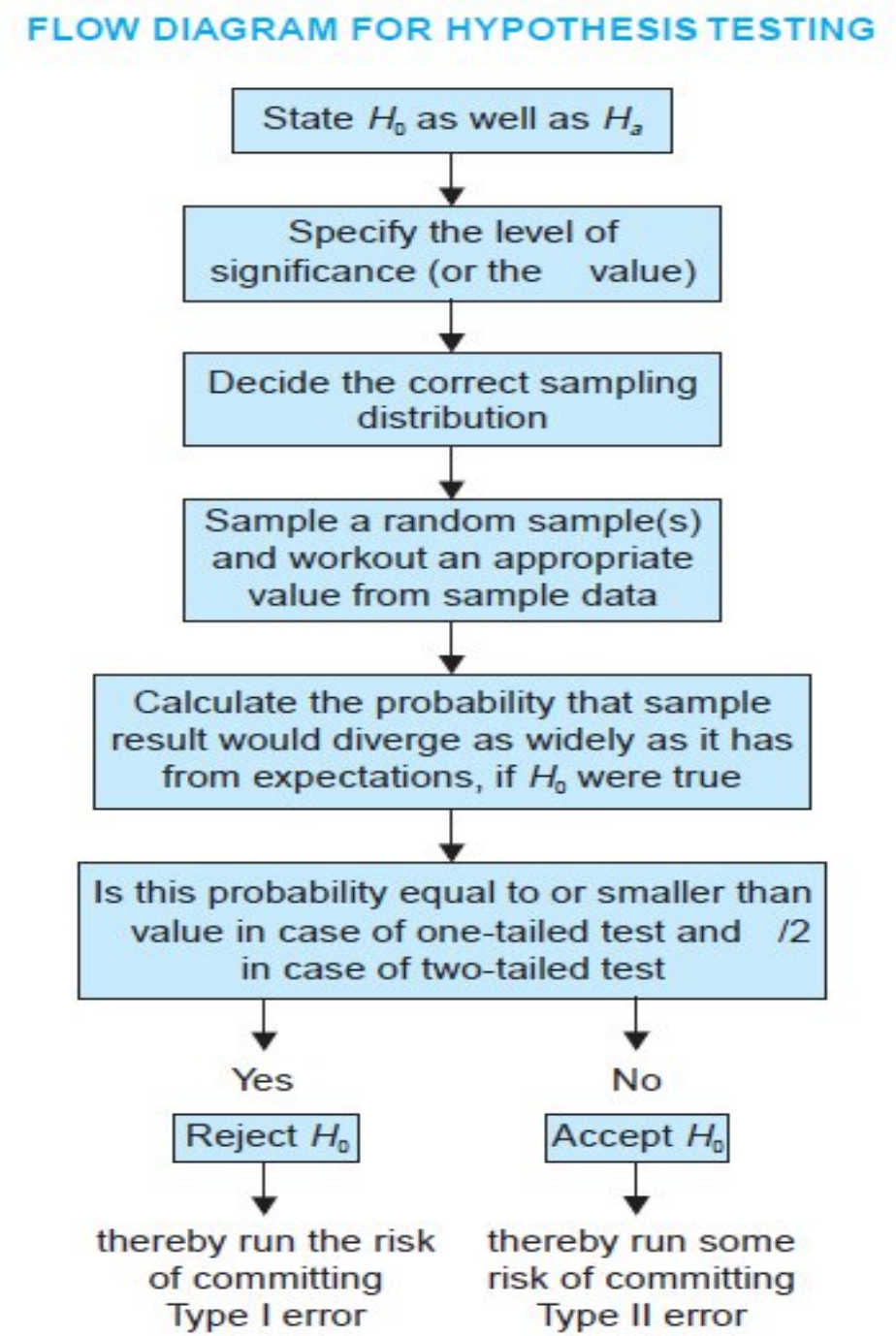
Procedure For Hypothesis Testing

(vi) *Comparing the probability:* comparing the probability thus calculated with the specified value for the significance level.

If the calculated probability is equal to or smaller than the α value in case of one-tailed test (and a /2 in case of two-tailed test), then reject the null hypothesis (i.e., accept the alternative hypothesis)

if the calculated probability is greater, then accept the null hypothesis.

Procedure For Hypothesis Testing



HYPOTHESIS TESTING OF MEANS

- Population normal, population infinite, sample size may be large or small but variance of the population is known, Ha may be one-sided or two-sided
- In such a situation *z*-test is used for testing hypothesis of mean

$$z = \frac{\overline{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

X-bar is the sample mean, μ is the population's assumed mean, σ_p is the population standard deviation, n is the sample size

HYPOTHESIS TESTING OF MEANS

- Population normal, population infinite, sample size small and variance of the population unknown, Ha may be one-sided or two-sided
- In such a situation *t* test is used

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}}$$

$$\sigma_s = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{(n-1)}}$$

σ_s is the sample standard deviation

HYPOTHESIS TESTING OF PROPORTIONS

Mean proportion of successes $= (n \cdot p)/n = p$

standard deviation of the proportion of successes $= \sqrt{\frac{p \cdot q}{n}}$

- If *n* is large, the binomial distribution tends to become normal distribution, and as such for proportion testing purposes we make use of the test statistic *z*

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

$\hat{p}$ is the sample proportion.

HYPOTHESIS TESTING OF VARIANCE

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

where σ_s^2 = variance of the sample

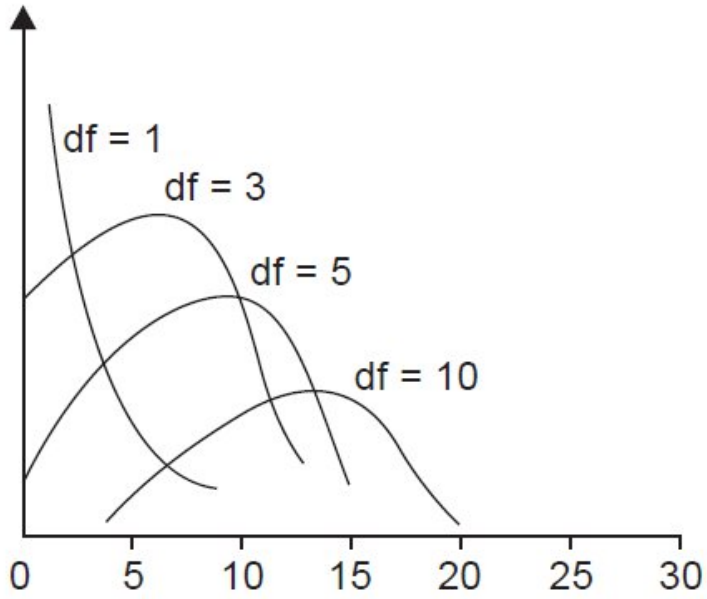
σ_p^2 = variance of the population

$(n-1)$ = degree of freedom, n being the number of items in the sample.

- The test we use for comparing a sample variance to some theoretical or hypothesized variance of population is different than *z*-test or the *t*-test.
- The test we use for this purpose is known as chi-square test

Chi-square test

- Chi-square test is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance.
- Applications - (i) test the goodness of fit; (ii) test the significance of association between two attributes, and (iii) test the homogeneity or the significance of population variance.
- The Chi-square distribution is not symmetrical and all the values are positive. One should know the degrees of freedom.
- The smaller the number of degrees of freedom, the more skewed is the distribution



Chi-square test

- The following conditions should be satisfied before chi-square test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the itmes in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known are know as linear constraints.

Steps in Chi-square test

(i) First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis. Usually in case of a 2×2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \left[\frac{\begin{array}{c}\text{(Row total for the row of that cell)} \times \\ \text{(Column total for the column of that cell)}\end{array}}{\text{(Grand total)}}\right]$$

(ii) Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(O_{ij} - E_{ij})^2$.

(iii) Divide the quantity $(O_{ij} - E_{ij})^2$ obtained as stated above by the corresponding expected frequency to get $(O_{ij} - E_{ij})^2/E_{ij}$ and this should be done for all the cell frequencies or the group frequencies.

(iv) Find the summation of $(O_{ij} - E_{ij})^2/E_{ij}$ values or what we call $\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. This is the required χ^2 value.

Chi-square test – Important characteristics

(i) This test (as a non-parametric test) is based on frequencies and not on the parameters like mean and standard deviation.

(ii) The test is used for testing the hypothesis and is not useful for estimation.

(iii) This test possesses the additive property as has already been explained.

(iv) This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.

(v) This test is an important non-parametric test as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

Analysis of Variance (ANOVA) and Co-variance (ANOCOVA)

- **ANOVA** - a procedure for testing the difference among different groups of data for homogeneity.
- "The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes."
- There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes.
- A method of analyzing the variance to which a response is subject into its various components corresponding to various sources of variation.
- Through this technique one can explain whether various varieties of seeds or fertilizers or soils differ significantly in the context of agriculture researches.
- The differences in various types of feed prepared for a particular class of animal or various types of drugs manufactured for curing a specific disease may be studied and judged to be significant or not through the application of ANOVA technique.
- A manager of a big concern can analyze the performance of various salesmen of his concern in order to know whether their performances differ significantly

Analysis of Variance (ANOVA) and Co-variance (ANOCOVA)

- **ANOVA** - to investigate any number of factors which are hypothesized or said to influence the dependent variable.
- To investigate the differences amongst various categories within each of these factors which may have a large number of possible values.
- The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.
- If we take only one factor and investigate the differences amongst its various categories having numerous possible values, we are said to use one-way ANOVA
- If we investigate two factors at the same time, then we use two-way ANOVA.
- In a two or more way ANOVA, the interaction (i.e., inter-relation between two independent variables/factors), if any, between two independent variables affecting a dependent variable can as well be studied for better decisions

Analysis of Variance (ANOVA) and Co-variance (ANOCOVA)

- Assumptions of ANOVA
  - Each group sample is drawn from a normally distributed population
  - All populations have a common variance
  - All samples are drawn independently of each other
  - Within each sample, the observations are sampled randomly and independently of each other
  - Factor effects are additive

Analysis of Variance (ANOVA) and Co-variance (ANOCOVA)

- **ANALYSIS OF CO-VARIANCE (ANOCOVA)**
- The object of experimental design is to ensure that the results observed may be attributed to the treatment variable and to no other causal circumstances.
- Consider an independent variable X and dependent variable Y.
- Z is an uncontrolled variable and is correlated with Y
- The researcher may wish to control the influence of *Z*, then he should use the technique of analysis of covariance for a valid evaluation of the outcome of the experiment.

Analysis of Variance (ANOVA) and Co-variance (ANOCOVA)

- **ANALYSIS OF CO-VARIANCE (ANOCOVA)**
- While applying the ANOCOVA technique, the influence of uncontrolled variable is usually removed by simple linear regression method
- The residual sums of squares are used to provide variance estimates which in turn are used to make tests of significance.
- In other words, covariance analysis consists in subtracting from each individual score (*Yi*) that portion of it *Yi´* that is predictable from uncontrolled variable (*Zi*) and then computing the usual analysis of variance on the resulting (*Y* – *Y´*)'s,

Thank you

. Note: Unit 3, part 2 consists of problems. Notes of the same in the pdf format will be provided