Chapter – VII

Impact Evaluation Tools and Basic Statistics

7.1 Evaluation for Impact Assessment

KVK programmes are investments where we expend capital resources to create sustainable functional units at gross root level from which we can expect to realize the benefits over an extended period of time. According to USDA, Evaluation is the process of determining how well one is doing in what one is trying to do. Evaluation when applied to the field of extension may be defined as a process of systematic appraisal by which we determine the value, worth or consequences of the extension programme/activity. Most KVK programmes are similar of extension programmes which are at the interface between the demand and supply systems of agricultural technology. Most of the evaluation study conducted in extension was of mostly comparison of production yield before and after the implementation of the programme. However, it must be understood that evaluation is not simply a measurement of achievements, which is usually done after a programme is executed. A complete evaluation for KVK programmes is one which aims at the full length enumeration of both tangible and intangible costs and benefits involved. Both tangible costs and benefits are easy to identify but it is not so for intangible ones. Knowledge, once disseminated by the extension service and acquired by farmers, has a tangible measurable product only if applied (Orievel, 1983; Feder and Slade, 1985).

The application of such knowledge by farmers is generally termed as adoption and is usually measured by adoption rates, that is, the proportion of farmers applying knowledge of a particular technology that they have acquired from extension agents. Moreover, adoption cannot take place without knowledge. It is not an automatic consequence of the acquisition of knowledge as many other influences may affect a farmer's decision to adopt. Among these are the profitability of the technology, the availability of key inputs, credit and complementary knowledge (Fitzhugh et. al.1982). Moreover, many of the farm production technologies are often not available to the resource-poor farmers living in remote areas, and clients of these services are semi-commercial farmers who benefit from such technology (Perviaz and Hendrik, 1989). So it is not entirely correct and complete method of evaluating the extension through the adoption rate alone. An application of the economic concept in the evaluation of extension projects brings about the total net cost and benefit of the project

7.2 Various Stages/ Phases in Economic Evaluation of Extension

Where does economic evaluation fit into extension programmes? Economic evaluation is a part and parcel of all phases in an extension programme right from its initial planning to implementation and completion. The economic concept of extension is applied:

(1) at project selection,

(2) during implementation, and

(3) after completion of the project

7.3 Evaluation Criteria during Project Selection

An extension project can be either selected or rejected once its cost and benefits are identified and valued. Moreover, the realistic estimation of costs and benefits is a pre-requisite for the successful evaluation of an extension project at its selection stage. Most of the extension projects suffer from incomplete identification usually resulting into over-estimation of benefits and under-estimation of costs (Evenson, 1978). The compounding and discounting are the techniques to enable the benefits the costs streams (or cash inflow and cash outflow) from different projects at a particular point of time to examine the profitability of each project and relative profitability of the projects.

Usually, a wide range of criteria are applied to evaluate a project during the selection stage for choosing the investment proposals. These criteria are classified into (1) Non-discounting criteria which includes urgency, payback period and accounting rate of return for estimation, and (2) Discounting criteria which includes the Net present worth , Internal rate of return, Benefit-cost ratio and Net Benefit Investment Ratio as its tool. The figures obtained during the initial stage of project selection are purely on the basis of conceptual ideas. This has to be confirmed before actually implementing the project on a large scale. This could be done through conducting on field extension trials and evaluating the results.

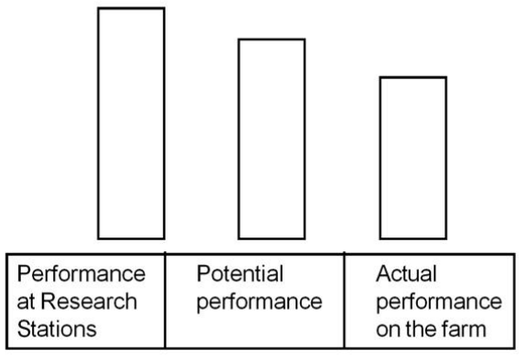
7.4 Economic Analysis of On Farm Trials (OFTs):

The consequences expected out of selected extension projects and technologies that are to be transferred to the farmers can be assessed by carrying out the evaluation of On Farm Trials (OFTs) under different client group settings. The objective here is to ascertain the initial estimated cost and benefits of the project on its large scale implementation. Thus OFT will help us as a screening tool in eliminating the extension projects and technologies which are not viable before going for a large scale implementation. The initial cost-benefit estimates worked out at the time of project selection is purely based on the projected and conceptual figures whereas results of the OFT gives us the estimates nearer to the real costs and benefits of the project on its large scale implementation.

Although the need to develop bio-economic models to screen new technologies has been stressed, data are so limited that such complex models can be developed only in advanced countries. They become mere academic exercises when applied to developing country situation. Even though the OFT can guide research, evaluation as well as to identify the weaknesses in the extension system, such models were rarely conceived and tried in the practical situation because the models suffer with various limitations. Demonstration farm models have widely been used to conduct the economic analysis of OFT.

The demonstration farm approach has helped to determine the economic value of new dairy technology for different sizes of farms (Patel, 1981). The yield-gap model gives the difference between the yields that are technically and economically feasible and the actual yields obtained by majority of the farmers. This model presented by Gomez and Santos (1981) is mainly used to explain the differences in the performance of technology at the farmers' farm and research station (Refer figure). What is the effect of extension in reducing this gap? Only an effective evaluation will give the answer.

Production differences



Environmental Differences

i. Biological Differences : Species, Diseases, Parasites, Nutrition.

ii. Managerial and Socio-Economic Constraints: Knowledge, preference, practices, credit, market costs and input availability.

Fig.: Yield Gap Model

The profitability of an extension programme should be verified through giving special attention to the identification of net costs and benefits of the technology to be transferred. In the preliminary stage, profitability should be estimated at the optimal input levels, with appropriate discount for risk. In addition, farmers should be consulted to identify the inputs, such as cash, labour, training requirements that play a part on the decision to experiment with the new technology which is to be implemented through the existing extension system. An evaluation that takes into account only the cash inputs (such as gross margin analysis) may be appropriate for certain type of technologies only (Pervaiz and Hendrik, 1989). Table 1 shows how economic tools can be matched for screening the technologies involved in farm/animal production before incorporating into the extension system.

Table1: Economic tools for screening the technologies

Technology	Farmers criteria	Tools of analysis
Effects of new yoke design on draft power	Weight and durability	Benefit – cost analysis
Effect of feeding practices on milk production	Changes in milk yield	1) Marginal analysis 2) Production function 3) Linear programming
Health related	Disease prevention and control	Benefit – cost analysis
Breeding and Management	Genetic potential production	Gross-margin analysis
Effects of feed supplementation	Physical performance and weight gain	1) Partial budgeting 2) Financial analysis

Source : Pervaiz and Hendrik (1989).

7.5 Cost-Benefit Analysis in Extension

"Every possible thing in this world is at some cost". The net cost and benefit in any extension programme includes both tangible and intangible costs. According to this, we cannot ignore the social cost (which is intangible) which the society incurs on the execution of the programme. Even though it is intangible, it is considered for the cost estimation because of the social value attached to it. The difficulty involved in measuring the social cost benefit is anybody's imagination. Yet without the inclusion of these intangible costs the efforts to evaluate the extension largely remains non-functional.

The social cost is anything the society pays/sacrifices save monetary contributions for the execution of an extension programme. Likewise the social benefits are those benefits other than its monetary benefits derived from the outcome of the project on its execution. Mostly, the social cost and benefits are indirect in nature.

Vaccination campaigns and knowledge about the animal diseases directly reduced the number of animals affected by the diseases and indirectly improved the relationship between the village rival groups as no longer the farmers believe the ailment of his animals as the handiwork of his village witchcraft and rival group. The extension programme brings knowledge to the farmers and their family. In broader sense, the cost of bringing out the knowledge to the family can be worked out but what the knowledge brings/does to the farm family is multi-dimensional in nature in which the economic benefit is one of its dimensions. The other dimensions are, due to increase in production, the farmer becomes economically well-off and his sons and daughters could get educated and take up new occupations. What about the services rendered by these benefited classes to their society? Is it an outcome of extension or not? How we can cover all these dimensions while attempting to measure the social cost and benefit of the extension? Is it possible to cover all its dimensions or not? If yes, up to what extent, if not what are those impeding forces and alternatives perceived? These are the questions, which need to be answered. Still there are no definite criteria or standardized procedures or ways to measure the social cost and benefit in the extension. As already stated, the economic cost-benefit is one aspect of the whole which we have to measure for determining a project's worth. It is the responsibility of the social scientists to take up this issue for attempting a complete evaluation of extension in this direction.

The final step in the evaluation required that the value of the increase in farm output resulting from our estimates of the productivity differential attributable to extension, be set against the additional costs incurred to make the additional output possible. To do so, the familiar technique of cost benefit analysis has to be applied to the extension system. As the analysis has to be undertaken ex-post availability of a complete series of either costs or benefits for the entire life of the project will help in arriving at an accurate evaluation result. If the costs and benefits for the entire life of the project is not available, we have to make number of assumptions relevant of an ex-ante analysis (Feder and Slade, 1985).

The stream of incremental extension cost should be constructed using data on the actual costs of the project during the initial years. In scenarios where the project life was assured to be less than the life span of physical structures and equipments, appropriate residual values were calculated and deducted from the costs (Feder and Slade, 1985), For arriving at an actual impact of extension,

the econometric techniques which accounted for differences in the quantities of variables and fixed inputs, the type of soils, human capital, and the production environment are used to estimate the percentage output differentials between the two areas with differences. If variable inputs are not taken into account then the output differentials includes the effect of extension on farm efficiency as well as the effect on the use of inputs, provided that price differential are also controlled (Feder and Slade, 1985). A complete accounting for the effects of extension should, therefore, take both types of effect into account.

7.6 Basic Statistical Tools

A branch of knowledge can be called as science only on condition that it can be studied through scientific methods. Science goes with the method and not with the subject matters. Scientific method consists of systematic observation, classification and interpretation of data. The main difference between our day to day generalization and the conclusion usually recognized as scientific method lies in the degree of formality, rigorousness, verifiability and general validity of the latter (Lunmdberg, 1960). Science may be defined in terms of six major processes that take place within it. These are testing, verification, definition, classification, organisation, orientation, which includes prediction and application. The scientific method is marked by the following features;

a) Careful and accurate observation, collection and classification of facts.

b) Observation of factual correlation and sequence.

c) The validity and reliability features.

d) Measurements and statistical analysis.

In this part the above features involved in the scientific method is briefly taken up for discussion.

7.7 Methods of Observation and Collection of Facts

7.7.1 'Ex post facto' method of observation

The term 'Ex post facto' was originally used by Chapin and Greenwood to mean a quasi experiment in which an attempt was made to control independent variables by matching and symbolic methods. The term 'Ex post facto' method refers to such method of observation or empirical inquiry in which the scientist doesn't have direct control of independent variables because their manifestations have already occurred or because they are inherently not manipulable. Inferences about the relationships among the variables are made from concomitant variation of independent and dependent variables without any direct intervention. In this method no direct control is possible and neither experimental manipulation nor random assignment can be used by the researcher. In the ex-post facto method of observation he does not have any manipulative control of independent variable.

7.8 Collection of Data or Facts

7.8.1 Survey

In social sciences the facts about a society or a social phenomenon is collected through survey technique or method. The survey is in brief a method of process by which quantitative facts

are collected from a given population. If the survey method is applied to collect a given quantitative data from all the units in a selected population then it is called as complete enumerative survey method. When the population size is big, it is not possible to collect data or observe the fact from every unit of the population. In such case sample survey technique is used.

7.8.2 Sampling Technique

Sample is a small group or unit which is taken as the representative of the whole. The whole group from which the sample has been drawn is referred as the population and the small group or unit drawn from the population to represent the population for the study is called as the sample. Sampling method has the advantages of saving time and money, in addition to the possibility of a detailed study. The accuracy of the results, high validity and reliability of the method make the sampling technique highly useful.

Methods used for drawing the sample from a given population are given under:

1) Random sampling
2) Purposive sampling
3) Stratified sampling
4) Quota sampling
5) Multistage sampling and
6) Convenience sampling

7.8.2.1 Random Sampling

Is the method of selection which assures each individual or element or unit in a given population an equal chance of being chosen. Lottery method, Tippets number system and selection from sequential list are the commonly used random sampling techniques.

7.8.2.2 Purposive Sampling

When the researcher purposively selects certain units or elements from the population, it is known as purposive sampling technique. In purposive selection the number of groups or units is selected in such a way that the selected units together yield the same average as the totality with respect to those characteristics which are under observation.

7.8.2.3 Stratified Sampling

It is a combination of both random sampling and purposive selection. The population is divided into a number of strata or groups. Then from each group or strata the required number of samples is drawn randomly.

7.8.2.4 Quota Sampling

In this method initially the population is divided into different strata. After the stratification, the number of sample to be selected from each stratum will be decided. This number is known as quota. Then the fixed quota of samples is drawn from each stratum.

7.8.2.5 Multistage Sampling

This method is suitable when sample is to be selected from a large population. The selection of the sample is made in different stages. For example, at the first stage the districts in a state are grouped according to region wise and from each of the region one district is selected. Then from each of the district, one taluk representing each area of the district is selected. The villages in the selected taluks are divided into developed and developing and from each of the category two villages are selected. Finally, from the selected villages, the house holds are grouped into various categories and from each category a fixed number of sample households are picked up randomly. Here, the selections of final sample units are made in number of stages. This method is a combination of stratified sampling and random sampling technique. By this method in a large population the representation from every unit can be obtained.

7.8.2.3 Convenience Sampling

Sample is selected according to the convenience of the researcher. The convenience may be anything such as availability, accessibility of the units etc. This method is generally unsystematic, or opportunistic.

The reliability of the sample selected can be tested by means of drawing a parallel sample or by comparing the measurement of the sample with those of the known population scores or by drawing a sub sample from the main sample.

7.8.3 Schedule and Questionnaire

The schedule and questionnaire are the forms containing some statements expressed in wide varied formats. The basic difference between the schedule and questionnaire is, in terms of schedule the researcher will personally administer it to the subject for collection of data or make observations. The questionnaire is the one when such formatted statements are mailed to the subjects for getting the response. Generally, the schedule and questionnaire are classified into structured or non-structured one. The structured format contains a definite, concrete and pre-ordinated question. The non structured format which is often known as interview guide is very much general and provides guidelines and instructions only.

The purpose of a schedule or questionnaire is to obtain the required information from a subject. Different types of schedules are observation schedule, rating schedule, document schedule and interview schedule. A good schedule should aid the required communication between the subject and researcher. It should also get the right response from the subject. The schedule should be simple and convey accurate meaning without any ambiguity. The size of the schedule should be optimum. Too small schedule will cut short the information and lengthy schedule will be difficult to use and at the end huge volume of data will complex the coding, analysis and interpretation process.

7.9 Observation Methods

Science begins with observation and must ultimately return to observation for its final validation. The observation is classified into participant observation and non-participant observation, controlled and uncontrolled observations. In the participant observation, the observer participates

with the activities of the subject under study and collects the required facts. In non participant observation the observer does not actually participate in the group activities and keep himself detached from the subject under observation and collect the required facts. In the former method the observer identifies himself with the group and its activities and in the later case he maintains distance and remains as an outsider.

7.9.1 Controlled and uncontrolled Observations

The observation may be controlled or uncontrolled. When the observations are made under natural surroundings and the activities are performed in their normal course without being influenced or manipulated by external force, it is known as uncontrolled observations. In case of controlled observation the experimenter exercises his control over the phenomenon and observation. The main purpose of controlled observation is to check any bias due to faulty perception and to avoid the influence of external factors which are not meant or connected with the experiment.

7.10 Measurement and Statistical Analysis

Measurement is the basic act of all scientific research. The outcome of all experiment is observed and measured. On qualification the data is subjected to statistical analysis to draw meaningful inferences from the figures or numbers obtained as a result of basic observation and measurement. Statistical analysis gives the meaning to observation. It establishes the reliability and validity of the results obtained. In this part, the fundamentals of measurement, reliability and validity of the data and basic statistical methods are discussed in brief.

7.10.1 Measurement

Assignment of numbers/numerals to objects or events according to rules is called as measurement. A numeral is a symbol of the form 1, 2, 3 or I, II, III. It has no quantitative meaning unless we give it such a meaning; it is simply a symbol of a special kind. Numerals are used because measurement ordinarily uses numerals which after being assigned quantitative meaning become numbers. A number, then, is a numeral that has been assigned quantitative meaning. General equation for any measurement procedure is f = {[x,y]; x =any object, and y=numeral}. This is read as “The function, or the rule of correspondence, is equal to the set of ordered pairs (x,y) such that x, is an object and each corresponding y is a numeral (Kerlinger, 1964).

7.10.2 Postulates of Measurement

Nine postulates of measurement given by Campbell are briefly given under. Out of the nine postulates of measurement, the first three postulates measures the identity. The fourth and fifth postulates measure the order and the remaining last four postulates deals with the activity. The postulates are:

1) a=b or a # b; it indicates the identify of the number.

2) If a = b, then b = a; here the relationship is symmetrical.

3) If a = b, and b = c then a = c; things equal to same thing are equal to each other.

4) a > b or b > a; indicates the order.

5) If a > b and b > c; then a >c; helps to put the variables in a rank order continuum.

6) If a = p and b >0 (Zero) then a+b > p; addition of zero leaves a number invariable.

7) a+b = b+a: Then order in which things are added makes no difference in the result.

8) If a = p and b = q then a+b = p+q (i.e) then identical objects may be substituted for one another in addition.

9) If (a+b)+c = a+(b+c) i.e. the order of combinations or associations make no difference in addition.

7.10.3 The Levels of Measurement

There are four general levels of measurement; nominal, ordinal, interval and ratio. These four levels lead to four kinds of scale. They are discussed hereunder:

7.10.3.1 Nominal scale

Simple and pre-requisite to all scales. Its major part is equivalence. In this scale we take into consideration the categories on the basis of the numbers and objectives.

7.10.3.2 Ordinal Scale

Ordinal type of scales take into consideration the relationship “greater than” or “higher than”. It is the relationship of asymmetrical which is possible for us to put our objectives in a psychological continuum. But in this scale, we are not interested to find out how much greater or how much higher (i.e) the distance between the two levels may not be equal but it is possible only to put in an order. In the interval scale we consider the distance.

7.10.3.3 Interval Scale

Here, we have the characteristics of nominal and interval scales. In addition, it has the consideration of distance between the two points i.e. the distance between a & b and b& c. In addition we have an arbitrary zero and we do not know from where this zero starts.

7.10.3.4 Ratio Scale

It includes all the qualities of ordinal, nominal and interval scale. In addition to this we have one absolute zero. But in social scale construction, it is not possible to go for such higher scale for measurement. In the nominal and ordinal type of scales mostly non-parametric tests are used. In the interval and ratio scale both parametric and non parametric type tests are used.

7.11 Reliability and Validity

7.11.1 Reliability

It means the accuracy or the precision of the measurement. It indicates the dependability or consistency or predictability of the measuring instrument or the method of approach followed in a research. If we measure the same concept twice after some interval and get more or less same result both the time, then we can say that our measurement has stability, dependability and predictability. Before accepting a method or measurement we have to estimate the reliability. The reliability can be estimated by the following methods.

1) Test Retest method

2) Parallel or Equivalent form method

3) Split Half method

7.11.1.1 Test Retest Method

In the test-retest method the same scale of measurement is applied two times. The results are then correlated by applying the Pearson-Product Movement test. This co-efficient or correlation is called as coefficient of stability. While using the test-retest method the main constraint is time. If the distance or time between the two tests is longer then the score may be altered due to the changes that have taken place in the subject quality in the mean time due to external factors. So the resultant "r" value is unreliable. If the time interval is too short then the "r" value is high and so we get a very high reliability of the scale. So, we have to avoid too short and too long time interval between two tests.

7.11.1.2 Parallel Form or Equivalent Form Method

Here the whole measuring scale is divided into two groups taking into consideration that all the items in one group is equivalent to the second group. On the basis of this they are called equivalent. The two groups of test should be administered under similar condition. Then the 'r' value is calculated between the two groups. The 'r' is called as correlation of equivalent.

7.11.1.3 Split Half Method

This method is extension of the equivalent form method. Here instead of dividing the whole scale into two halves, we divide the scale on the basis of odd and even numbers basis. If an attitude measuring scale have 20 statements, the statement number 1,3,5,7,9……19 forms one group and even number statements 2,4,8……20 forms the other group. The statements are administered to the subject and corresponding score for each group is calculated. Then correlation co-efficient between the two groups using the Spearman Brown formula is obtained.

7.11.1.4 Standard Error of Measurement

If we administer the same scale under the same condition repeatedly, every time we get the score more of less same but may not be actually same. Then we have to calculate the true score by applying the correction factor. Calculate the mean of the scores you obtained in 4 to 5 times of

measurement and then derive the standard deviation. Then the standard error of the measurement is calculated by the following means.

$$Se = s\sqrt{1-rtt}$$

Where Se = Standard error of measurement

S = Standard deviation and

rtt = Reliability co-efficient of the test

7.11.2 Validity

It means, are we measuring the same thing which we are supposed to measure? The test of measurement is valid only if it measures the same item which it intends to measure. The one way of establishing the validity of the measuring tool is to analyse for its content validity which means the representativeness or sampling adequacy of the content of the measuring instrument to measure the particular logic or concept. Nunally (1967) indicated two major standards for ensuring content validity: (i) a representative collection of items, and (ii) sensible method of test construction. The measuring instrument need to be constructed in accordance with the steps enunciated or recommended for the particular measuring technique.

7.12 Measures of Central Tendency

7.12.1 Average

An average is a quantitative figure that expresses the quantitative nature of a population. An average is a figure or expression that conveys the qualities or characteristics of any group. Average can convey the common qualities of the group, or in other words those characteristics that are common to all or atleast most of the units. The different types of averages used can be conveniently categorized into the following ways.

1) Average of location
   a) Mode
   b) Median
2) Mathematical Average
   a) Arithmetic Average
   b) Geometric Average
   c) Harmonic Average

7.12.2 Mode

Mode is the value of the item that has highest frequency. It is the measurement that is directly applicable to a large number of cases. The calculation of mode in a given data depends on the frequency of the items. The data is to be grouped in discrete or continuous series and the item value with higher frequency would be the mode.

7.12.3 Median

Median is the measurement of the middle item arranged in ascending or descending order. The underlying assumption of median is that measurement increase by regular order and thus total measurement above the middle item is the same as the total measurement below it. Naturally the size of the middle item is the representative of average size.

7.13 Chi Square Test or Goodness of Fit

The chi-square test and its application to the theory of statistical inference were first introduced by Karl Pearson. Chi-square is a measure of the degree to which a series of observed frequencies deviates from corresponding theoretical or hypothetical frequencies.

Determination of reliability on the basis of chi-square test is also known as 'goodness of fit'. This is because we try to fit the theoretical frequency distribution upon the observed frequencies. If this fit is quite close we can say that the fit is good or in other words there is no significant difference between the observed and theoretical frequencies. If on the other hand the disparity is very great, the fit is not considered good and we shall consider the difference to be significant.

7.14 Statistical concepts for KVK Scientists

Statistics is the science of collecting, organizing, analyzing, and interpreting data.

There are 2 types of data sets in statistics. They are population and sample.

Population is a data set consisting of *all* outcomes, measurements, or responses of interest while

Sample is a data set which is a subset of the population data set.

Examples:

- If we are interested in measuring the salaries of KVK SMS in India, the population data set would be a list of the salaries of every SMS in India. A sample data set could be obtained by selecting 100 SMS from across the country and listing their salaries.
- A survey organization wants to know whether Indians favour increased defence spending. The population data set would consist of the responses of every citizen. A common way of choosing a sample data set would be to randomly call 1000 citizens and gather their responses to the question of whether they favour increased defense spending.

7.15 Types of Measurements

Parameter - a numerical measurement made using the population data set

Statistic - a numerical measurement made using a sample data set

Examples:

- Using the SMS salary data sets, we could calculate the average salary for the SMS. The average calculated from the population data set would be the parameter. The average calculated from the sample of 100 SMS would be a statistic.
- Using the opinion poll data on defence spending, we could calculate the percentage of Indians who favor increased defence spending. The actual percentage of all citizens who favored increased defence spending would be the parameter. The percentage of the 1000 citizens in our sample who favored increased spending would be a statistic.

Notice that unless the population is very small it is probably impossible to gather the population data set, and so it is usually impossible to calculate the parameter we are interested in.

The main idea of the science of statistics is that we can get around this difficulty by selecting a sample, calculating the sample statistic, and use the sample statistic to make an estimate of the parameter.

Unfortunately, statistical estimates can never be 100% certain. (But they can be 90% or 95% or 99% certain)

7.16 Types of Data

7.16.1 Qualitative Data - non-numerical characteristics or labels

Examples: Eye Color, First Name, Favorite Movie, Political Party

7.16.2 Quantitative Data - numerical measurements or quantities

Examples: Height, Weight, Income

7.17 Levels of Measurement

7.17.1 Nominal Data

Can be qualitative only. Data values serve as labels, but the labels have no meaningful order.

Examples: Blood type, Graduation subject, Breed of a dog.

7.17.2 Ordinal Data

Ordinal data can be qualitative or quantitative. Data values serve as labels but the labels have a natural meaningful order. Differences between values, however, are meaningless.

Examples: Semester Grade, ICC Cricket Rankings.

7.17.3 Interval Data

Interval data are always quantitative. Data values are numerical. So they have a natural meaningful order, and differences between data values are meaningful. The ratio of two data values, however, is meaningless. This occurs when zero is an arbitrary measurement rather than actually indicating "nothing".

Examples: Temperature, Year of Birth

7.17.4 Ratio Data

Ratio data are always quantitative. Data values are numerical, have order, and both differences and ratios between values are meaningful. Zero measurement indicates absence of the quantity being measured.

Examples: Weight, Height, Volume, Number of Children

7.18 Frequency Distributions

A **frequency distribution** is a table used to describe a data set. A frequency table lists intervals or ranges of data values called **data classes** together with the number of data values from the set that are in each class. This number is called the **frequency** of the class.

Example: Statistics exam grades. Suppose that 20 statistics students' scores in an exam are as follows:

97, 92, 88, 75, 83, 67, 89, 55, 72, 78, 81, 91, 57, 63, 67, 74, 87, 84, 98, 46

We can construct a frequency table with classes 90-99, 80-89, 70-79 etc. by counting the number of grades in each grade range.

Class	*Frequency (f)*
90-99	4
80-89	6
70-79	4
60-69	3
50-59	2
40-49	1

Note that the sum of the frequency column is equal to 20, the number of test scores.

Additional Terminology

Lower Class Limit – The least value that can belong to a class.

Upper Class Limit – The greatest value that can belong to a class.

Class Width – The difference between the upper (or lower) class limits of consecutive classes. All classes should have the same class width.

Class Midpoint – The middle value of each data class. To find the class midpoint, average the upper and lower class limits.

7.19 Mathematical Notation

The following symbols and variables normally carry meanings as given below. (unless otherwise specified)

Variables

x = data value

n = number of values in a sample data set

N = number of values in a population data set

f = frequency of a data class

Symbol

$\sum$ indicates the sum of all values for the following variable or expression.

Example: Using our notation, we can write the statement that the sum of the frequencies in a frequency table should equal the number of values in the data set as follows:

$$\sum f = n$$

7.20 Cumulative Frequency

The **cumulative frequency** of a data class is the number of data elements in that class and all previous classes. (can be done ascending or descending)

Example:

Class	*Frequency (f)*	*Cumulative Frequency*
90-99	4	4
80-89	6	10
70-79	4	14
60-69	3	17
50-59	2	19
40-49	1	20

Notice that the last entry in the cumulative frequency column is $n = 20$.

7.21 Relative Frequency

The **relative frequency** of a data class is the percentage of data elements in that class. We can calculate the relative frequency for each class as follows:

$$\text{relative frequency} = \frac{f}{n}$$

Example:

Class	Frequency (*f*) Frequency	Cumulative Frequency (*f* / *n*)	Relative
90-99	4	4	.20
80-89	6	10	.30
70-79	4	14	.20
60-69	3	17	.15
50-59	2	19	.10
40-49	1	20	.05

Note: The sum of the relative frequencies should be 1.

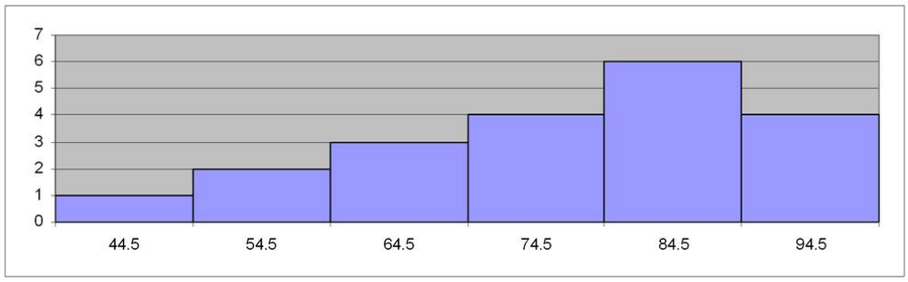
$$\sum \frac{f}{n} = 1$$

7.22 Histograms

A **histogram** is a graphical representation of the information in a frequency table using a bar graph.

The histogram should have the variable being measured in the data set as its horizontal axis, and the class frequency as the vertical axis. Each data class will be represented by a vertical bar whose height is the frequency of the class and whose width is the class width.

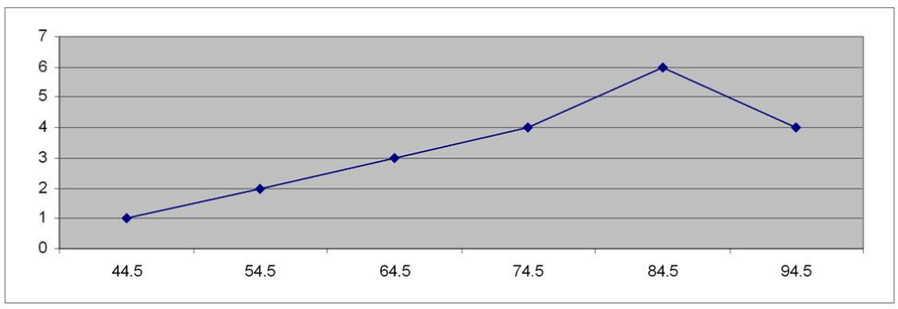
Notice that the bar for each class is centered at the class midpoint, and the bars for successive classes touch.



7.23 Frequency Polygon

A **frequency polygon** is a line graph representation of the information in a frequency table.

Like a histogram, the vertical axis represents frequency and the horizontal axis represents the variable being measured in the data set. To construct the graph, a point is plotted for each class at its midpoint and with height given by the frequency of the class, the points are then connected by straight lines.



7.24 Measures of Central Tendency

A **measure of central tendency** is a value used to represent the typical or "average" value in a data set.

Three Common Measures of Central Tendency

Mean – the sum of all data values divided by the number of values in the data set. The mean of a sample data set is denoted by $\overline{x}$ and the mean of a population data set by the Greek letter μ.

$$\overline{x} = \frac{\sum x}{n} \qquad \mu = \frac{\sum x}{N}$$

- **Median** – the value which separates the largest 50% of data values from the lowest 50%. To calculate the median, place data values in number order. If *n* is odd, the middle value is the median. If *n* is even, the mean of the two middle values is the median.
- **Mode** – the data value (or values) which appears the largest number of times in the set. If no data value is repeated, we say that there is no mode.

Properties of Mean, Median and Mode

- Mean is the most commonly used measure of central tendency.
- One drawback of the mean is that it is heavily influenced by a few very high or very low data values. In these cases it is more common to use the median.
- The mode has the advantage that it can be used to measure data sets even if they contain only qualitative data. A disadvantage is that a data set may not have a mode.

Weighted Means

A **weighted mean** is used when we want some data values in a set to factor more often into the calculation of the mean than others.

In this case, we attach a numerical **weight** to each value and calculate the mean as follows:

$$\overline{x} = \frac{\sum (x \cdot w)}{\sum w}$$

Note: This is equivalent to counting each data value the number of times given by its weight.

Examples:

- Grade point average. We assign the letter grades the number values A=4, B=3, C=2, D=1, F=0, and then each grade value is counted into the GPA according to the number of credits earned with that grade.
- Course grade. The final grade in this course is calculated according to the following scale: Homework counts for 15%, 3 exams count 20% each, and the final exam is worth 25%. We can weight the score for each component of the final grade with its percentage to calculate the final grade.
- Yield data from OFT and FLD of KVKs also uses weighted means.

Estimating a Mean from a Frequency Table

Given the frequency distribution of a data set, we can make the best estimate of the mean for the data set by using a weighted mean.

1. Calculate the **class midpoint** for each data class. These will be our data values for calculating the weighted mean.

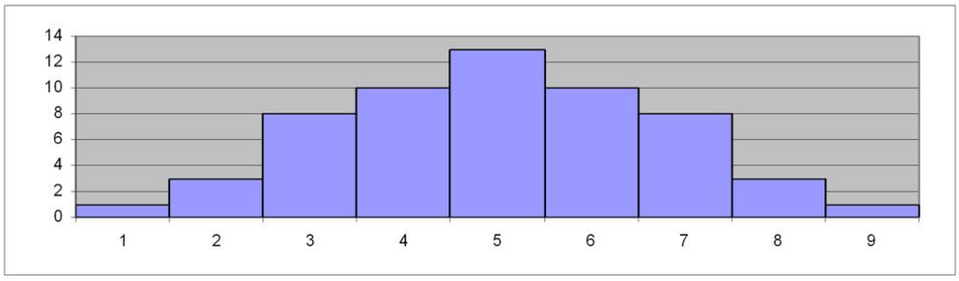
2. Use the **frequency** of the data class as the **weight** for each data class midpoint.

3. Calculate the weighted mean by the weighted mean formula, or:

$$\overline{x} = \frac{\sum (x_{mid} \cdot f)}{\sum f}$$

7.25 Shapes of Data Distributions

Symmetric – The data distribution is approximately the same shape on either side of a central dividing line.

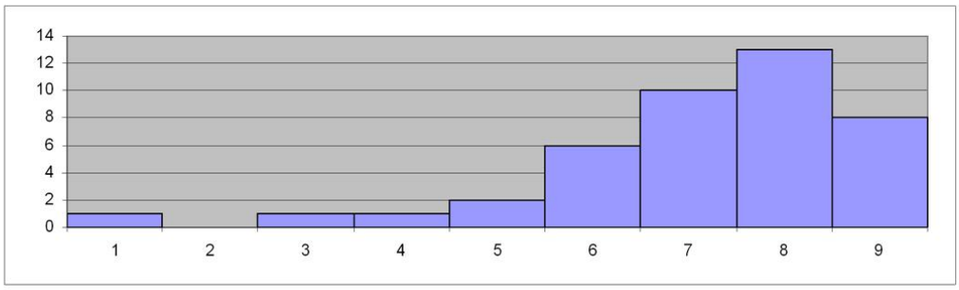


The mean and median (and mode if unimodal) are equal in a symmetric distribution.

Examples: Men's Heights

Left-Skewed – A few data values are much lower than the majority of values in the set. (Tail extends to the left)

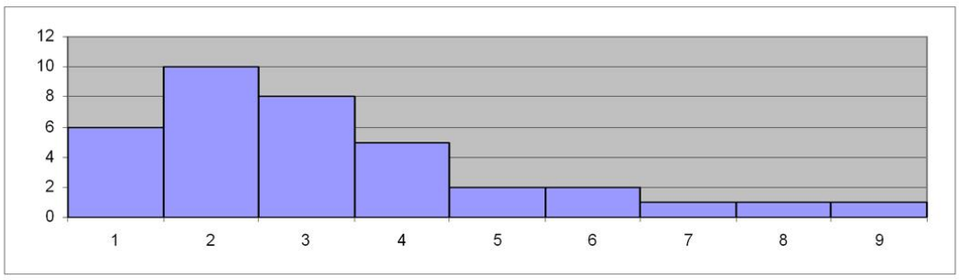
Generally the mean is less than the median (and mode) in a left-skewed distribution.



Example: Yield of a variety with a few farms doing poorly

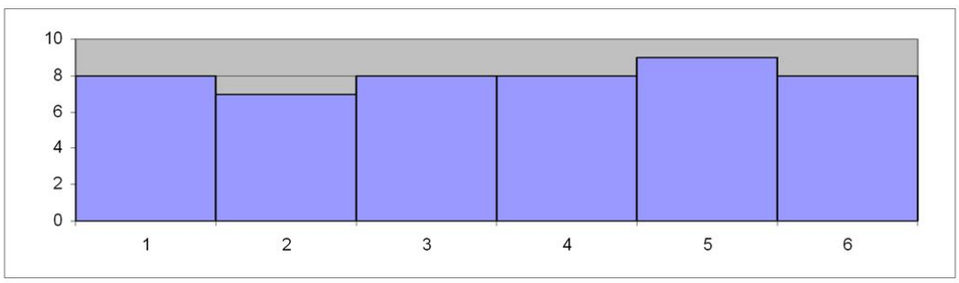
Right-Skewed – A few data values are much higher than the majority of values in the set. (Tail extends to the right)

Generally the mean is greater than the median (and mode) in a right-skewed distribution.



Examples: Income from poultry farming

Uniform – All data values are equally represented.



Example: Number of eggs laid in a farm

7.26 Variation

Variation in a data set is the amount of difference between data values.

In a data set with little variation, almost all data values would be close to one another. The histogram of such a data set would be narrow and tall.

Example: Training Scores: 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5

In a data set with a great deal of variation, the data values would be spread widely. The histogram of this data set would be wide and low.

Example: Attribute Scores: 1, 3, 4, 5, 6, 6, 7, 8, 8, 9, 10

7.27 Common Measures of Variation

1. **Range** – the difference between the largest and smallest data values in a data set.
2. **Standard Deviation** – The most commonly used measure of variation. Its a measure of the "average" distance of a data value from the mean for the data set.

Standard deviation is calculated using two different formulae depending on whether the data set being considered is a population data set or a sample data set.

Population standard deviation is represented by the small Greek letter sigma σ and is calculated using the following formula:

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

Sample standard deviation is represented by *s* and is calculated using the following formula:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

3. **Variance** – the square of the standard deviation. Population variance is represented by σ^2 and sample variance by s^2

Calculating Standard Deviation Using the Formula

1. Calculate the mean of the data set.
2. Subtract the mean from each data value in the set. These values are called the **deviations** of the data values.
3. Square each of the deviations calculated in Step 2.
4. Sum the squares calculated from Step 3.
5. Divide the sum from Step 4 by the population size for population standard deviation or the sample size minus 1 for sample standard deviation.
6. Take the square root of the result of Step 5.

Estimating Standard Deviation using a Frequency Table

Given the frequency distribution of a data set, we can make the best estimate of the standard deviation for the data set by using the same technique as for mean.

1. Calculate the class midpoint for each data class. These will be our data values for calculating the standard deviation.

2. Use the frequency of the data class as the weight for each data class midpoint.

3. Calculate the standard deviation by using the formula:

$$s = \sqrt{\frac{\sum (x_{mid} - \bar{x})^2 \cdot f}{n-1}}$$

Theorems Involving Standard Deviation

The standard deviation of a data set is an important quantity because it limits the number of data values that can be very far (high or low) from average.

The Empirical Rule (68-95-99.7 Rule)

- Applies only to bell-shaped distributions.
- Approximately 68% of data values must be within 1 standard deviation of the mean.
- Approximately 95% of data values must be within 2 standard deviation of the mean.
- Approximately 99.7% of data values must be within 3 standard deviation of the mean.

Example: Men's Heights have a bell-shaped distribution with a mean of 69.2 inches and a standard deviation of 2.9 inches.

7.28 Measures of Position

Fractiles divide a data set into consecutive intervals so that each interval has (at least approximately) the same number of data values. The most common fractiles are:

- **Percentiles** – divide a data set into 100 parts. For example, the 36th percentile is the value which separates the lowest 36% of data values from the highest 64% of data values and is denoted by P_{36}.
- **Quartiles** – divide a data set into fourths. For example, the first quartile Q_1 divides the lowest quarter of a data set from the upper three quarters.
- **Deciles** – divide a data set into 10 parts. For example, the 7th decile is the value which separates the lowest 7/10 of data values from the highest 3/10 of data values and is denoted D_7.

Note: There are 99 percentiles P_1-P_{99}, 3 quartiles Q_1-Q_3, and 9 deciles D_1-D_9.

Note: $P_{50} = Q_2 = D_5$ = Median

7.29 The Standard Score

The **standard score** (or **z-score**) of a data value is the number of standard deviations that the value lies above or below the mean.

Standard Scores can be calculated using the formula:

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

Note: The z-score of a value is positive if the value is above the mean and negative if it is below the mean. The mean itself always has a z-score of 0.

A data value is considered to be **unusual** if it is more than two standard deviations from the mean. A data value is unusually high if it has a z-score larger than 2 and unusually low if it has a z-score of less than -2.

7.30 Statistical Hypotheses

A **statistical hypothesis** is a mathematical claim about a population parameter.

Examples:

- The mean height of women is less than 65 inches tall.
- The percentage of NE farmers favoring organic farming is 97%.
- The average fertilizer use by NE farmers is less than 10 kgs.
- At least 5% of Indian farmers earn more than Rs. 100,000 per year.

We could write the claims above as $\mu < 65$, p=.97, $\mu > 10,000$, and $p \geq .05$

Hypothesis Testing - Basic Procedure

If we wanted to know whether any of the above hypotheses are true, we would conduct a *hypothesis test*. When we test a statistical hypothesis, we follow the following basic procedure:

1. Draw a random sample for the random variable in question.
2. Determine if the results from the sample data are consistent or inconsistent with the hypothesis.
3. If the sample data is significantly different from the claimed hypothesis, we would reject the hypothesis as being false. If the data is not significantly different, we would not reject the hypothesis.

Formal Hypothesis Tests

In a formal hypothesis test, the opposite claims would be given the names **null hypothesis** and **alternative hypothesis**. The null hypotheses is denoted by H_0 and the alternative hypothesis is denoted by H_a. The null and alternative hypotheses need to be assigned as follows:

The null hypothesis is the hypothesis being tested. It must:

- be the hypothesis we want to **reject**
- contain the condition of equality

The alternative hypothesis is always the opposite of the null hypothesis. It must:

- be the hypothesis we want to **support**
- not contain the condition of equality

A formal hypothesis test will always conclude with a decision to reject based on sample data, or the decision that there is not strong enough evidence to reject .

7.31 Types of Error

Whenever sample data is used to make an estimate of a population parameter, there is always a probability of error due to drawing an unusual sample. There are two main types of error that occur in hypothesis tests.

Type I Error – A sample is chosen whose sample data leads to the rejection of the null hypothesis when, in fact, H_o is true.

Type II Error – A sample is chosen whose sample data leads to not rejecting the null hypothesis when, in fact, H_o is false.

	H_o **True**	H_o **False**
H_o **Rejected**	**Type I Error**	**Correct Decision**
H_o **Not Rejected**	**Correct Decision**	**Type II Error**

7.32 Level of Significance

In hypothesis tests, a conservative approach is usually taken toward the rejection of the null hypothesis. That is, we want the probability of making a Type I Error to be small.

The maximum acceptable probability is usually chosen from the beginning of the hypothesis test, and is called the **level of significance** for the test. The level of significance is denoted by , and the most commonly used values are $\alpha = .10, \alpha = .05,$ and $\alpha = .01$.

The probability of making a Type II Error in a hypothesis test is denoted by β. Once α is determined, the value of β is also fixed, but the calculation of this value β is beyond the scope of this chapter.

7.33 Types of Tests

There are three basic types of hypothesis tests:

Left-tailed Test – used when the null hypothesis being tested is a claim that the population parameter is **at least** a given value. Note that the alternative hypothesis then claims that the parameter is **less than** (<) the value.

Example:

$$H_0 : \mu \geq 35,000$$

$$H_a : \mu < 35,000$$

We would reject in the case above if our sample mean was significantly less than 35,000. That is, if our sample mean was in the **left tail** of the distribution of all sample means.

Right-tailed Test – used when the null hypothesis being tested is a claim that the population parameter is **at most** a given value. Note that the alternative hypothesis then claims that the parameter is **greater than** (>) the value.

Example:

$$H_0 : \mu \leq 35,000$$
$$H_a : \mu > 35,000$$

We would reject in this case if our sample mean was significantly more than 35,000. That is, if our sample mean was in the **right tail** of the distribution of all sample means.

Two-tailed Test – used when the null hypothesis being tested is a claim that the population parameter is **equal to** (=) a given value. Note that the alternative hypothesis then claims that the parameter is **not equal to** the value.

Example: The Census claims that the percentage of Shillong residents with a bachelor's degree or higher is 24.4%. We would write the null and alternative hypotheses for this claim as:

$$H_0 : p = .244$$

$$H_a : p \neq .244$$

In this case, we would have to reject if our sample percentage was either significantly more than 24.4%, or significantly less than 24.4%. That is, if our sample proportion was in **either tail** of the distribution of all sample proportions.

Testing a Claim about the Mean Using a Large Sample

When a hypothesis test involves a claim about a population mean, then we will draw a sample and look at the sample mean to test the claim. If the sample drawn is large enough, then the Central Limit Theorem applies, and the distribution of sample means is approximately normal. As usual, we also have that $\mu_{\bar{x}} = \mu$ and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}.$$

Note: Since *s* and *n* are known from the sample data, so we have a good estimate of $\sigma_{\bar{x}}$, but we do not know μ since this is the parameter we are testing a claim about. In order to have a value for μ, we will **always assume that the null hypothesis is true** in any hypothesis test.

Since the null hypothesis must be of one of the following types:

$$\mu = \mu_0, \ \mu \geq \mu_0, \text{ or } \mu \leq \mu_0$$

where μ_0 is a constant, we will always assume for the purpose of our test that $\mu = \mu_0$.

7.34 The Standardized Test Statistic

There are two methods we will use to determine whether to reject or not reject the null hypothesis, but in both cases it will be more convenient to convert our sample mean $\bar{x}$ to a z-score which will be called our **standardized test statistic**.

Since we are assuming $\mu = \mu_0$, we also have $\mu_{\bar{x}} = \mu_0$, and so:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

As long as $\mu = \mu_0$ as assumed, the distribution of the standardized test statistic *z* defined above will be the Standard Normal Distribution.

Example: Suppose we believe that the mean body temperature of healthy adults is less than the commonly accepted measurement of F. A sample of 60 healthy adults is drawn with an average temperature of $\bar{x} = 98.2^\circ$ F and with a sample standard deviation of $s = 1.1^\circ$.

Our hypotheses in this case would be:

$$H_0 : \mu \geq 98.6$$

$$H_a : \mu < 98.6$$

So we have a left-tailed test with $\mu_0 = 98.6$.

Based on our sample data, our standardized test statistic is:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{98.2 - 98.6}{1.1/\sqrt{60}} = \frac{-0.4}{.142} \approx -2.82$$

The P-Value Method

The **P-value** of a test is the probability of drawing a random sample whose standardized test statistic is at least as contrary to the claim of the null hypothesis as that observed in the sample group.

Example: In the hypothesis test for body temperature given above, we had:

$$H_0 : \mu \geq 98.6$$

$$H_a : \mu < 98.6$$

Our sample had a mean temperature of $\bar{x} = 98.2^\circ$ which is contrary to the null hypothesis. Only a sample group with an average temperature less than 98.2 would be stronger evidence against H_0. Thus the P-value of this test is $P(\bar{x} \leq 98.2)$. Since the z-score of $\bar{x}$ is just our standardized test statistic z which has the Standard Normal Distribution, $P(\bar{x} \leq 98.2) = P(z \leq -2.82) = .0024$.

Since the probability of drawing a sample as contrary to the null hypothesis as the observed sample (assuming H_0 is true) is small, we would decide to reject H_0.

Calculating P-Values

In the example above, we calculated the P-value of the test by finding the area to the **left** of the standardized test statistic *z* on the standard normal curve. Notice that the example above was also a **left-tailed** test, and that any hypothesis test which is left-tailed will have the P-value calculated exactly as above.

Similarly, for a **right-tailed** test, we would calculate the P-value by finding the area to the **right** of the standardized test statistic.

For a **two-tailed test**, the null hypothesis is always claiming that $\mu = \mu_0$, and so the sample data is contrary to this claim if the sample mean is either much higher or much lower than μ_0. The P-value for a two tailed test then is the area in both tails of the normal distribution more extreme than the standardized test statistic. Since the normal distribution is symmetric, this is just **twice the area in one tail**.

Deciding to Reject the Null Hypothesis

In the examples above, we saw that a very small P-value would lead us to reject the null hypothesis, and a high P-value would not.

Since the P-value of a test is the probability of randomly drawing a sample at least as contrary to H_0 as the observed sample, we can also think of the P-value as the probability that we will be wrong if we choose to reject H_0 based on our sample data. The P-value then is the probability of making a Type I Error.

Recall that the maximum acceptable probability of making a Type I Error is the level of significance, and is usually determined at the outset of the hypothesis test.

The rule we will use to decide whether to reject H_0 is:

Reject H_0 if $P \leq \alpha$

Do not reject H_0 if $P > \alpha$

7.35 Rejection Regions and Critical Values

A second method to determine whether to reject the null hypothesis is to use **rejection regions** and **critical values**.

A **rejection region** for a hypothesis test is the range of values for the standardized test statistic which would cause us to decide to reject the null hypothesis. **Critical values** for a hypothesis test are the z-scores which separate the rejection region(s) from the non-rejection region. The critical values will be denoted by z_0.

The rejection region for a test is determined by the type of test (left/right/two tailed) and the level of significance α for the test. For a left-tailed test, the rejection region is a region in the left tail of the normal distribution, for a right tailed test, it is in the right tail, and for a two tailed test, there are two equal rejection regions in either tail.

Since the level of confidence is the maximum acceptable probability of a Type I Error, we want the area under the normal curve in the rejection region to have an area of α. We can use this area to find our critical values.

Once we establish the critical values and rejection region, if the standardized test statistic for a sample data set falls in the rejection region, we will reject the null hypothesis.

Example: In our body temperature example, we were using a left-tailed test. If the level of significance was $\alpha = .05$, then the rejection region would be the values in the lowest 5% of the standard normal distribution. Looking up .05 in the standard normal table, we see that this corresponds to a z-score of at most -1.645, so this would be our critical value z_0, and so our rejection region is $z \le -1.645$. Since our standardized test statistic $z = -2.82$ falls into this region, we would choose to reject H_0.

Testing a Claim about the Mean Using a Small Sample

Recall that if a sample of values for a normal random variable is drawn and if the sample size is less than 30, and the population standard deviation is unknown, then the random variable has the Student t-distribution with $n-1$ degrees of freedom.

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

When testing a claim about the mean using sample data from a small sample, we should therefore use the appropriate t-distribution instead of the standard normal distribution to determine our standardized test statistic, critical values, rejection regions, and P-values.

The standardized test statistic for a t-distribution test will be given by:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

7.36 Finding Critical Values for the t-distributions

The process for locating the rejection regions for a t-distribution hypothesis test are the same as for the normal distribution tests. The critical values however will be different.

To find the critical value(s) t_0 for a test, first determine if the test is one-tailed or two-tailed, and the level of significance α. The critical values can be found in the t-distribution table in the front of the book by looking up the entry in the column giving the level of significance and the row giving the degree of freedom.

Note that:

- For a right-tailed test, t_0 is the value in the table
- For a left-tailed test, t_0 is the negative of the value in the table

- For a two-tailed test, there are two critical values t_0 both the value and its opposite

Example: In a knowledge test scores are normally distributed. A sample of scores for 16 trainees has mean score of $\bar{x} = 522.8$ with a sample standard deviation of $s = 154.5$.

Suppose we wished to support the claim that the average knowledge score exceeds 500 using a $\alpha = .05$ level of significance.

The null and alternative hypotheses in this case would be:

$$H_0 : \mu \leq 500$$

$$H_a : \mu > 500$$

So the test is right-tailed with $\mu_0 = 500$.

Using the t-distribution table with One Tail $\alpha = .05$ and 15 degrees of freedom, we get a critical value of $t_0 = 1.753$. The rejection region is thus: $t \geq 1.753$.

The standardized test statistic is given by:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{522.8 - 500}{154.5/\sqrt{16}} = \frac{22.8}{38.625} \approx .59$$

Because the standardized test statistic is not in the rejection region, we would not reject H_0, and so the sample data is not sufficient to support the claim that the mean exceeds 500 at the .05 level of significance.

7.37 Testing a Claim about a Population Proportion

Recall that if a random sample is drawn and if the sample proportion $\hat{p}$ is measured and $n\hat{p} \geq 5$, $n\hat{q} \geq 5$, then the distribution of $\hat{p}$ is approximately normal with $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{pq/n}$.

When testing a claim about a population proportion, the null hypothesis has one of the following forms:

$p = p_0$, $p \geq p_0$, or $p \leq p_0$

So as with the mean, we will assume $p = p_0$ and we will use the following standardized test statistic for a proportion test:

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

This random variable should have the standard normal distribution, and so we will calculate all of our rejection regions, critical values, and P-values using the standard normal distribution as we did when testing a mean using a large sample.

Example: An irrigation pump manufacturer tests pump motors coming off the production line. In one sample of 577 motors, 37 were found to have defects. The company wants to claim that the proportion of motors that are defective is only 4%. Can the company's claim be rejected at the $\alpha = .01$ level of significance?

The null and alternative hypotheses in this case would be:

$$H_0 : p = .04$$

$$H_a : \mu \neq .04$$

So the test is two-tailed with $p_0 = .04$.

Since $\hat{p} = 37/577 \approx .064$, the standardized test statistic is given by:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{.064 - .04}{\sqrt{(.04)(.96)/577}} \approx \frac{.024}{.008} = 3.0$$

Looking up $z = 3.00$ in the standard normal table we get a value of .9987, so $P(z \geq 3.00) = 1 - .9987 = .0013$ and since we have a two-tailed test, the P-value is just twice this amount or .0026. Since this is less than $\alpha = .01$, we can reject the company's claim.

7.38 Correlation

A **correlation** is a relationship between two statistical variables measured from the same population. In this chapter, we will consider only **linear correlation** which comes in three types:

Positive Linear Correlation - *high* values for one variable tend to correspond to *high* values for the second variable.

Examples:

- Height vs. Weight for adults
- Fertilizer dosage vs. Plant growth

Negative Linear Correlation - *high* values for one variable tend to correspond to *low* values for the second variable.

Examples:

- Age vs. Resale Value of a tractor
- Room temperature vs. keeping quality of fruits

Non Linear Correlation - no relationship between the variables or a non-linear relationship.

Examples:

- Plant height vs. Size of farm
- Day of the year vs. Hours of daylight

7.39 Scatter Diagrams

One way to determine the type of linear correlation between two variables is by means of a **scatter diagram**. To construct a scatter diagram, we plot the value of one variable along the x-axis and the other along the y-axis, and then for each member of our population or sample group, we plot a point corresponding to the measurements of the individual.

We can then determine the type of linear correlation as follows:

Positive Linear Correlation

General trend in the plotted points is from **bottom left to top right.**

Negative Linear Correlation

General trend in the plotted points is from **top left to bottom right.**

Non Linear Correlation

No general trend in plotted points, or a non-linear trend.

The **strength** of the linear correlation can be judged by looking at how closely the points approximate a straight line.

Example: The following table shows the Height (x) vs. Femur Length (y) measurements (both in inches) for 10 men:

x	70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6
y	42.5	40.2	44.4	42.8	40	47.3	43.4	40.1	42.1	36

The diagram shows a positive linear correlation between the variables.

7.40 Coefficient of Correlation

A more precise method of determining the type and strength of a linear correlation is to calculate the **coefficient of linear correlation** (denoted by *r*) for the two variables using the formula:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

The coefficient of linear correlation will always be a number between -1 and 1, with a positive value indicating a positive correlation and a negative value a negative correlation. A coefficient of $r = 1$ for a data set indicates perfect positive linear correlation, and $r = -1$ indicates perfect negative linear correlation, while $r = 0$ would indicate no linear correlation. The closer the value of *r* is to ± 1, the stronger the correlation, and the closer to zero, the weaker the correlation.

7.41 Calculating the Coefficient of Correlation

The coefficient of correlation between two variables is most easily calculated by constructing a table (see example below) with columns that contain the *x* and *y* variable values for each individual, the value of *xy* for each individual, and the values of x^2 and y^2 for each individual.

The sum of each column is found, and these sums can then be substituted into the formula above to find *r*.

Example: Using our previous data set of height vs femur length for 10 men, we get the table:

Variable	*x*	*y*	*xy*	x^2	y^2
	70.8	42.5	3009	5012.64	1806.25
	66.2	40.2	2661.24	4382.44	1616.04
	71.7	44.4	3183.48	5140.89	1971.36
	68.7	42.8	2940.36	4719.69	1831.84
	67.6	40	2704	4569.76	1600
	69.2	47.3	3273.16	4788.64	2237.29
	66.5	43.4	2886.1	4422.25	1883.56

Variable	x	y	xy	x^2	y^2
	67.2	40.1	2694.72	4515.84	1608.01
	68.3	42.1	2875.43	4664.89	1772.41
	65.6	36	2361.6	4303.36	1296
Sum	681.8	418.8	28589.09	46520.4	17622.76

The coefficient of correlation for the variables is thus:

$$r = \frac{10(28589.09) - (681.8)(418.8)}{\sqrt{10(46520.4) - (681.8)^2}\sqrt{10(17622.76) - (418.8)^2}} =$$

$$\frac{353.06}{\sqrt{352.76}\sqrt{834.16}} = \frac{353.06}{542.4558} \approx .651$$

7.42 Significance of the Coefficient of Correlation

When the coefficient of correlation is calculated from sample data sets, there is a chance that a linear correlation will be found when, in fact, no correlation exists between the population variables. Therefore, before deciding that a linear correlation exists between two variables when using sample data, we will run a test for significance.

The population parameter representing the coefficient of correlation for population data is denoted by ρ, and we use the sample coefficient r to determine if the hypothesis $H_0 : \rho = 0$ can be rejected.

7.43 Linear Regression

If a pair of variables has a significant linear correlation, then the relationship between the data values can be roughly approximated by a linear equation. The process of finding the linear equation which best fits the data values is known as **linear regression** and the line of best fit is called the **regression line**.

It is a fact of linear algebra and analysis that the least squares line of best fit to a set of data values has an equation of the form $\hat{y} = mx + b$ where:

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad \text{and} \quad b = \overline{y} - m\overline{x} = \frac{(\sum y) - m(\sum x)}{n}$$

Conclusion

Since KVK work is highly social science oriented, a basic understanding of statistical tools is a must for every KVK scientist. Keeping this in view the chapter has presented the fundamental concepts of statistics along with details of basic statistical tests. The same may be utilized in deriving meaningful conclusions from OFT and FLD datas obtained by KVK scientist for better interpretation of the same.