Virtual University of Pakistan

Statistics and Probability

STA301

TABLE OF CONTENTS

LECTURE NO. 1

WHAT IS STATISTICS?

- That science which enables us to draw conclusions about various phenomena on the basis of real data collected on sample-basis
- A tool for data-based research
- Also known as Quantitative Analysis
- A lot of application in a wide variety of disciplines Agriculture, Anthropology, Astronomy, Biology, Economic, Engineering, Environment, Geology, Genetics, Medicine, Physics, Psychology, Sociology, Zoology …. Virtually every single subject from Anthropology to Zoology …. A to Z!
- Any scientific enquiry in which you would like to base your conclusions and decisions on real-life data, you need to employ statistical techniques!
- Now a day, in the developed countries of the world, there is an active movement for of Statistical Literacy.

THE NATURE OF THIS DISCIPLINE

DESCRIPTIVE STATISTICS

↓

PROBABILITY

↓

INFERENTIAL STATISTICS

MEANINGS OF 'STATISTICS'

The word "Statistics" which comes from the Latin words *status*, meaning a political state, originally meant information useful to the state, for example, information about the sizes of population sand armed forces. But this word has now acquired different meanings.

- In the *first place*, the word *statistics* refers to "numerical facts systematically arranged". In this sense, the word statistics is always used in plural. We have, for instance, statistics of prices, statistics of road accidents, statistics of crimes, statistics of births, statistics of educational institutions, etc. In all these examples, the word statistics denotes a set of numerical data in the respective fields. This is the meaning the man in the street gives to the word *Statistics* and most people usually use the word *data* instead.
- In the *second place*, the word *statistics* is defined as a discipline that includes procedures and techniques used to collect process and analyze numerical data to make inferences and to research decisions in the face of

uncertainty. It should of course be borne in mind that uncertainty does not imply ignorance but it refers to the incompleteness and the instability of data available. In this sense, the word statistics is used in the singular. As it embodies more of less all stages of the general process of learning, sometimes called *scientific method*, statistics is characterized as a science. Thus the word *statistics* used in the plural refers to a set of numerical information and in the singular, denotes the science of basing decision on numerical data. It should be noted that statistics as a subject is mathematical in character.

- *Thirdly*, the word statistics are numerical quantities calculated from sample observations; a single quantity that has been so collected is called a *statistic*. The mean of a sample for instance is a statistic. The word statistics is plural when used in this sense.

CHARACTERISTICS OF THE SCIENCE OF STATISTICS

Statistics is a discipline in its own right. It would therefore be desirable to know the characteristic features of statistics in order to appreciate and understand its general nature. Some of its important characteristics are given below:

- Statistics deals with the behaviour of aggregates or large groups of data. It has nothing to do with what is happening to a particular individual or object of the aggregate.
- Statistics deals with aggregates of observations of the same kind rather than isolated figures.
- Statistics deals with variability that obscures underlying patterns. No two objects in this universe are exactly alike. If they were, there would have been no statistical problem.
- Statistics deals with uncertainties as every process of getting observations whether controlled or uncontrolled, involves deficiencies or chance variation. That is why we have to talk in terms of probability.
- Statistics deals with those characteristics or aspects of things which can be described numerically either by counts or by measurements.
- Statistics deals with those aggregates which are subject to a number of random causes, *e.g.* the heights of persons are subject to a number of causes such as race, ancestry, age, diet, habits, climate and so forth.
- Statistical laws are valid *on the average* or in the long run. There is n guarantee that a certain law will hold in all cases. Statistical inference is therefore made in the face of uncertainty.
- Statistical results might be misleading the incorrect if sufficient care in collecting, processing and interpreting the data is not exercised or if the statistical data are handled by a person who is not well versed in the subject mater of statistics.

THE WAY IN WHICH STATISTICS WORKS

As it is such an important area of knowledge, it is definitely useful to have a fairly good idea about the way in which it works, and this is exactly the purpose of this introductory course.
The following points indicate some of the main functions of this science:

- Statistics assists in summarizing the larger set of data in a form that is easily understandable.
- Statistics assists in the efficient design of laboratory and field experiments as well as surveys.
- Statistics assists in a sound and effective planning in any field of inquiry.
- Statistics assists in drawing general conclusions and in making predictions of how much of a thing will happen under given conditions.

IMPORTANCE OF STATISTICS IN VARIOUS FIELDS

As stated earlier, Statistics is a discipline that has finds application in the most diverse fields of activity. It is perhaps a subject that should be used by everybody. Statistical techniques being powerful tools for analyzing numerical data are used in almost every branch of learning. In all areas, statistical techniques are being increasingly used, and are developing very rapidly.

- A modern administrator whether in public or private sector leans on statistical data to provide a factual basis for decision.

- A politician uses statistics advantageously to lend support and credence to his arguments while elucidating the problems he handles.
- A businessman, an industrial and a research worker all employ statistical methods in their work. Banks, Insurance companies and Government all have their statistics departments.

- A social scientist uses statistical methods in various areas of socio-economic life a nation. It is sometimes said that "a social scientist without an adequate understanding of statistics, is often like the blind man groping in a dark room for a black cat that is not there".

THE MEANING OF DATA

The word "data" appears in many contexts and frequently is used in ordinary conversation. Although the word carries something of an aura of scientific mystique, its meaning is quite simple and mundane. It is Latin for "those that are given" (the singular form is "datum"). Data may therefore be thought of as the *results of observation*.

EXAMPLES OF DATA

- Data are collected in many aspects of everyday life.

- Statements given to a police officer or physician or psychologist during an interview are data.

- So are the correct and incorrect answers given by a student on a final examination.

- Almost any athletic event produces data.

- The time required by a runner to complete a marathon,

- The number of errors committed by a baseball team in nine innings of play.

- And, of course, data are obtained in the course of scientific inquiry:

- the positions of artifacts and fossils in an archaeological site,

- The number of interactions between two members of an animal colony during a period of observation,

- The spectral composition of light emitted by a star.

OBSERVATIONS AND VARIABLES

In statistics, an *observation* often means any sort of numerical recording of information, whether it is a physical measurement such as height or weight; a classification such as heads or tails, or an answer to a question such as yes or no.

VARIABLES

A characteristic that varies with an individual or an object is called a *variable*. For example, age is a variable as it varies from person to person. A variable can assume a number of values. The given set of all possible values from which the variable takes on a value is called its Domain. If for a given problem, the domain of a variable contains only one value, then the variable is referred to as a *constant*.

QUANTITATIVE AND QUALITATIVE VARIABLES

Variables may be classified into quantitative and qualitative according to the form of the characteristic of interest. A variable is called a *quantitative variable* when a characteristic can be expressed numerically such as age, weight, income or number of children. On the other hand, if the characteristic is non-numerical such as education, sex, eye-colour, quality, intelligence, poverty, satisfaction, etc. the variable is referred to as a *qualitative variable*. A qualitative characteristic is also called an *attribute*. An individual or an object with such a characteristic can be counted or enumerated after having been assigned to one of the several mutually exclusive classes or categories.

DISCRETE AND CONTINUOUS VARIABLES

A quantitative variable may be classified as discrete or continuous. *A discrete* variable is one that can take only a discrete set of integers or whole numbers, which is the values, are taken by jumps or breaks. A discrete variable represents *count* data such as the number of persons in a family, the number of rooms in a house, the number of deaths in an accident, the income of an individual, etc.

A variable is called a *continuous* variable if it can take on any value-fractional or integral—within a given interval, *i.e.* its domain is an interval with all possible values without gaps. A continuous variable represents measurement data such as the age of a person, the height of a plant, the weight of a commodity, the temperature at a place, etc.

A variable whether countable or measurable, is generally denoted by some symbol such as X or Y and X_i or X_j represents the ith or j^{th} value of the variable. The subscript i or j is replaced by a number such as 1,2,3, … when referred to a particular value.

MEASUREMENT SCALES

By *measurement*, we usually mean the assigning of number to observations or objects and scaling is a process of measuring. The four scales of measurements are briefly mentioned below:

NOMINAL SCALE

The classification or grouping of the observations into mutually exclusive qualitative categories or classes is said to constitute a *nominal scale*. For example, students are classified as male and female. Number 1 and 2 may also be used to identify these two categories. Similarly, rainfall may be classified as heavy moderate and light. We may use number 1, 2 and 3 to denote the three classes of rainfall. The numbers when they are used only to identify the categories of the given scale carry no numerical significance and there is no particular order for the grouping.

ORDINAL OR RANKING SCALE

It includes the characteristic of a nominal scale and in addition has the property of *ordering* or *ranking* of measurements. For example, the performance of students (or players) is rated as excellent, good fair or poor, etc. Number 1, 2, 3, 4 etc. are also used to indicate ranks. The only relation that holds between any pair of categories is that of "greater than" (or more preferred).

INTERVAL SCALE

A measurement scale possessing a constant interval size (distance) but not a true zero point, is called an *interval scale*. Temperature measured on either the Celsius or the Fahrenheit scale is an outstanding example of interval scale because the same difference exists between 20^o C (68^o F) and 30^o C (86^o F) as between 5^o C (41^o F) and 15^o C (59^o F). It cannot be said that a temperature of 40 degrees is twice as hot as a temperature of 20 degree, i.e. the ratio 40/20 has no meaning. The arithmetic operation of addition, subtraction, etc. is meaningful.

RATIO SCALE

It is a special kind of an interval scale where the sale of measurement has a true *zero* point as its origin. The ratio scale is used to measure weight, volume, distance, money, etc. The, key to differentiating interval and ratio scale is that the zero point is meaningful for ratio scale.

ERRORS OF MEASUREMENT

Experience has shown that a continuous variable can never be measured with perfect fineness because of certain habits and practices, methods of measurements, instruments used, etc. the measurements are thus always recorded correct to the nearest units and hence are of limited accuracy. The *actual* or *true* values are, however, assumed to exist. For example, if a student's weight is recorded as 60 kg (correct to the nearest kilogram), his true weight in fact lies between 59.5 kg and 60.5 kg, whereas a weight recorded as 60.00 kg means the true weight is known to lie between 59.995 and 60.005 kg. Thus there is a difference, however small it may be between the measured value and the true value. This sort of departure from the true value is technically known as the *error of measurement.* In other words, if the observed value and the true value of a variable are denoted by x and $x + \varepsilon$ respectively, then the difference $(x + \varepsilon) - x$, i.e. ε is the error. This error involves the unit of measurement of x and is therefore called an *absolute error*. An absolute error divided by the true value is called the *relative error*. Thus the relative error $= \frac{\varepsilon}{x + \varepsilon}$, which when multiplied by 100, is *percentage error*. These errors are independent of the units of measurement of x. It ought to be noted that an error has both magnitude and direction and that the word *error* in statistics does not mean mistake which is a chance inaccuracy.

BIASED AND RANDOM ERRORS

An error is said to be *biased* when the observed value is consistently and constantly higher or lower than the true value. Biased errors arise from the personal limitations of the observer, the imperfection in the instruments used or some other conditions which control the measurements. These errors are not revealed by repeating the measurements. They are cumulative in nature, that is, the greater the number of measurements, the greater would be the magnitude of error. They are thus more troublesome. These errors are also called *cumulative* or *systematic errors*.

An error, on the other hand, is said to be unbiased when the deviations, i.e. the excesses and defects, from the true value tend to occur equally often. Unbiased errors and revealed when measurements are repeated and they tend to cancel out in the long run. These errors are therefore *compensating* and are also known as *random errors* or *accidental errors*.

LECTURE NO. 2

Steps involved in a Statistical Research-Project

- Collection of Data:
  - Primary Data
  - Secondary Data
- Sampling:
  - Concept of Sampling
  - Non-Random Versus Random Sampling
  - Simple Random Sampling
  - Other Types of Random Sampling

STEPS INVOLVED IN ANY STATISTICAL RESEARCH

- Topic and significance of the study
- Objective of your study
- Methodology for data-collection
  - Source of your data
  - Sampling methodology
  - Instrument for collecting data

As far as the objectives of your research are concerned, they should be stated in such a way that you are absolutely clear about the goal of your study --- EXACTLY WHAT IT IS THAT YOU ARE TRYING TO FIND OUT?
As far as the *methodology for DATA-COLLECTION* is concerned, you need to consider:

- Source of your data (the statistical population)
- Sampling Methodology
- Instrument for collecting data

COLLECTION OF DATA

The most important part of statistical work is perhaps the collection of data. Statistical data are collected either by a COMPLETE enumeration of the whole field, called CENSUS, which in many cases would be too costly and too time consuming as it requires large number of enumerators and supervisory staff, or by a PARTIAL enumeration associated with a SAMPLE which saves much time and money.

PRIMARY AND SECONDARY DATA

Data that have been originally collected (raw data) and have not undergone any sort of statistical treatment, are called *PRIMARY data*. Data that have undergone any sort of treatment by statistical methods at least ONCE, i.e. the data that have been collected, classified, tabulated or presented in some form for a certain purpose, are called *SECONDARY data*.

COLLECTION OF PRIMARY DATA
One or more of the following methods are employed to collect primary data:

- Direct Personal Investigation
- Indirect Investigation
- Collection through Questionnaires
- Collection through Enumerators
- Collection through Local Sources

DIRECT PERSONAL INVESTIGATION

In this method, an investigator collects the information personally from the individuals concerned. Since he interviews the informants himself, the information collected is generally considered quite accurate and complete. This method may prove very costly and time-consuming when the area to be covered is vast. However, it is useful for laboratory experiments or localized inquiries. Errors are likely to enter the results due to personal bias of the investigator.

INDIRECT INVESTIGATION

Sometimes the direct sources do not exist or the informants hesitate to respond for some reason or other. In such a case, third parties or witnesses having information are interviewed. Moreover, due allowance is to be made for the personal bias. This method is useful when the information desired is complex or there is reluctance or indifference on the part of the informants. It can be adopted for extensive inquiries.

COLLECTION THROUGH QUESTIONNAIRES

A questionnaire is an inquiry form comprising of a number of pertinent questions with space for entering information asked. The questionnaires are usually sent by mail, and the informants are requested to return the questionnaires to the investigator after doing the needful within a certain period. This method is cheap, fairly expeditious and good for extensive inquiries. But the difficulty is that the majority of the respondents (i.e. persons who are required to answer the questions) do not care to fill the questionnaires in, and to return them to the investigators. Sometimes, the questionnaires are returned incomplete and full of errors. Students, in spite of these drawbacks, this method is considered as the STANDARD method for routine business and administrative inquiries.

It is important to note that the questions should be few, brief, very simple, and easy for all respondents answer, clearly worded and not offensive to certain respondents.

COLLECTION THROUGH ENUMERATORS

Under this method, the information is gathered by employing trained enumerators who assist the informants in making the entries in the schedules or questionnaires correctly. This method gives the most reliable information if the enumerator is well-trained, experienced and tactful. Students, it is considered the BEST method when a large-scale governmental inquiry is to be conducted. This method can generally not be adopted by a private individual or institution as its cost would be prohibitive to them.

COLLECTION THROUGH LOCAL SOURCES

In this method, there is no formal collection of data but the agents or local correspondents are directed to collect and send the required information, using their own judgment as to the best way of obtaining it. This method is cheap and expeditious, but gives only the estimates.

COLLECTION OF SECONDARY DATA

The secondary data may be obtained from the following sources:

- Official, e.g. the publications of the Statistical Division, Ministry of Finance, the Federal and Provincial Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labour, etc.
- Semi-Official, e.g., State Bank of Pakistan, Railway Board, Central Cotton Committee, Boards of Economic Inquiry, District Councils, Municipalities, etc.
- Publications of Trade Associations, Chambers of Commerce, etc
- Technical and Trade Journals and Newspapers
- Research Organizations such as universities, and other institutions

Let us now consider the POPULATION from which we will be collecting our data. In this context, the first important question is: Why do we have to resort to Sampling?
The answer is that: If we have available to us every value of the variable under study, then that would be an ideal and a perfect situation. But, the problem is that this ideal situation is very rarely available --- very rarely do we have access to the entire population.
The census is an exercise in which an attempt is made to cover the entire population. But, as you might know, even the most developed countries of the world cannot afford to conduct such a huge exercise on an annual basis!
More often than not, we have to conduct our research study on a sample basis. In fact, the goal of the science of Statistics is to draw conclusions about large populations on the basis of information contained in small samples.

'POPULATION'

A statistical population is the collection of every member of a group possessing the same basic and defined characteristic, but varying in amount or quality from one member to another.

EXAMPLES

- **Finite population:**
  - IQ's of all children in a school.
- **Infinite population:**
  - Barometric pressure:
    (There are an indefinitely large number of points on the surface of the earth).
  - A flight of migrating ducks in Canada

(Many finite pops are so large that they can be treated as effectively infinite). The examples that we have just considered are those of existent populations.
A hypothetical population can be defined as the aggregate of all the conceivable ways in which a specified event can happen.

For Example:

- 1)All the possible outcomes from the throw of a die – however long we throw the die and record the results, we could always continue to do so far a still longer period in a theoretical concept – one which has no existence in reality.
- 2) The No. of ways in which a football team of 11 players can be selected from the 16 possible members named by the Club Manager.

We also need to differentiate between the sampled population and the target population. Sampled population is that from which a sample is chosen whereas the population about which information is sought is called the target population thus our population will consist of the total no. of students in all the colleges in the Punjab.

Suppose on account of shortage of resources or of time, we are able to conduct such a survey on only 5 colleges scattered throughout the province. In this case, the students of all the colleges will constitute the target pop whereas the students of those 5 colleges from which the sample of students will be selected will constitute the sampled population. The above discussion regarding the population, you must have realized how important it is to have a very well-defined population.

The next question is: How will we draw a sample from our population?

The answer is that: In order to draw a random sample from a finite population, the first thing that we need is the complete list of all the elements in our population.

This list is technically called the **FRAME.**

SAMPLING FRAME

A sampling frame is a complete list of all the elements in the population. For example:

- The complete list of the BCS students of Virtual University of Pakistan on February 15, 2003
- Speaking of the sampling frame, it must be kept in mind that, as far as possible, our frame should be free from various types of defects:
- does not contain inaccurate elements
- is not incomplete
- is free from duplication, and
- Is not out of date.

Next, let's talk about the sample that we are going to draw from this population.

As you all know, a sample is only a part of a statistical population, and hence it can represent the population to only to some extent. Of course, it is intuitively logical that the larger the sample, the more likely it is to represent the population. Obviously, the limiting case is that: when the sample size tends to the population size, the sample will tend to be identical to the population. But, of course, in general, the sample is much smaller than the population.

The point is that, in general, statistical sampling seeks to determine how accurate a description of the population the sample and its properties will provide. We may have to compromise on accuracy, but there are certain such advantages of sampling because of which it has an extremely important place in data-based research studies.

ADVANTAGES OF SAMPLING

1. Savings in time and money.

- Although cost per unit in a sample is greater than in a complete investigation, the total cost will be less (because the sample will be so much smaller than the statistical population from which it has been drawn).
- A sample survey can be completed faster than a full investigation so that variations from sample unit to sample unit over time will largely be eliminated.
- Also, the results can be processed and analyzed with increased speed and precision because there are fewer of them.

2. More detailed information may be obtained from each sample unit.

3. Possibility of follow-up:

(After detailed checking, queries and omissions can be followed up --- a procedure which might prove impossible in a complete survey).

4. Sampling is the only feasible possibility where tests to destruction are undertaken or where the population is effectively infinite.

The next two important concepts that need to be considered are those of sampling and non-sampling errors.

SAMPLING & NON-SAMPLING ERRORS

1. SAMPLING ERROR

The difference between the estimate derived from the sample (i.e. the statistic) and the true population value (i.e. the parameter) is technically called the sampling error. For example,

Sampling error = $\bar{X} - \mu$

Sampling error arises due to the fact that a sample cannot exactly represent the pop, even if it is drawn in a correct manner

2. NON-SAMPLING ERROR

Besides sampling errors, there are certain errors which are not attributable to sampling but arise in the process of data collection, even if a complete count is carried out.
Main sources of non sampling errors are:

- The defect in the sampling frame.
- Faulty reporting of facts due to personal preferences.
- Negligence or indifference of the investigators
- Non-response to mail questionnaires.

These (non-sampling) errors can be avoided through

- Following up the non-response,
- Proper training of the investigators.
- Correct manipulation of the collected information,

Let us now consider exactly what is meant by 'sampling error': We can say that there are two types of non-response --- partial non-response and total non-response. *'Partial non-response'* implies that the respondent refuses to answer some of the questions. On the other hand, '*total non-response'* implies that the respondent refuses to answer any of the questions. Of course, the problem of late returns and non-response of the kind that I have just mentioned occurs in the case of HUMAN populations. Although refusal of sample units to cooperate is encountered in interview surveys, it is far more of a problem in mail surveys. It is not uncommon to find the response rate to mail questionnaires as low as 15 or 20%.The provision of INFORMATION ABOUT THE PURPOSE OF THE SURVEY helps in stimulating interest, thus increasing the chances of greater response. Particularly if it can be shown that the work will be to the ADVANTAGE of the respondent IN THE LONG RUN.
Similarly, the respondent will be encouraged to reply if a pre-paid and addressed ENVELOPE is sent out with the questionnaire. But in spite *of* these ways of reducing non-response, we are bound to have some amount of non-response. Hence, a decision has to be taken about how many RECALLS should be made.
The term 'recall' implies that we approach the respondent more than once in order to persuade him to respond to our queries.
Another point worth considering is:
How long the process of data collection should be continued? Obviously, no such process can be carried out for an indefinite period of time! In fact, the longer the time period over which the survey is conducted, the greater will be the potential VARIATIONS in attitudes and opinions of the respondents. Hence, a well-defined cut-off date generally needs to be established. Let us now look at the various ways in which we can select a sample from our population. We begin by looking at the difference between non-random and RANDOM sampling. First of all, what do we mean by non-random sampling?

NONRANDOM SAMPLING

'Nonrandom sampling' implies that kind of sampling in which the population units are drawn into the sample by using one's personal judgment. This type of sampling is also known as purposive sampling. Within this category, one very important type of sampling is known as Quota Sampling.

QUOTA SAMPLING

In this type of sampling, the selection of the sampling unit from the population is no longer dictated by chance. A sampling frame is not used at all, and the choice of the actual sample units to be interviewed is left to the discretion of the interviewer. However, the interviewer is restricted by quota controls. For example, one particular interviewer may be told to interview ten married women between thirty and forty years of age living in town X, whose husbands are professional workers, and five unmarried professional women of the same age living in the same town. Quota sampling is often used in commercial surveys such as consumer market-research. Also, it is often used in public opinion polls.

ADVANTAGES OF QUOTA SAMPLING

- There is no need to construct a frame.
- It is a very quick form of investigation.
- Cost reduction.

DISADVANTAGES

- It is a subjective method. One has to choose between objectivity and convenience.
- If random sampling is not employed, it is no longer theoretically possible to evaluate the sampling error.
- (Since the selection of the elements is not based on probability theory but on the personal judgment of the interviewer, hence the precision and the reliability of the estimates can not be determined objectively i.e. in terms of probability.)
- Although the purpose of implementing quota controls is to reduce bias, bias creeps in due to the fact that the interviewer is FREE to select particular individuals within the quotas. (Interviewers usually look for persons who either agree with their points of view or are personally known to them or can easily be contacted.)
- Even if the above is not the case, the interviewer may still be making unsuitable selection of sample units.
- (Although he may put some qualifying questions to a potential respondent in order to determine whether he or she is of the type prescribed by the quota controls, some features must necessarily be decided arbitrarily by the interviewer, the most difficult of these being social class.)

If mistakes are being made, it is almost impossible for the organizers to detect these, because follow-ups are not possible unless a detailed record of the respondents' names, addresses etc. has been kept.
Falsification of returns is therefore more of a danger in quota sampling than in random sampling. In spite of the above limitations, it has been shown by F. Edwards that a well-organized quota survey with well-trained interviewers can produce quite adequate results.
Next, let us consider the concept of random sampling.

RANDOM SAMPLING

The theory of statistical sampling rests on the assumption that the selection of the sample units has been carried out in a random manner.
By random sampling we mean sampling that has been done by adopting the lottery method.

TYPES OF RANDOM SAMPLING

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling
- Cluster Sampling
- Multi-stage Sampling, etc.

In this course, I will discuss with you the simplest type of random sampling i.e. simple random sampling.

SIMPLE RANDOM SAMPLING

In this type of sampling, the chance of any one element of the parent pop being included in the sample is the same as for any other element. By extension, it follows that, in simple random sampling, the chance of any one sample appearing is the same as for any other. There exists quite a lot of misconception regarding the concept of random sampling. Many a time, haphazard selection is considered to be equivalent to simple random sampling.
For example, a market research interviewer may select women shoppers to find their attitude to brand X of a product by stopping one and then another as they pass along a busy shopping area --- and he may think that he has accomplished simple random sampling!

Actually, there is a strong possibility of bias as the interviewer may tend to ask his questions of young attractive women rather than older housewives, or he may stop women who have packets of brand X prominently on show in their shopping bags!.
In this example, there is no suggestion of INTENTIONAL bias! From experience, it is known that the human being is a poor random selector --- one who is very subject to bias.
Fundamental psychological traits prevent complete objectivity, and no amount of training or conscious effort can eradicate them. As stated earlier, random sampling is that in which population units are selected by the lottery method.

As you know, the traditional method of writing people's names on small pieces of paper, folding these pieces of paper and shuffling them is very cumbersome!
A much more convenient alternative is the use of *RANDOM NUMBERS TABLES.*
A random number table is a page full of digits from zero to 9. These digits are printed on the page in a TOTALLY random manner i.e. there is no systematic pattern of printing these digits on the page.

ONE THOUSAND RANDOM DIGITS

2	3	1	5	7	5	4	8	5	9	0	1	8	3	7	2	5	9	9	3	7	6	2	4	9	7	0	8	8	6	9	5	2	3	0	3	6	7	4	4
0	5	5	4	5	5	5	0	4	3	1	0	5	3	7	4	3	5	0	8	9	0	6	1	1	8	3	7	4	4	1	0	9	6	2	2	1	3	4	3
1	4	8	7	1	6	0	3	5	0	3	2	4	0	4	3	6	2	2	3	5	0	0	5	1	0	0	3	2	2	1	1	5	4	3	8	0	8	3	4
3	8	9	7	6	7	4	9	5	1	9	4	0	5	1	7	5	8	5	3	7	8	8	0	5	9	0	1	9	4	3	2	4	2	8	7	1	6	9	5
9	7	3	1	2	6	1	7	1	8	9	9	7	5	5	3	0	8	7	0	9	4	2	5	1	2	5	8	4	1	5	4	8	8	2	1	0	5	1	3
1	1	7	4	2	6	9	3	8	1	4	4	3	3	9	3	0	8	7	2	3	2	7	9	7	3	3	1	1	8	2	2	6	4	7	0	6	8	5	0
4	3	3	6	1	2	8	8	5	9	1	1	0	1	6	4	5	6	2	3	9	3	0	0	9	0	0	4	9	9	4	3	6	4	0	7	4	0	3	6
9	3	8	0	6	2	0	4	7	8	3	8	2	6	8	0	4	4	9	1	5	5	7	5	1	1	8	9	3	2	5	8	4	7	5	5	2	5	7	1
4	9	5	4	0	1	3	1	8	1	0	8	4	2	9	8	4	1	8	7	6	9	5	3	8	2	9	6	6	1	7	7	7	3	8	0	9	5	2	7
3	6	7	6	8	7	2	6	3	3	3	7	9	4	8	2	1	5	6	9	4	1	9	5	9	6	8	6	7	0	4	5	2	7	4	8	3	8	8	0
0	7	0	9	2	5	2	3	9	2	2	4	6	2	7	1	2	6	0	7	0	6	5	5	8	4	5	3	4	4	6	7	3	3	8	4	5	3	2	0
4	3	3	1	0	0	1	0	8	1	4	4	8	6	3	8	0	3	0	7	5	2	5	5	5	1	6	1	4	8	8	9	7	4	2	9	4	6	4	7
6	1	5	7	0	0	6	3	6	0	0	6	1	7	3	6	3	7	7	5	6	3	1	4	8	9	5	1	2	3	3	5	0	1	7	4	6	9	9	3
3	1	3	5	2	8	3	7	9	9	1	0	7	7	9	1	8	9	4	1	3	1	5	7	9	7	6	4	4	8	6	2	5	8	4	8	6	9	1	9
5	7	0	4	8	8	6	5	2	6	2	7	7	9	5	9	3	6	8	2	9	0	5	2	9	5	6	5	4	6	3	5	0	6	5	3	2	2	5	4
0	9	2	4	3	4	4	2	0	0	6	8	7	2	1	0	7	1	3	7	3	0	7	2	9	7	5	7	3	6	0	9	2	9	8	2	7	6	5	0
9	7	9	5	5	3	5	0	1	8	4	0	8	9	4	8	8	3	2	9	5	2	2	3	0	8	2	5	2	1	2	2	5	3	2	6	1	5	8	7
9	3	7	3	2	5	9	5	7	0	4	3	7	8	1	9	8	8	8	5	5	6	6	7	1	6	6	8	2	6	9	5	9	9	6	4	4	5	6	9
7	2	6	2	1	1	1	2	2	5	0	0	9	2	2	6	8	2	6	4	3	5	6	6	6	5	9	4	3	4	7	1	6	8	7	5	1	8	6	7
6	1	0	2	0	7	4	4	1	8	4	5	3	7	1	2	0	7	9	4	9	5	9	1	7	3	7	8	6	6	9	9	5	3	6	1	9	3	7	8
9	7	8	3	9	8	5	4	7	4	3	3	0	5	5	9	1	7	1	8	4	5	4	7	3	5	4	1	4	4	2	2	0	3	4	2	3	0	0	0
8	9	1	6	0	9	7	1	9	2	2	2	2	3	2	9	0	6	3	7	3	5	0	5	5	4	5	4	8	9	8	8	4	3	8	1	6	3	6	1
2	5	9	6	6	8	8	2	2	0	6	2	8	7	1	7	9	2	6	5	0	2	8	2	3	5	2	8	6	2	8	4	9	1	9	5	4	8	8	3
8	1	4	4	3	3	1	7	1	9	0	5	0	4	9	5	4	8	0	6	7	4	6	9	0	0	7	5	6	7	6	5	0	1	7	1	6	5	4	5
1	1	3	2	2	5	4	9	3	1	4	2	3	6	2	3	4	3	8	6	0	8	6	2	4	9	7	6	6	7	4	2	2	4	5	2	3	2	4	5

Actually, Random Number Tables are constructed according to certain mathematical principles so that each digit has the same chance of selection. Of course, nowadays randomness may be achieved electronically. Computers have all those programmes by which we can generate random numbers.

EXAMPLE

The following frequency table of distribution gives the ages of a population of 1000 teen-age college students in a particular country.
Select a sample of 10 students using the random numbers table. Find the sample mean age and compare with the population mean age.

Student-Population of a College

Age (X)	No. of Students (f)
13	6
14	61
15	270
16	491
17	153
18	15
19	4
	1000

How will we proceed to select our sample of size 10 from this population of size 1000?

The first step is to allocate to each student in this population a sampling number. For this purpose, we will begin by constructing a column of cumulative frequencies.

AGE X	No. of Students f	Cumulative Frequency cf
13	6	6
14	61	67
15	270	337
16	491	828
17	153	981
18	15	996
19	4	1000
	1000	

Now that we have the cumulative frequency of each class, we are in a position to allocate the sampling numbers to all the values in a class. As the frequency as well as the cumulative frequency of the first class is 6, we allocate numbers 000 to 005 to the six students who belong to this class.

AGE X	No. of Students f	cf	Sampling Numbers
13	6	6	000 – 005
14	61	67	
15	270	337	
16	491	828	
17	153	981	
18	15	996	
19	4	1000	
	1000		

As the cumulative frequency of the second class is 67 while that of the first class was 6, therefore we allocate sampling numbers 006 to 066 to the 61 students who belong to this class.

AGE X	No. of Students f	cf	Sampling Numbers
13	6	6	000 – 005
14	61	67	006 – 066
15	270	337	
16	491	828	
17	153	981	
18	15	996	
19	4	1000	
	1000		

As the cumulative frequency of the third class is 337 while that of the second class was 67, therefore we allocate sampling numbers 007 to 337 to the 270 students who belong to this class.

AGE X	No. of Students f	cf	Sampling Numbers
13	6	6	000 – 005
14	61	67	006 – 066
15	270	337	067 – 336
16	491	828	
17	153	981	
18	15	996	
19	4	1000	
	1000		

Proceeding in this manner, we obtain the column of sampling numbers.

AGE X	**No. of Students f**	**cf**	**Sampling Numbers**
13	**6**	**6**	**000 – 005**
14	**61**	**67**	**006 – 066**
15	**270**	**337**	**067 – 336**
16	**491**	**828**	**337 – 827**
17	**153**	**981**	**828 – 980**
18	**15**	**996**	**981 – 995**
19	**4**	**1000**	**996 - 999**
	1000		

The column implies that the first student of the first class has been allocated the sampling number 000, the second student has been allocated the sampling 001, and, proceeding in this fashion, the last student i.e. the 1000th student has been allocated the sampling number 999.

The question is: Why did we not allot the number 0001 to the first student and the number 1000 to the 1000th student? The answer is that we could do that but that would have meant that every student would have been allocated a four-digit number, whereas by shifting the number backward by 1, we are able to allocate to every student a three-digit number --- which is obviously simpler.
The next step is to SELECT 10 RANDOM NUMBERS from the random number table. This is accomplished by closing one's eyes and letting one's finger land anywhere on the random number table. In this example, since all our sampling numbers are three-digit numbers, hence we will read three digits that are adjacent to each other at that position where our finger landed. Suppose that we adopt this procedure and our random numbers come out to be 041, 103, 374, 171, 508, 652, 880, 066, 715, 471
Selected Random Numbers:
041, 103, 374, 171, 508, 652, 880, 066, 715, 471
Thus the corresponding ages are:
14, 15, 16, 15, 16, 16, 17, 15, 16, 16

EXPLANATION

Our first selected random number is 041 which mean that we have to pick up the 42nd student. The cumulative frequency of the first class is 6 whereas the cumulative frequency of the second class is 67. This means that definitely the 42nd student does not belong to the first class but does belong to the second class.

AGE X	No. of Students f	cf
13	6	6
14	61	67
15	270	337
16	491	828
17	153	981
18	15	996
19	4	1000
	1000	

The age of each student in this class is 14 years; hence, obviously, the age of the 42nd student is also 14 years. This is how we are able to ascertain the ages of all the students who have been selected in our sampling. You will recall that in this example, our aim was to draw a sample from the population of college students, and to compare the sample's mean age with the population mean age. The population mean age comes out to be 15.785 years.

AGE X	No. of Students f	fX
13	6	78
14	61	854
15	270	4050
16	491	7856
17	153	2601
18	15	270
19	4	76
	1000	15785

The population mean age is :

$$\mu = \frac{\sum fx}{\sum f} = \frac{15785}{1000} = 15.785 \ years$$

The above formula is a slightly modified form of the basic formula that you have done ever-since school-days i.e. the mean is equal to the sum of all the observations divided by the total number of observations.
Next, we compute the sample mean age.
Adding the 10 values and dividing by 10, we obtain:
Ages of students selected in the sample (in years):
14, 15, 16, 15, 16, 16, 17, 15, 16, 16
Hence the sample mean age is: 15.6, comparing the sample mean age of 15.6 years with the population mean age of 15.785 years, we note that the difference is really quite slight, and hence the sampling error is equal to

Sampling Error

$$\bar{X} - \mu = 15.6 - 15.785$$
$$= -0.185 \ year$$

And the reason for such a small error is that we have adopted the RANDOM sampling method. The basic advantage of random sampling is that the probability is very high that the sample will be a good representative of the population from which it has been drawn, and any quantity computed from the sample will be a good estimate of the corresponding quantity computed from the population! Actually, a sample is supposed to be a MINIATURE REPLICA of the population. As stated earlier, there are various other types of random sampling.

OTHER TYPES OF RANDOM SAMPLING

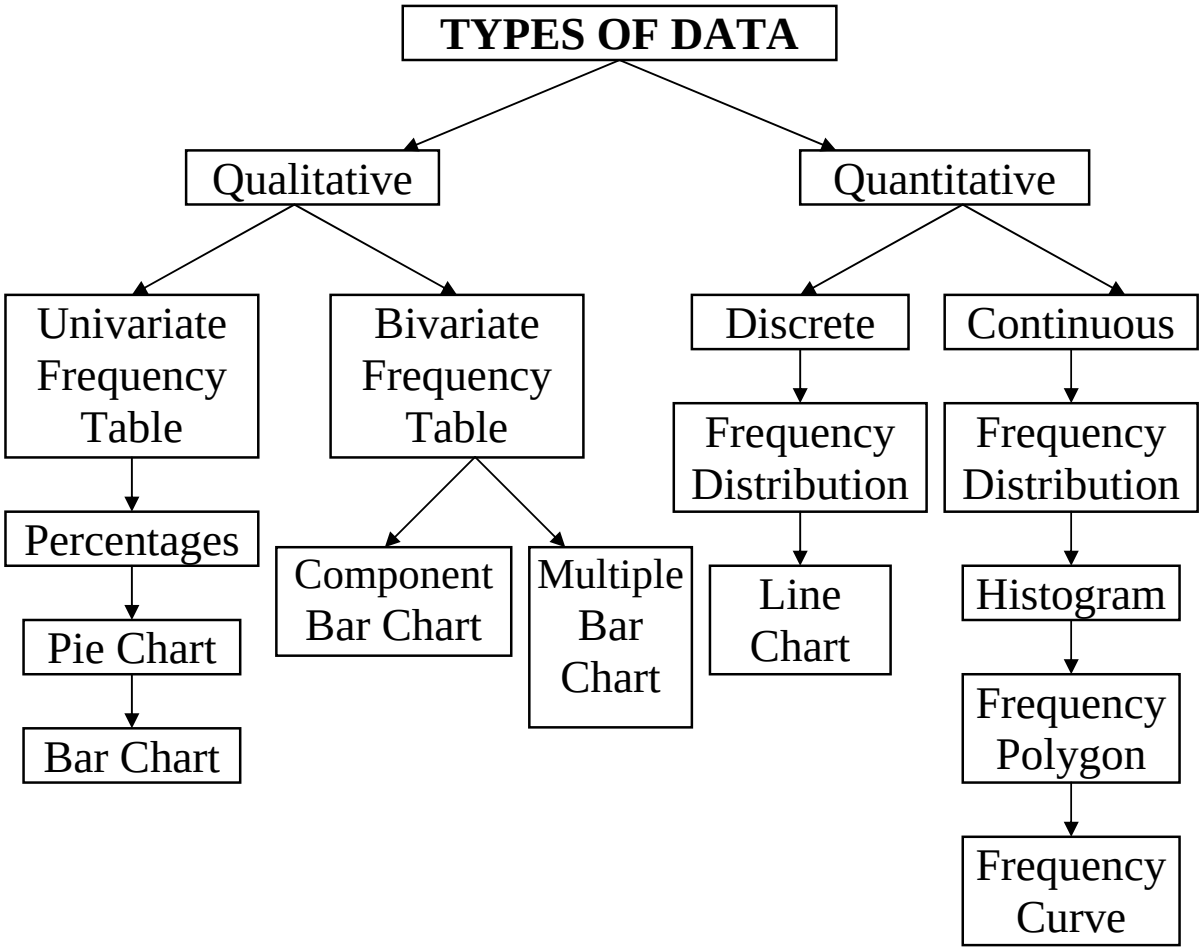
- ·Stratified sampling (if the population is heterogeneous)
- Systematic sampling (practically, more convenient than simple random sampling)
- Cluster sampling (sometimes the sampling units exist in natural clusters)
- Multi-stage sampling

All these designs rest upon random or quasi-random sampling. They are various forms of PROBABILITY sampling --- that in which each sampling unit has a known (but not necessarily equal) probability of being selected.
Because of this knowledge, there exist methods by which the precision and the reliability of the estimates can be calculated OBJECTIVELY.
It should be realized that in practice, several sampling techniques are incorporated into each survey design, and only rarely will simple random sample be used, or a multi-stage design be employed, without stratification.
The point to remember is that whatever method be adopted, care should be exercised at every step so as to make the results as reliable as possible.

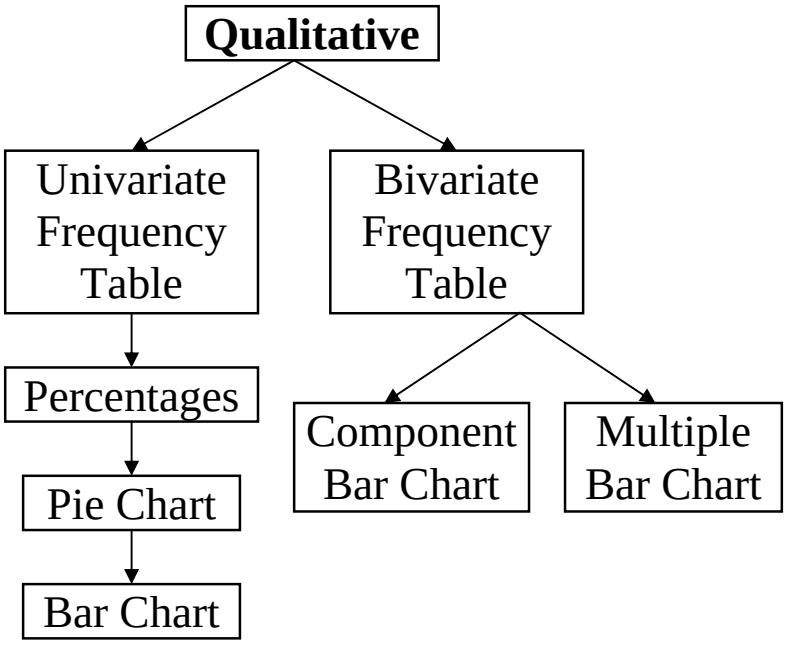
LECTURE NO. 3

- Tabulation
- Simple bar chart
- Component bar chart
- Multiple bar chart
- Pie chart

As indicated in the last lecture, there are two broad categories of data … qualitative data and quantitative data. A variety of methods exist for summarizing and describing these two types of data. The tree-diagram below presents an outline of the various techniques



In today's lecture, we will be dealing with various techniques for summarizing and describing qualitative data.



We will begin with the univariate situation, and will proceed to the bivariate situation.

EXAMPLE

Suppose that we are carrying out a survey of the students of first year studying in a co-educational college of Lahore. Suppose that in all there are 1200 students of first year in this large college. We wish to determine what proportion of these students have come from Urdu medium schools and what proportion has come from English medium schools. So we will interview the students and we will inquire from each one of them about their schooling. As a result, we will obtain a set of data as you can now see on the screen.
We will have an array of observations as follows:

U, U, E, U, E, E, E, U, ……

(U : URDU MEDIUM)
(E : ENGLISH MEDIUM)

Now, the question is what should we do with this data?
Obviously, the first thing that comes to mind is to count the number of students who said "Urdu medium" as well as the number of students who said "English medium". This will result in the following table:

Medium of Institution	No. of Students (f)
Urdu	719
English	481
	1200

The technical term for the numbers given in the second column of this table is "frequency". It means "how frequently something happens?" Out of the 1200 students, 719 stated that they had come from Urdu medium schools. So in this example, the frequency of the first category of responses is 719 whereas the frequency of the second category of responses is 481.

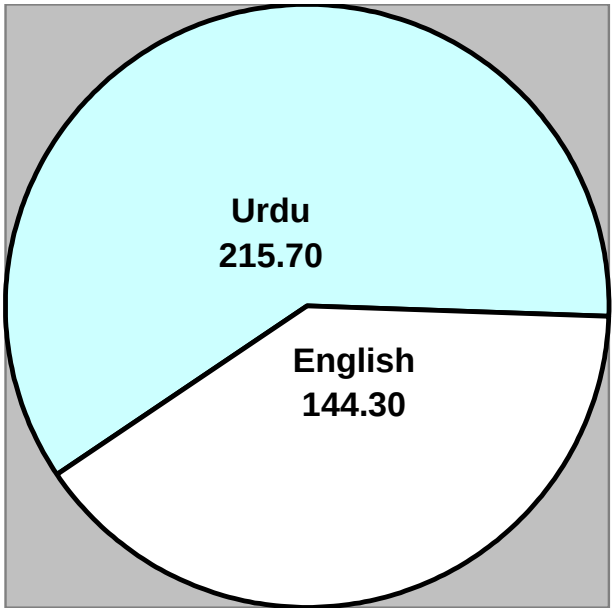
It is evident that this information is not as useful as if we compute the proportion or percentage of students falling in each category. Dividing the cell frequencies by the total frequency and multiplying by 100 we obtain the following:

Medium of Institution	**f**	**%**
Urdu	719	59.9 = 60%
English	481	40.1 = 40%
	1200	

What we have just accomplished is an example of a univariate frequency table pertaining to qualitative data.
Let us now see how we can represent this information in the form of a diagram.
One good way of representing the above information is in the form of a pie chart.
A pie chart consists of a circle which is divided into two or more parts in accordance with the number of distinct categories that we have in our data.
For the example that we have just considered, the circle is divided into two sectors, the larger sector pertaining to students coming from Urdu medium schools and the smaller sector pertaining to students coming from English medium schools.
How do we decide where to cut the circle?
The answer is very simple! All we have to do is to divide the cell frequency by the total frequency and multiply by 360. This process will give us the exact value of the angle at which we should cut the circle.

PIE CHART

Medium of Institution	**f**	**Angle**
Urdu	719	215.7^0
English	481	144.3^0
	1200	



SIMPLE BAR CHART:

The next diagram to be considered is the simple bar chart.

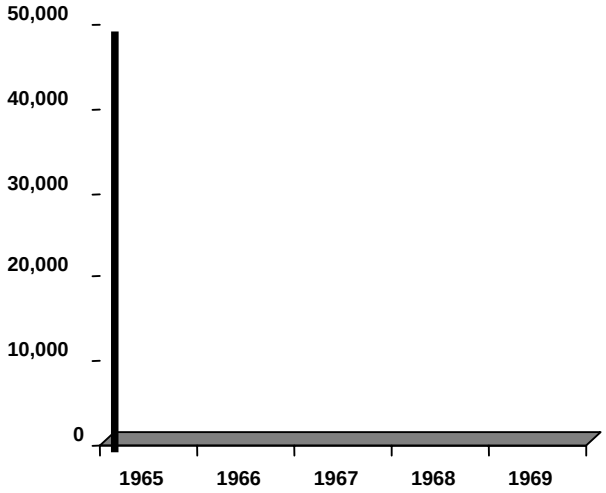
A simple bar chart consists of horizontal or vertical bars of equal width and lengths proportional to values they represent.

As the basis of comparison is one-dimensional, the widths of these bars have no mathematical significance but are taken in order to make the chart look attractive.
Let us consider an example.

Suppose we have available to us information regarding the turnover of a company for 5 years as given in the table below:

Years	1965	1966	1967	1968	1969
Turnover (Rupees)	35,000	42,000	43,500	48,000	48,500

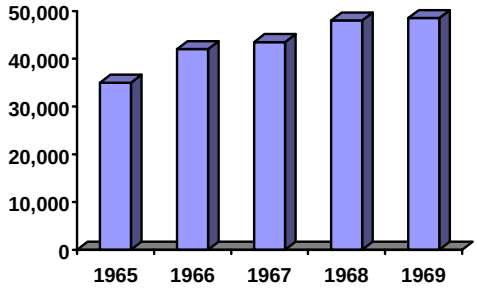
In order to represent the above information in the form of a bar chart, all we have to do is to take the year along the x-axis and construct a scale for turnover along the y-axis.



Next, against each year, we will draw vertical bars of equal width and different heights in accordance with the turn-over figures that we have in our table.

As a result we obtain a simple and attractive diagram as shown below.
When our values do not relate to time, they should be arranged in ascending or descending order before-charting.

BIVARIATE FREQUENCY TABLE



What we have just considered was the univariate situation. In each of the two examples, we were dealing with one single variable. In the example of the first year students of a college, our lone variable of interest was 'medium of schooling'. And in the second example, our one single variable of interest was turnover. Now let us expand the discussion a little, and consider the bivariate situation.

Going back to the example of the first year students, suppose that alongwith the enquiry about the Medium of Institution, you are also recording the sex of the student.
Suppose that our survey results in the following information:

Student No.	Medium	Gender
1	U	F
2	U	M
3	E	M
4	U	F
5	E	M
6	E	F
7	U	M
8	E	M
:	:	:
:	:	:

Now this is a bivariate situation; we have two variables, medium of schooling and sex of the student.
In order to summarize the above information, we will construct a table containing a box head and a stub as shown below:

Sex / Med.	MALE	**Female**	**Total**
Urdu			
English			
Total			

The top row of this kind of a table is known as the boxhead and the first column of the table is known as stub.
Next, we will count the number of students falling in each of the following four categories:

1. Male student coming from an Urdu medium school.
2. Female student coming from an Urdu medium school.
3. Male student coming from an English medium school.
4. Female student coming from an English medium school.

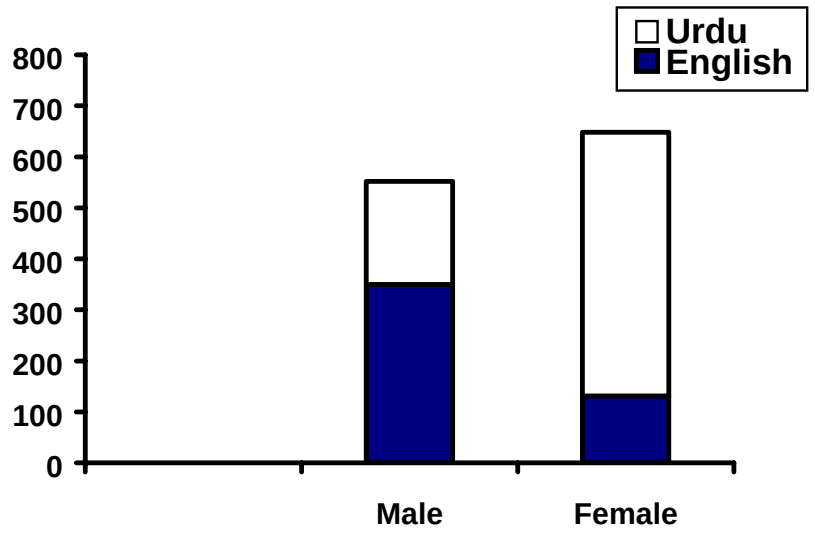
As a result, suppose we obtain the following figures:

Sex / Med.	MALE	**Female**	**Total**
Urdu	202	517	719
English	350	131	481
Total	552	648	1200

What we have just accomplished is an example of a bivariate frequency table pertaining to two qualitative variables.

COMPONENT BAR CHAR:

Let us now consider how we will depict the above information diagrammatically.
This can be accomplished by constructing the component bar chart (also known as the subdivided bar chart) as shown below:



In the above figure, each bar has been divided into two parts. The first bar represents the total number of male students whereas the second bar represents the total number of female students.

As far as the medium of schooling is concerned, the lower part of each bar represents the students coming from English medium schools. Whereas the upper part of each bar represents the students coming from the Urdu medium schools. The advantage of this kind of a diagram is that we are able to ascertain the situation of both the variables at a glance.

We can compare the number of male students in the college with the number of female students, and at the same time we can compare the number of English medium students among the males with the number of English medium students among the females.

MULTIPLE BAR CHARTS

The next diagram to be considered is the multiple bar charts. Let us consider an example.
Suppose we have information regarding the imports and exports of Pakistan for the years 1970-71 to 1974-75 as shown in the table below:

Years	**Imports (Crores of Rs.)**	**Exports (Crores of Rs.)**
1970-71	370	200
1971-72	350	337
1972-73	840	855
1973-74	1438	1016
1974-75	2092	1029

Source: State Bank of Pakistan

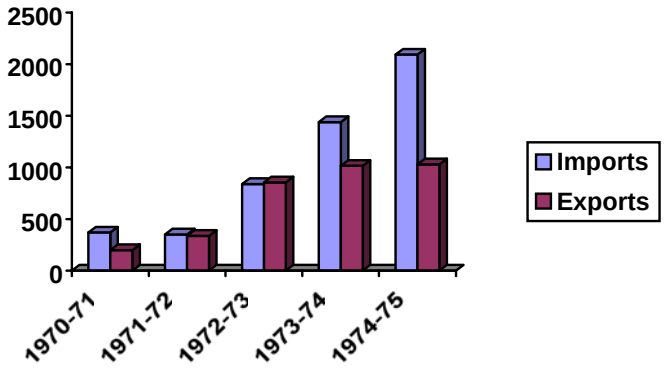
A multiple bar chart is a very useful and effective way of presenting this kind of information.

This kind of a chart consists of a set of grouped bars, the lengths of which are proportionate to the values of our variables, and each of which is shaded or colored differently in order to aid identification. With reference to the above example, we obtain the multiple bar chart shown below:

Multiple Bar Chart Showing

Imports & Exports
of Pakistan 1970-71 to 1974-75

This is a very good device for the comparison of two different kinds of information.



If, in addition to information regarding imports and exports, we also had information regarding production, we could have compared them from year to year by grouping the three bars together.
The question is, what is the basic difference between a component bar chart and a multiple bar chart?
The component bar chart should be used when we have available to us information regarding totals and their components.

For example, the total number of male students out of which some are Urdu medium and some are English medium. The number of Urdu medium male students and the number of English medium male students add up to give us the total number of male students.
On the contrary, in the example of exports and imports, the imports and exports do not add up to give us the totality of some one thing!

LECTURE NO. 4

In THIS Lecture, we will discuss the frequency distribution of a continuous variable & the graphical ways of representing data pertaining to a continuous variable i.e. histogram, frequency polygon and frequency curve.
You will recall that in Lecture No. 1, it was mentioned that a continuous variable takes values over a continuous interval (e.g. a normal Pakistani adult male's height may lie anywhere between 5.25 feet and 6.5 feet).
Hence, in such a situation, the method of constructing a frequency distribution is somewhat different from the one that was discussed in the last lecture.

EXAMPLE:

Suppose that the Environmental Protection Agency of a developed country performs extensive tests on all new car models in order to determine their mileage rating. Suppose that the following 30 measurements are obtained by conducting such tests on a particular new car model.

EPA MILEAGE RATINGS ON 30 CARS (MILES PER GALLON)		
36.3	42.1	44.9
30.1	37.5	32.9
40.5	40.0	40.2
36.2	35.6	35.9
38.5	38.8	38.6
36.3	38.4	40.5
41.0	39.0	37.0
37.0	36.7	37.1
37.1	34.8	33.9
39.9	38.1	39.8

EPA: Environmental Protection Agency

There are a few steps in the construction of a frequency distribution for this type of a variable.

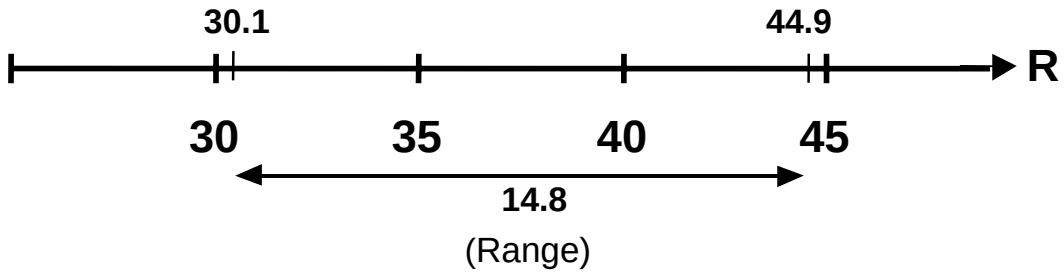
CONSTRUCTION OF A FREQUENCY DISTRIBUTION

Step-1

Identify the smallest and the largest measurements in the data set.
In our example:
Smallest value (X0) = 30.1,
Largest Value (Xm) = 44.9,

Step-2

Find the range which is defined as the difference between the largest value and the smallest value
In our example:
Range = Xm – X0
= 44.9 – 30.1
= 14.8
Let us now look at the graphical picture of what we have just computed.



Step-3

Decide on the number of classes into which the data are to be grouped.

(By classes, we mean small sub-intervals of the total interval which, in this example, is 14.8 units long.)There are no hard and fast rules for this purpose. The decision will depend on the size of the data. When the data are sufficiently large, the number of classes is usually taken between 10 and 20.In this example, suppose that we decide to form 5 classes (as there are only 30 observations).

Step-4

Divide the range by the chosen number of classes in order to obtain the approximate value of the class interval i.e. the width of our classes.

Class interval is usually denoted by h. Hence, in this example

Class interval = h = 14.8 / 5 = 2.96

Rounding the number 2.96, we obtain 3, and hence we take h = 3.

This means that our big interval will be divided into small sub-intervals, each of which will be 3 units long.

Step-5

Decide the lower class limit of the lowest class. Where should we start from?

The answer is that we should start constructing our classes from a number equal to or slightly less than the smallest value in the data.

In this example, smallest value = 30.1

So we may choose the lower class limit of the lowest class to be 30.0.

Step-6

Determine the lower class limits of the successive classes by adding h = 3 successively. Hence, we obtain the following table:

Class Number	Lower Class Limit
1	30.0
2	30.0 + 3 = 33.0
3	33.0 + 3 = 36.0
4	36.0 + 3 = 39.0
5	39.0 + 3 = 42.0

Step-7

Determine the upper class limit of every class. The upper class limit of the highest class should cover the largest value in the data. It should be noted that the upper class limits will also have a difference of h between them. Hence, we obtain the upper class limits that are visible in the third column of the following table.

Class Number	Lower Class Limit	Upper Class Limit
1	30.0	32.9
2	30.0 + 3 = 33.0	32.9 + 3 = 35.9
3	33.0 + 3 = 36.0	35.9 + 3 = 38.9
4	36.0 + 3 = 39.0	38.9 + 3 = 41.9
5	39.0 + 3 = 42.0	41.9 + 3 = 44.9

Hence we obtain the following classes:

Classes
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
39.0 – 41.9
42.0 – 44.9

The question arises: why did we not write 33 instead of 32.9? Why did we not write 36 instead of 35.9? and so on.

The reason is that if we wrote 30 to 33 and then 33 to 36, we would have trouble when tallying our data into these classes. Where should I put the value 33? Should I put it in the first class, or should I put it in the second class? By writing 30.0 to 32.9 and 33.0 to 35.9, we avoid this problem. And the point to be noted is that the class interval is still 3, and not 2.9 as it appears to be. This point will be better understood when we discuss the concept of class boundaries … which will come a little later in today's lecture.

Step-8

After forming the classes, distribute the data into the appropriate classes and find the frequency of each class, in this example:

Class	Tally	Frequency
30.0 – 32.9	||	2
33.0 – 35.9	||||	4
36.0 – 38.9	𝍸 𝍸 ||||	14
39.0 – 41.9	𝍸 |||	8
42.0 – 44.9	||	2
	Total	30

This is a simple example of the frequency distribution of a continuous or, in other words, measurable variable.

CLASS BOUNDARIES:

The true class limits of a class are known as its class boundaries. In this example:

Class Limit	Class Boundaries	Frequency
30.0 – 32.9	29.95 – 32.95	2
33.0 – 35.9	32.95 – 35.95	4
36.0 – 38.9	35.95 – 38.95	14
39.0 – 41.9	38.95 – 41.95	8
42.0 – 44.9	41.95 – 44.95	2
	Total	30

It should be noted that the difference between the upper class boundary and the lower class boundary of any class is equal to the class interval h = 3.
32.95 minus 29.95 is equal to 3, 35.95 minus 32.95 is equal to 3, and so on.
A key point in this entire discussion is that the class boundaries should be taken up to one decimal place more than the given data. In this way, the possibility of an observation falling exactly on the boundary is avoided. (The observed value will either be greater than or less than a particular boundary and hence will conveniently fall in its appropriate class).Next, we consider the concept of the relative frequency distribution and the percentage frequency distribution.
Next, we consider the concept of the relative frequency distribution and the percentage frequency distribution.
This concept has already been discussed when we considered the frequency distribution of a discrete variable.
Dividing each frequency of a frequency distribution by the total number of observations, we obtain the relative frequency distribution.
Multiplying each relative frequency by 100, we obtain the percentage of frequency distribution.
In this way, we obtain the relative frequencies and the percentage frequencies shown below

Class Limit	Frequency	Relative Frequency	%age Frequency
30.0 – 32.9	2	2/30 = 0.067	6.7
33.0 – 35.9	4	4/30 = 0.133	13.3
36.0 – 38.9	14	14/30 = 0.467	4.67
39.0 – 41.9	8	8/30 = 0.267	26.7
42.0 – 44.9	2	2/30 = 0.067	6.7
	30		

The term **'relative frequencies'** simply means that we are considering the frequencies of the various classes relative to the total number of observations.
The advantage of constructing a relative frequency distribution is that comparison is possible between two sets of data having similar classes.
For example, suppose that the Environment Protection Agency perform tests on two car models A and B, and obtains the frequency distributions shown below:

MILEAGE	FREQUENCY	
	Model A	Model B
30.0 – 32.9	2	7
33.0 – 35.9	4	10
36.0 – 38.9	14	16
39.0 – 41.9	8	9
42.0 – 44.9	2	8
	30	50

In order to be able to compare the performance of the two car models, we construct the relative frequency distributions in the percentage form:

MILEAGE	Model A	Model B
30.0-32.9	2/30 x 100 = 6.7	7/50 x 100 = 14
33.0-35.9	4/30 x 100 = 13.3	10/50 x 100 = 20
36.0-38.9	14/30 x 100 = 46.7	16/50 x 100 = 32
39.0-41.9	8/30 x 100 = 26.7	9/50 x 100 = 18
42.0-44.9	2/30 x 100 = 6.7	8/50 x 100 = 16

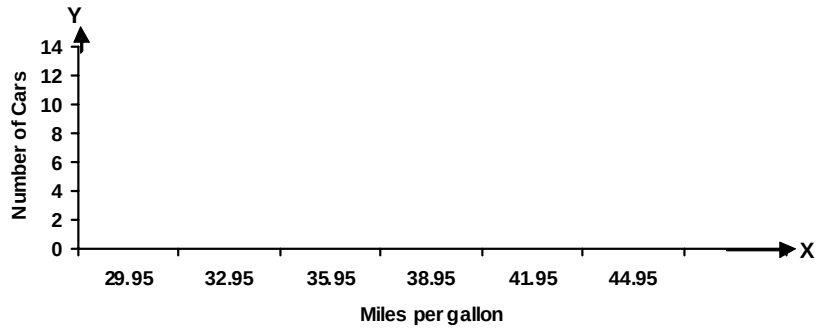
From the table it is clear that whereas 6.7% of the cars of model A fall in the mileage group 42.0 to 44.9, as many as 16% of the cars of model B fall in this group. Other comparisons can similarly be made.
Let us now turn to the visual representation of a continuous frequency distribution. In this context, we will discuss three different types of graphs i.e. the histogram, the frequency polygon, and the frequency curve.

HISTOGRAM:

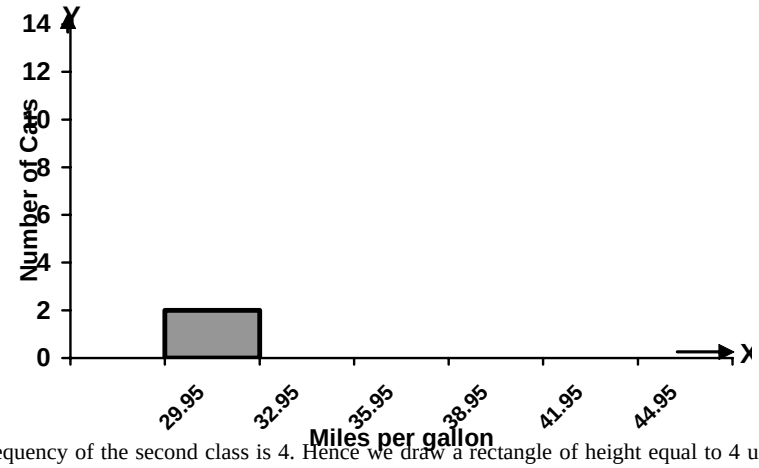
A histogram consists of a set of adjacent rectangles whose bases are marked off by class boundaries along the X-axis, and whose heights are proportional to the frequencies associated with the respective classes.
It will be recalled that, in the last lecture, we were considering the mileage ratings of the cars that had been inspected by the Environment Protection Agency. Our frequency table came out as shown below:

Class Limit	Class Boundaries	Frequency
30.0 – 32.9	29.95 – 32.95	2
33.0 – 35.9	32.95 – 35.95	4
36.0 – 38.9	35.95 – 38.95	14
39.0 – 41.9	38.95 – 41.95	8
42.0 – 44.9	41.95 – 44.95	2
	Total	30

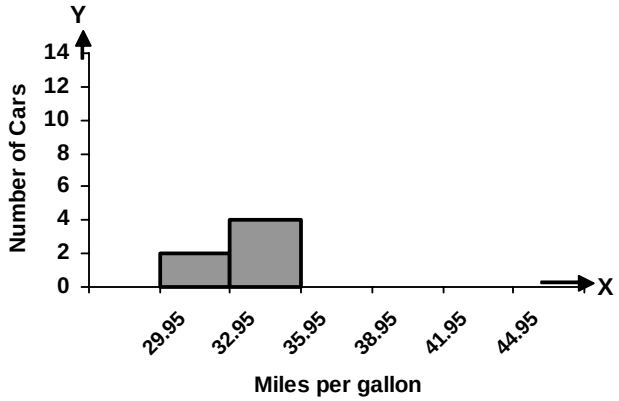
In accordance with the procedure that I just mentioned, we need to take the class boundaries along the X axis
We obtain



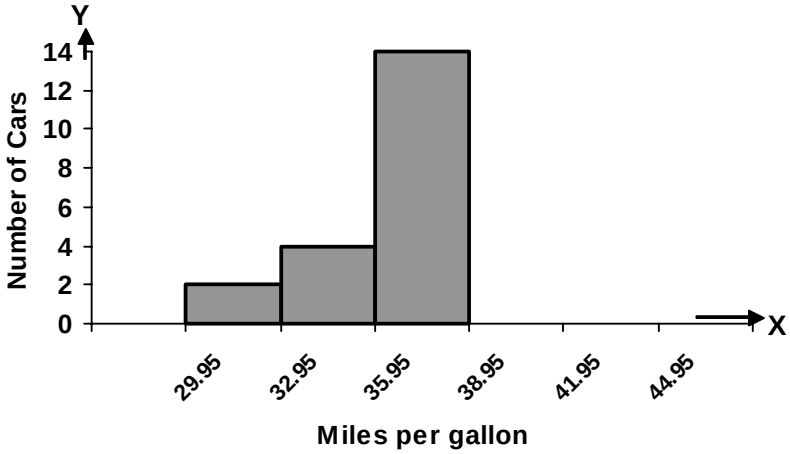
Now, as seen in the frequency table, the frequency of the first class is 2. As such, we will draw a rectangle of height equal to 2 units and obtain the following figure:



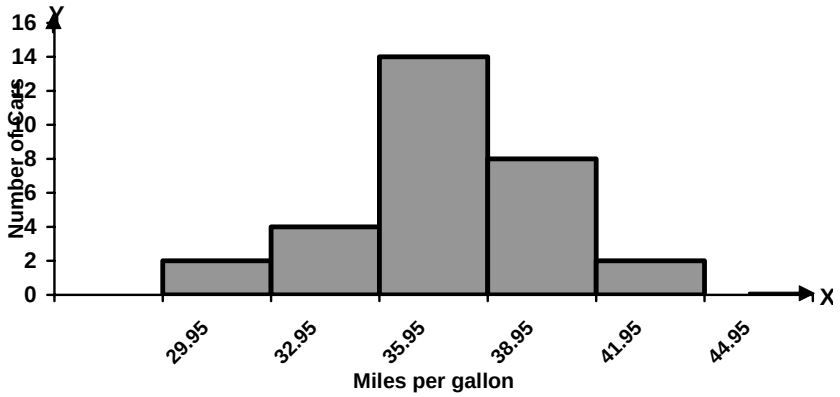
The frequency of the second class is 4. Hence we draw a rectangle of height equal to 4 units against the second class, and thus obtain the following situation:



The frequency of the third class is 14. Hence we draw a rectangle of height equal to 14 units against the third class, and thus obtain the following picture:



Continuing in this fashion, we obtain the following attractive diagram:



This diagram is known as the histogram, and it gives an indication of the overall pattern of our frequency distribution.

FREQUENCY POLYGON:

A frequency polygon is obtained by plotting the class frequencies against the mid-points of the classes, and connecting the points so obtained by straight line segments. In our example of the EPA mileage ratings, the classes are

Class Boundaries
29.95 – 32.95
32.95 – 35.95
35.95 – 38.95
38.95 – 41.95
41.95 – 44.95

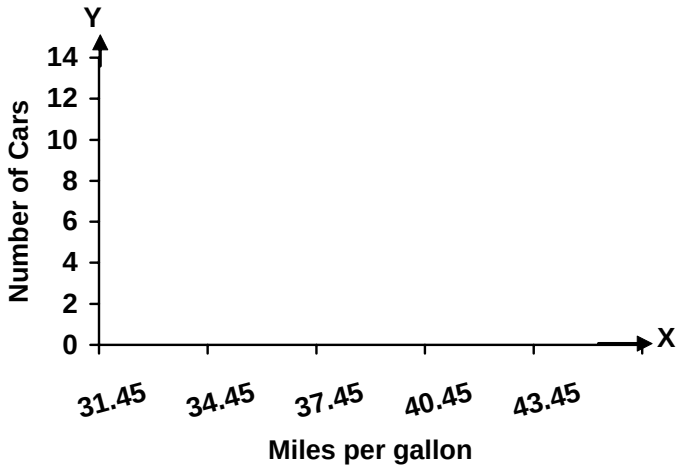
These mid-points are denoted by X.
Now let us add two classes to my frequency table, one class in the very beginning, and one class at the very end.

Class Boundaries	Mid-Point (X)	Frequency (f)
26.95 – 29.95	28.45	
29.95 – 32.95	31.45	2
32.95 – 35.95	34.45	4
35.95 – 38.95	37.45	14
38.95 – 41.95	40.45	8
41.95 – 44.95	43.45	2
44.95 – 47.95	46.45	

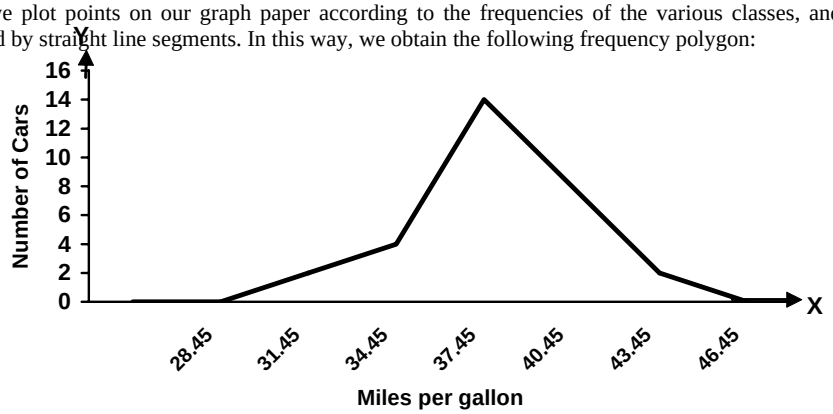
The frequency of each of these two classes is 0, as in our data set, no value falls in these classes.

Class Boundaries	Mid-Point (X)	Frequency (f)
26.95 – 29.95	28.45	0
29.95 – 32.95	31.45	2
32.95 – 35.95	34.45	4
35.95 – 38.95	37.45	14
38.95 – 41.95	40.45	8
41.95 – 44.95	43.45	2
44.95 – 47.95	46.45	0

Now, in order to construct the frequency polygon, the mid-points of the classes are taken along the X-axis and the frequencies along the Y-axis, as shown below:

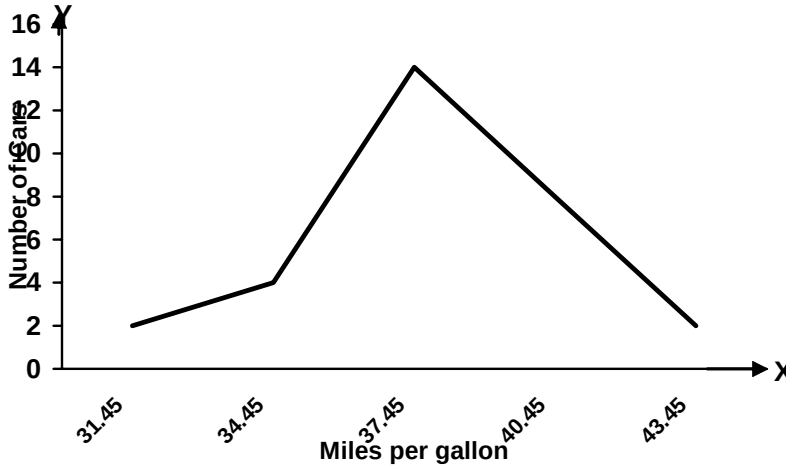


Next, we plot points on our graph paper according to the frequencies of the various classes, and join the points so obtained by straight line segments. In this way, we obtain the following frequency polygon:



It is well-known that the term 'polygon' implies a many-sided closed figure. As such, we want our frequency polygon to be a closed figure. This is exactly the reason why we added two classes to our table, each having zero frequency. Because of the frequency being zero, the line segment touches the X-axis both at the beginning and at the end, and our figure becomes a closed figure.

Had we not carried out this step, our graph would have been as follows:

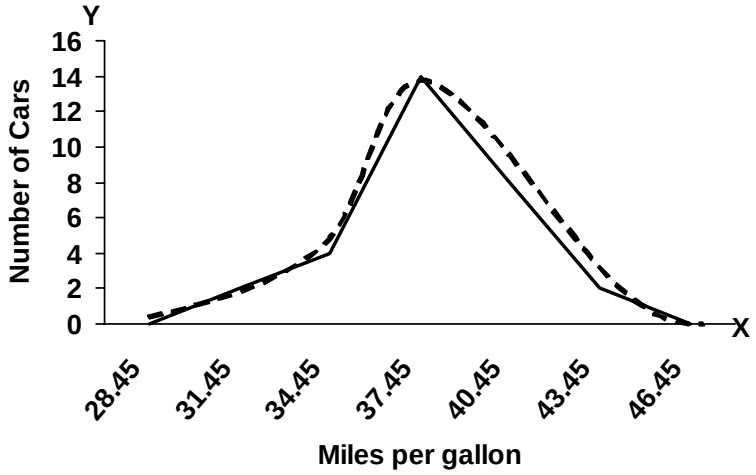


And since this graph is touching the X-axis, hence it cannot be called a frequency polygon (because it is not a closed figure)!

FREQUENCY CURVE:

When the frequency polygon is **smoothed**, we obtain what may be called the frequency curve.

In our example:



LECTURE NO. 5

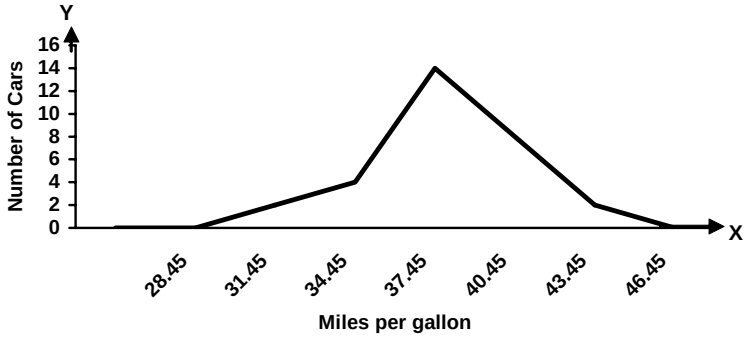
Today's lecture is in continuation with the last lecture, and today we will begin with various types of frequency curves that are encountered in practice. Also, we will discuss the cumulative frequency distribution and cumulative frequency polygon for a continuous variable.

FREQUENCY POLYGON:

A frequency polygon is obtained by plotting the class frequencies against the mid-points of the classes, and connecting the points so obtained by straight line segments. In our example of the EPA mileage ratings, the classes were:

Class Boundaries	Mid-Point (X)	Frequency (f)
26.95 – 29.95	28.45	
29.95 – 32.95	31.45	2
32.95 – 35.95	34.45	4
35.95 – 38.95	37.45	14
38.95 – 41.95	40.45	8
41.95 – 44.95	43.45	2
44.95 – 47.95	46.45	

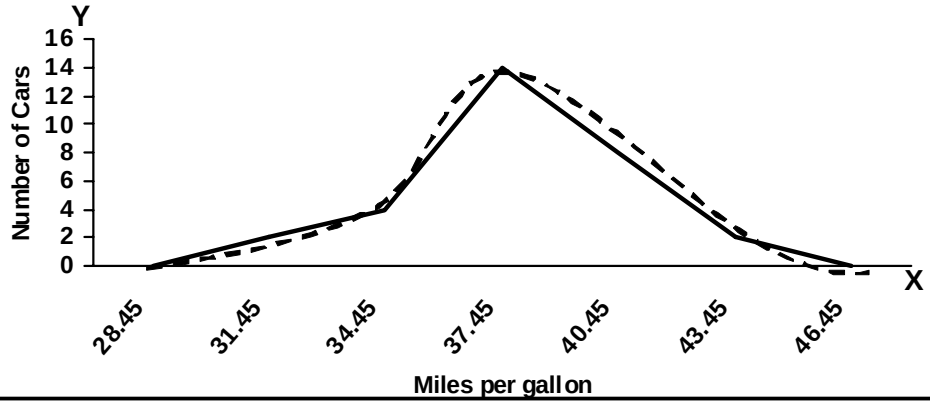
And our frequency polygon came out to be:



Also, it was mentioned that, when the frequency polygon is smoothed, we obtain what may be called the

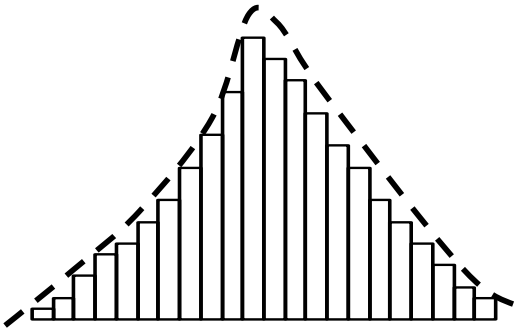
FREQUENCY CURVE

In our example:



In the above figure, the dotted line represents the frequency curve. It should be noted that it is not necessary that our frequency curve must touch all the points. The purpose of the frequency curve is simply to display the overall pattern of the distribution. Hence we draw the curve by the free-hand method, and hence it does not have to touch all the plotted points. It should be realized that the frequency curve is actually a theoretical concept.
If the class interval of a histogram is made very small, and the number of classes is very large, the rectangles of the histogram will be narrow as shown below:

The smaller the class interval and the larger the number of classes, the narrower the rectangles will be. In this way, the histogram approaches a smooth curve as shown below:



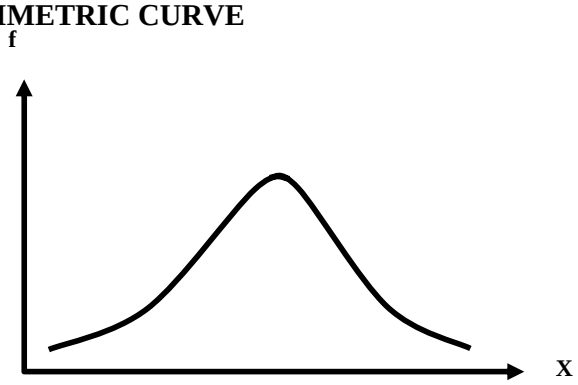
In spite of the fact that the frequency curve is a theoretical concept, it is useful in analyzing real-world problems. The reason is that very close approximations to theoretical curves are often generated in the real world so close that it is quite valid to utilize the properties of various types of mathematical curves in order to aid analysis of the real-world problem at hand.

VARIOUS TYPES OF FREQUENCY CURVES

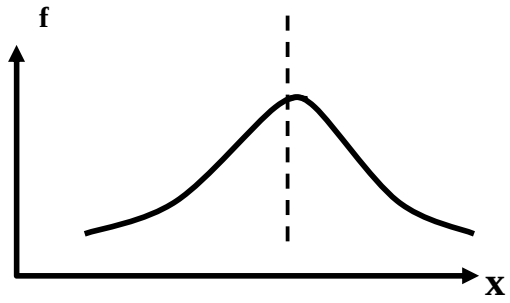
- the symmetrical frequency curve
- the moderately skewed frequency curve
- the extremely skewed frequency curve
- the U-shaped frequency curve

Let us discuss them one by one. First of all, the symmetrical frequency curve is of the following shape:

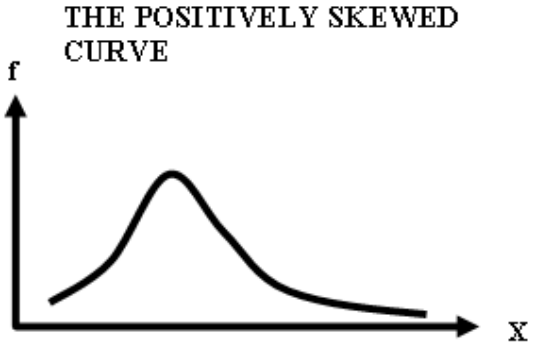
THE SYMMETRIC CURVE



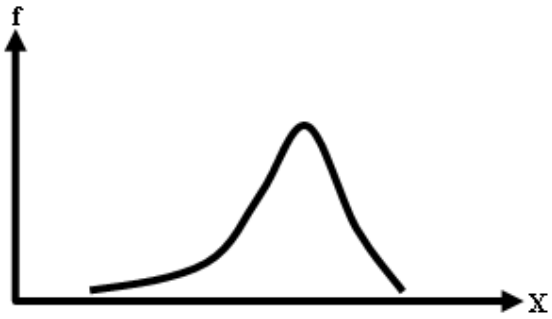
If we place a vertical mirror in the centre of this graph, the left hand side will be the mirror image of the right hand side.



Next, we consider the moderately skewed frequency curve. We have the positively skewed curve and the negatively skewed curve. The positively skewed curve is that one whose right tail is longer than its left tail, as shown below

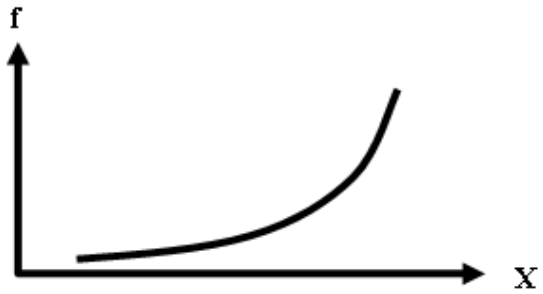


On the other hand, the negatively skewed frequency curve is the one for which the left tail is longer than the right tail.



Both of these that we have just considered are moderately positively and negatively skewed.
Sometimes, we have the extreme case when we obtain the EXTREMELY skewed frequency curve. An extremely negatively skewed curve is of the type shown below:

THE EXTREMELY NEGATIVELY SKEWED (J-SHAPED) CURVE



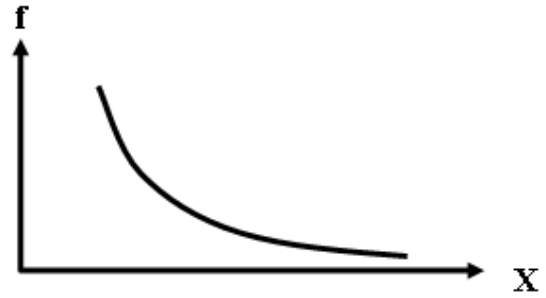
This is the case when the maximum frequency occurs at the end of the frequency table.
For example, if we think of the death rates of adult males of various age groups starting from age 20 and going up to age 79 years, we might obtain something like this:

DEATH RATES BY AGE GROUP

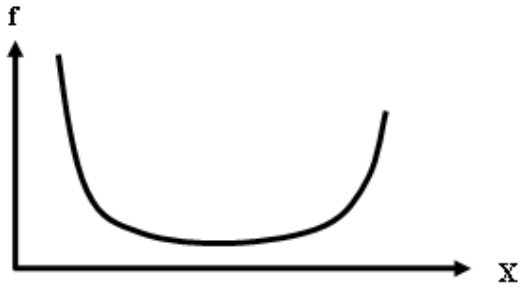
Age Group	No. of deaths per thousand
20 – 29	2.1
30 – 39	4.3
40 – 49	5.7
50 – 59	8.9
60 – 69	12.4
70 – 79	16.7

This will result in a J-shaped distribution similar to the one shown above.
Similarly, the extremely positively skewed distribution is known as the REVERSE J-shaped distribution.

THE EXTREMELY POSITIVELY SKEWED (REVERSE J-SHAPED) CURVE



A relatively LESS frequently encountered frequency distribution is the U-shaped distribution.



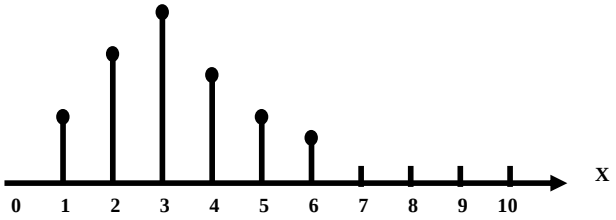
If we consider the example of the death rates not for only the adult population but for the population of ALL the age groups, we will obtain the **U-shaped distribution.**

Out of all these curves, the MOST frequently encountered frequency distribution is the moderately skewed frequency distribution. There are thousands of natural and social phenomena which yield the moderately skewed frequency distribution. Suppose that we walk into a school and collect data of the weights, heights, marks, shoulder-lengths, finger-lengths or any other such variable pertaining to the children of any one class.

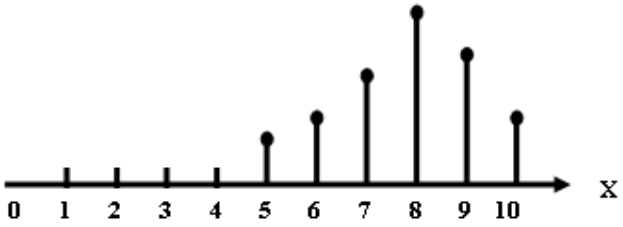
If we construct a frequency distribution of this data, and draw its histogram and its frequency curve, we will find that our data will generate a moderately skewed distribution. Until now, we have discussed the various possible shapes of the frequency distribution of a continuous variable. Similar shapes are possible for the frequency distribution of a discrete variable.

VARIOUS TYPES OF DISCRETE FREQUENCY DISTRIBUTION

I. Positively Skewed Distribution



II. Negatively Skewed Distribution



III. Symmetric Distribution

Let us now consider another aspect of the frequency distribution i.e.

CUMULATIVE FREQUENCY DISTRIBUTION

As in the case of the frequency distribution of a discrete variable, if we start adding the frequencies of our frequency table column-wise, we obtain the column of cumulative frequencies.

In our example, we obtain the cumulative frequencies shown below:

CUMULATIVE FREQUENCY DISTRIBUTION

Class Boundaries	Frequency	Cumulative Frequency
29.95 – 32.95	2	2
32.95 – 35.95	4	2+4 = 6
35.95 – 38.95	14	6+14 = 20
38.95 – 41.95	8	20+8 = 28
41.95 – 44.95	2	28+2 = 30
	30	

In the above table, 2+4 gives 6, 6+14 gives 20, and so on.
The question arises: "What is the purpose of making this column?"
You will recall that, when we were discussing the frequency distribution of a discrete variable, any particular cumulative frequency meant that we were counting the number of observations starting from the very first value of X and going up to THAT particular value of X against which that particular cumulative frequency was falling.

In case of a the distribution of a continuous variable, each of these cumulative frequencies represents the total frequency of a frequency distribution from the lower class boundary of the lowest class to the UPPER class boundary of THAT class whose cumulative frequency we are considering.

In the above table, the total number of cars showing mileage less than 35.95 miles per gallon is 6, the total number of car showing mileage less than 41.95 miles per gallon is 28, etc.

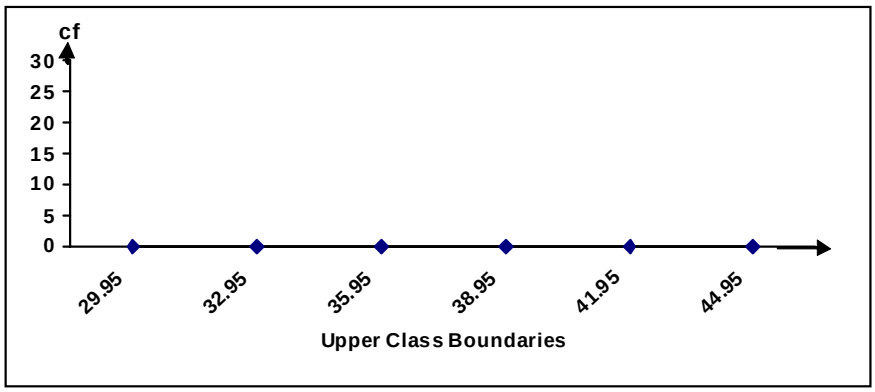
CUMULATIVE FREQUENCY DISTRIBUTION

Class Boundaries	Frequency	Cumulative Frequency
29.95 – 32.95	2	2
32.95 – 35.95	4	2+4 = 6
35.95 – 38.95	14	6+14 = 20
38.95 – 41.95	8	20+8 = 28
41.95 – 44.95	2	28+2 = 30
	30	

Such a cumulative frequency distribution is called a "**less than**" type of a cumulative frequency distribution. The graph of a cumulative frequency distribution is called a

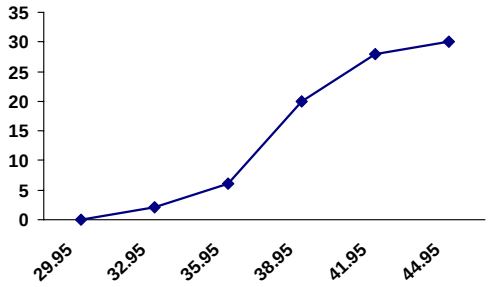
CUMULATIVE FREQUENCY POLYGON or OGIVE

A "less than" type ogive is obtained by marking off the upper class boundaries of the various classes along the X-axis and the cumulative frequencies along the y-axis, as shown below:



The cumulative frequencies are plotted on the graph paper against the upper class boundaries, and the points so obtained are joined by means of straight line segments.
Hence we obtain the cumulative frequency polygon shown below:

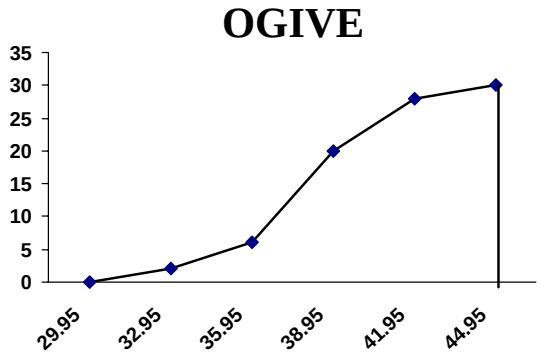
Cumulative Frequency Polygon or OGIVE



It should be noted that this graph is touching the X-Axis on the left-hand side. This is achieved by ADDING a class having zero frequency in the beginning of our frequency distribution, as shown below:

Class Boundaries	Frequency	Cumulative Frequency
26.95 – 29.95	0	0
29.95 – 32.95	2	0+2 = 2
32.95 – 35.95	4	2+4 = 6
35.95 – 38.95	14	6+14 = 20
38.95 – 41.95	8	20+8 = 28
41.95 – 44.95	2	28+2 = 30
	30	

Since the frequency of the first class is zero, hence the cumulative frequency of the first class will also be zero, and hence, automatically, the cumulative frequency polygon will touch the X-Axis from the left hand side. If we want our cumulative frequency polygon to be closed from the right-hand side also, we can achieve this by connecting the last point on our graph paper with the X-axis by means of a vertical line, as shown below:



In the example of EPA mileage ratings, all the data-values were correct to one decimal place. Let us now consider another example:

EXAMPLE:

For a sample of 40 pizza products, the following data represent cost of a slice in dollars (S Cost).

PRODUCT	S cost
Pizza Hut Hand Tossed	1.51
Domino's Deep Dish	1.53
Pizza Hut Pan Pizza	1.51
Domino's Hand Tossed	1.90
Little Caesars Pan! Pizza!	1.23

PRODUCT	S Cost
Boboli crust with Boboli sauce	1.00
Jack's Super Cheese	0.69
Pappalo's Three Cheese	0.75
Tombstone Original Extra Cheese	0.81
Master Choice Gourmet Four Cheese	0.90
Celeste Pizza For One	0.92
Totino's Party	0.64
The New Weight Watchers Extra Cheese	1.54
Jeno's Crisp'N Tasty	0.72
Stouffer's French Bread 2-Cheese	1.15

PRODUCT	S Cost
Ellio's 9-slice	0.52
Kroger	0.72
Healthy Choice French Bread	1.50
Lean Cuisine French Bread	1.49
DiGiorno Rising Crust	0.87
Tombstone Special Order	0.81
Pappalo's	0.73
Jack's New More Cheese!	0.64
Tombstone Original	0.77
Red Baron Premium	0.80

PRODUCT	Scost
Tony's Italian Style Pastry Cruse	0.83
Red Baron Deep Dish Singles	1.13
Totino's Party	0.62
The New Weight Watchers	1.52
Jeno's Crisp'N Tasty	0.71
Stouffer's French Bread	1.14
Celeste Pizza For One	1.11
Tombstone For One French Bread	1.11
Healthy Choice French Bread	1.46
Lean Cuisine French Bread	1.71

PRODUCT	Scost
Little Caesars Pizza! Pizza!	1.28
Pizza Hut Stuffed Crust	1.23
DiGiorno Rising Crust Four Cheese	0.90
Tombstone Speical Order Four Cheese	0.85
Red Baron Premium 4-Cheese	0.80

Source: "Pizza," Copyright 1997 by Consumers Union of United States, Inc., Yonkers, N.Y. 10703.

Example taken from
"Business Statistics – A First Course" by Mark L. Berenson & David M. Levine (International Edition), Prentice-Hall International, Inc., Copyright © 1998.

In order to construct the frequency distribution of the above data, the first thing to note is that, in this example, all our data values are correct to two decimal places. As such, we should construct the class limits correct to TWO decimal places, and the class boundaries correct to three decimal places.

As in the last example, first of all, let us find the maximum and the minimum values in our data, and compute the RANGE.

Minimum value X0 = 0.52
Maximum value Xm = 1.90
Hence:

Range = 1.90 - 0.52 = 1.38

Desired number of classes = 8
Hence:

Class interval h = RANGE/No. of classes

= 1.38 / 8 = 0.1725 ~ 0.20

Lower limit of the first class = 0.51

Hence, our successive class limits come out to be:

Class Limits
0.51 – 0.70
0.71 – 0.90
0.91 – 1.10
1.11 – 1.30
1.31 – 1.50
1.51 – 1.70
1.71 – 1.90

Stretching the class limits to the left and to the right, we obtain class boundaries as shown below:

Class Limits	Class Boundaries
0.51 – 0.70	0.505 – 0.705
0.71 – 0.90	0.705 – 0.905
0.91 – 1.10	0.905 – 1.105
1.11 – 1.30	1.105 – 1.305
1.31 – 1.50	1.305 – 1.505
1.51 – 1.70	1.505 – 1.705
1.71 – 1.90	1.705 – 1.905

By tallying the data-values in the appropriate classes, we will obtain a frequency distribution similar to the one that we obtained in the examples of the EPA mileage ratings.
By constructing the histogram of this data-set, we will be able to decide whether our distribution is symmetric, positively skewed or negatively skewed. This may please be attempted as an exercise.

LECTURE NO. 6

This plot was introduced by the famous statistician *John Tukey in 1977*. A frequency table has the disadvantage that the identity of individual observations is lost in grouping process. To overcome this drawback, John Tukey (1977) introduced this particular technique (**known as the Stem-and-Leaf Display**).

This technique offers a quick and novel way for simultaneously sorting and displaying data sets where each number in the data set is divided into two parts, a Stem and a Leaf.

A stem is the leading digit(s) of each number and is used in sorting, while a leaf is the rest of the number or the trailing digit(s) and shown in display. A vertical line separates the leaf (or leaves) from the stem.

For example, the number 243 could be split in two ways:

Leading Digit	Trailing Digits	OR	Leading Digit	Trailing Digit
2	43		24	3
Stem	Leaf		Stem	Leaf

How do we construct a stem and leaf display when we have a whole set of values? This is explained by way of the following example:

EXAMPLE:

The ages of 30 patients admitted to a certain hospital during a particular week were as follows:

48, 31, 54, 37, 18, 64, 61, 43, 40, 71, 51, 12, 52, 65, 53, 42, 39, 62, 74, 48, 29, 67, 30, 49, 68, 35, 57, 26, 27, 58.

Construct a stem-and-leaf display from the data and list the data in an array.

A scan of the data indicates that the observations range (in age) from 12 to 74. We use the first (or leading) digit as the stem and the second (or trailing) digit as the leaf. The first observation is 48, which has a stem of 4 and a leaf of 8, the second a stem of 3 and a leaf of 1, etc. Placing the leaves in the order in which they APPEAR in the data, we get the stem-and-leaf display as shown below:

Stem (Leading Digit)	Leaf (Trailing Digit)
1	8 2
2	9 6 7
3	1 7 9 0 5
4	8 3 0 2 8 9
5	4 1 2 3 7 8
6	4 1 5 2 7 8
7	1 4

But it is a common practice to ARRANGE the trailing digits in each row from smallest to highest. In this example, in order to obtain an array, we associate the leaves in order of size with the stems as shown below:

DATA IN THE FORM OF AN ARRAY (in ascending order):

12, 18, 26, 27, 29, 30, 31, 35, 37, 39, 40, 42, 43, 48, 48, 49, 51, 52, 53, 54, 57, 58, 61, 62, 64, 65, 67, 68, 71, 74.

Hence we obtain the stem and leaf plot shown below:

STEM AND LEAF DISPLAY

Stem (Leading Digit)	Leaf (Trailing Digit)
1	2 8
2	6 7 9
3	0 1 5 7 9
4	0 2 3 8 8 9
5	1 2 3 4 7 8
6	1 2 4 5 7 8
7	1 4

The stem-and-leaf table provides a useful description of the data set and, if we so desire, can easily be converted to a frequency table. In this example, the frequency of the class 10-19 is 2, the frequency of the class 20-29 is 3, and the frequency of the class 30-39 is 5, and so on.

STEM AND LEAF DISPLAY

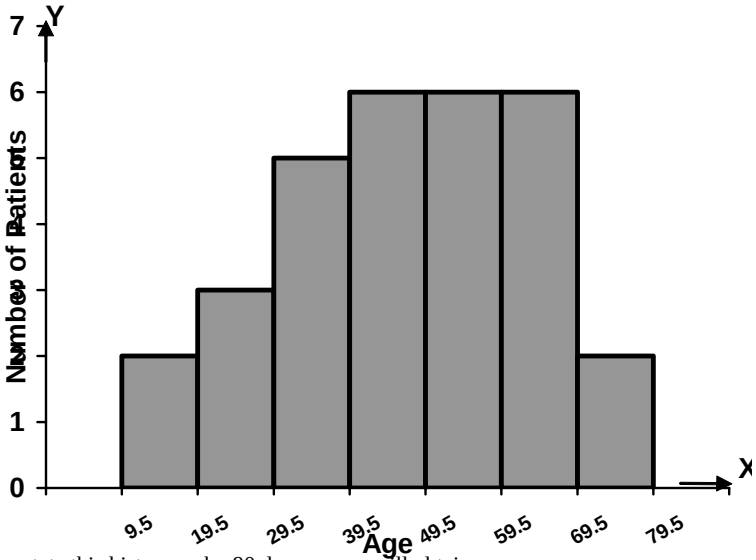
Stem (Leading Digit)	Leaf (Trailing Digit)
1	2 8
2	6 7 9
3	0 1 5 7 9
4	0 2 3 8 8 9
5	1 2 3 4 7 8
6	1 2 4 5 7 8
7	1 4

Hence, this stem and leaf plot conveniently converts into the frequency distribution shown below:

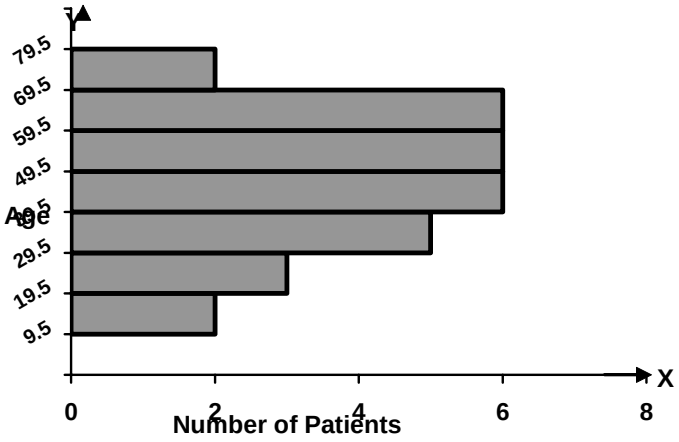
FREQUENCY DISTRIBUTION

Class Limits	Class Boundaries	Tally Marks	Frequency
10 – 19	9.5 – 19.5	//	2
20 – 29	19.5 – 29.5	///	3
30 – 39	29.5 – 39.5	~~////~~	5
40 – 49	39.5 – 49.5	~~////~~ /	6
50 – 59	49.5 – 59.5	~~////~~ /	6
60 – 69	59.5 – 69.5	~~////~~ /	6
70 - 79	69.5 – 79.5	//	2

Converting this frequency **distribution into a histogram,** we obtain:

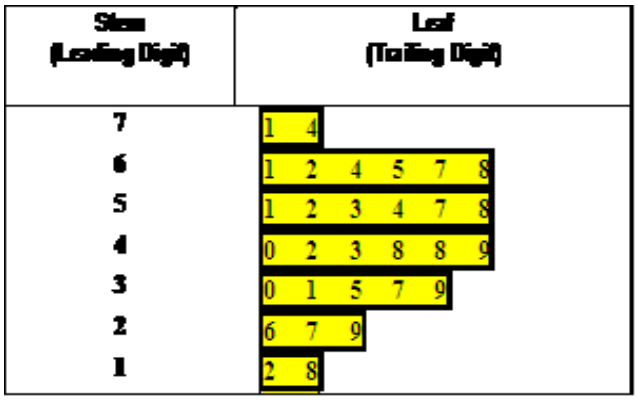


If we rotate this histogram by 90 degrees, we will obtain:



Let us re-consider the stem and leaf plot that we obtained a short while ago.

STEM AND LEAF DISPLAY



It is noteworthy that the shape of the stem and leaf display is exactly like the shape of our histogram. Let us now consider another example.

EXAMPLE

Construct a stem-and-leaf display for the data of mean annual death rates per thousand at ages 20-65 given below:
7.5, 8.2, 7.2, 8.9, 7.8, 5.4, 9.4, 9.9, 10.9, 10.8, 7.4, 9.7, 11.6, 12.6, 5.0, 10.2, 9.2, 12.0, 9.9, 7.3, 7.3, 8.4, 10.3, 10.1, 10.0, 11.1, 6.5, 12.5, 7.8, 6.5, 8.7, 9.3, 12.4, 10.6, 9.1, 9.7, 9.3, 6.2, 10.3, 6.6, 7.4, 8.6, 7.7, 9.4, 7.7, 12.8, 8.7, 5.5, 8.6, 9.6, 11.9, 10.4, 7.8, 7.6, 12.1, 4.6, 14.0, 8.1, 11.4, 10.6, 11.6, 10.4, 8.1, 4.6, 6.6, 12.8, 6.8, 7.1, 6.6, 8.8, 8.8, 10.7, 10.8, 6.0, 7.9, 7.3, 9.3, 9.3, 8.9, 10.1, 3.9, 6.0, 6.9, 9.0, 8.8, 9.4, 11.4, 10.9

Using the decimal part in each number as the leaf and the rest of the digits as the stem, we get the ordered stem-and-leaf display shown below:

STEM AND LEAF DISPLAY

Stem	Leaf
3	9
4	6 6
5	0 4 5
6	0 0 2 2 5 5 6 6 6 8 9
7	1 3 3 3 4 4 5 6 7 7 8 8 8 9
8	1 1 2 4 6 6 7 7 8 8 8 9 9
9	0 1 2 3 3 3 3 4 4 4 6 7 7 9 9
10	0 1 1 2 3 3 4 4 6 6 7 8 8 9 9
11	1 4 4 6 6 9
12	0 1 4 5 6 8 8
14	0

EXERCISE

- The above data may be converted into a stem and leaf plot (so as to verify that the one shown above is correct).
- Various variations of the stem and leaf display may be studied on your own.

The next concept that we are going to consider is the concept of the central tendency of a data-set. In this context, the first thing to note is that in any data-based study, our data is always going to be variable, and hence, first of all, we will need to describe the data that is available to us.

DESCRIPTION OF VARIABLE DATA:

Regarding any statistical enquiry, primarily we need some means of describing the situation with which we are confronted. A concise numerical description is often preferable to a lengthy tabulation, and if this form of description also enables us to form a mental image of the data and interpret its significance, so much the better.

MEASURES OF CENTRAL TENDENCY AND MEASURES OF DISPERSION

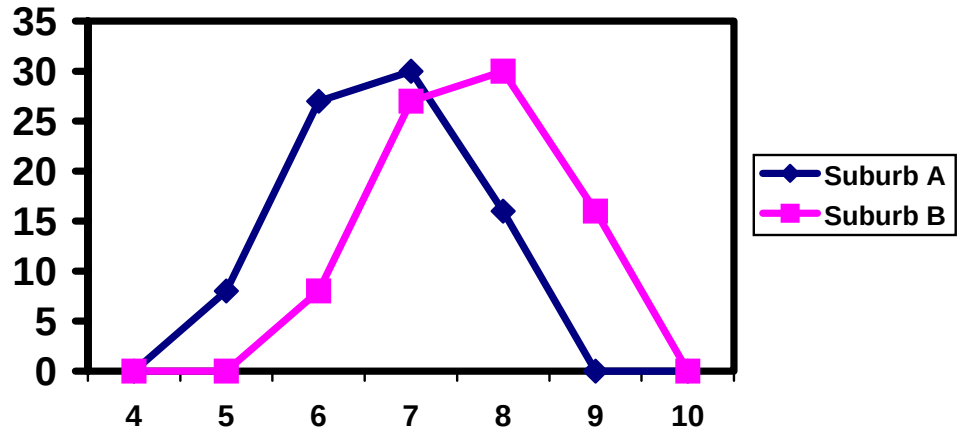
- Averages enable us to measure the central tendency of variable data
- Measures of dispersion enable us to measure its variability.

AVERAGES (I.E. MEASURES OF CENTRAL TENDENCY)

An average is a single value which is intended to represent a set of data or a distribution as a whole. It is more or less CENTRAL value ROUND which the observations in the set of data or distribution usually tend to cluster.
As a measure of central tendency (i.e. an average) indicates the location or general position of the distribution on the X-axis, it is also known as a measure of location or position.
Let us consider an example: Suppose that we have the following two frequency distributions:

EXAMPLE:

Looking at these two frequency distributions, we should ask ourselves what exactly is the distinguishing feature?
If we draw the frequency polygon of the two frequency distributions, we obtain



Inspection of these frequency polygons shows that they have exactly the same shape. It is their position relative to the horizontal axis (X-axis) which distinguishes them.
If we compute the mean number of rooms per house for each of the two suburbs, we will find that the average number of rooms per house in A is 6.67 while in B it is 7.67.
This difference of 1 is equivalent to the difference in position of the two frequency polygons.

Our interpretation of the above situation would be that there are LARGER houses in suburb B than in suburb A, to the extent that there are on the *average.*

VARIOUS TYPES OF AVERAGES:

There are several types of averages each of which has a use in specifically defined circumstances.
The most common types of averages are:

- The arithmetic mean,
- The geometric mean,

- The harmonic mean
- The median, and
- The mode

The ***Arithmetic, Geometric and Harmonic means*** are averages that are mathematical in character, and give an indication of the magnitude of the observed values.
The ***Median*** indicates the middle position while the mode provides information about the most frequent value in the distribution or the set of data. THE MODE:
The ***Mode*** is defined as that value which occurs most frequently in a set of data i.e. it indicates the most common result.

EXAMPLE:

Suppose that the marks of eight students in a particular test are as follows:
2, 7, 9, 5, 8, 9, 10, 9
Obviously, the most common mark is 9. In other words, Mode = 9.

MODE IN CASE OF RAW DATA PERTAINING TO A CONTINUOUS VARIABLE

In case of a set of values (pertaining to a continuous variable) that have not been grouped into a frequency distribution (i.e. in case of raw data pertaining to a continuous variable), the mode is obtained by counting the number of times each value occurs.

EXAMPLE:

Suppose that the government of a country collected data regarding the percentages of revenues spent on Research and Development by 49 different companies, and obtained the following figures:

Percentage of Revenues Spent on Research and Development

Company	Percentage	Company	Percentage
1	13.5	14	9.5
2	8.4	15	8.1
3	10.5	16	13.5
4	9.0	17	9.9
5	9.2	18	6.9
6	9.7	19	7.5
7	6.6	20	11.1
8	10.6	21	8.2
9	10.1	22	8.0
10	7.1	23	7.7
11	8.0	24	7.4
12	7.9	25	6.5
13	6.8	26	9.5

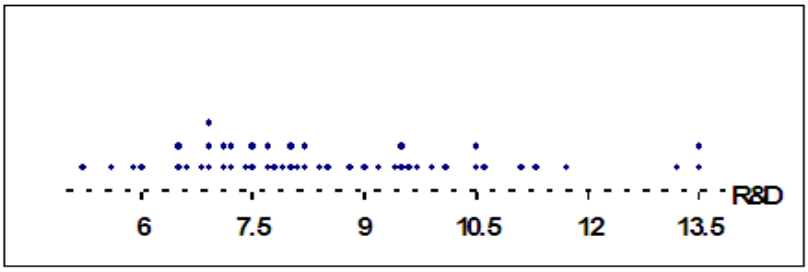
Company	Percentage	Company	Percentage
27	8.2	39	6.5
28	6.9	40	7.5
29	7.2	41	7.1
30	8.2	42	13.2
31	9.6	43	7.7
32	7.2	44	5.9
33	8.8	45	5.2
34	11.3	46	5.6
35	8.5	47	11.7
36	9.4	48	6.0
37	10.5	49	7.8
38	6.9		

We can represent this data by means of a plot that is called dot plot.

DOT PLOT:

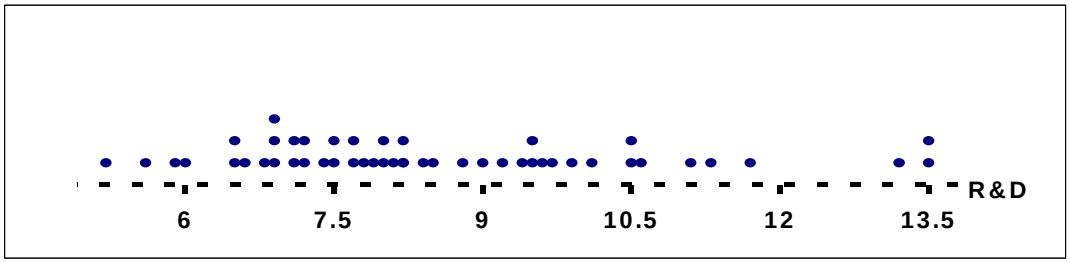
The horizontal axis of a dot plot contains a scale for the quantitative variable that we want to represent. The numerical value of each measurement in the data set is located on the horizontal scale by a dot. When data values repeat, the dots are placed above one another, forming a pile at that particular numerical location.
In this example

Dot Plot

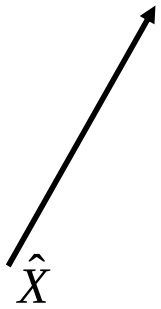


The above material has been taken from "Statistics for Business and Economics" by James T. McClave, P. George Benson and Terry Sincich (Seventh Edition), © 1998 Prentice-Hal International, **Inc.**

As is obvious from the above diagram, the value 6.9 occurs 3 times whereas all the other values are occurring either once or twice.
Hence the modal value is 6.9.



Dot Plot



= 6.9

Also, this dot plot shows that

- almost all of the R&D percentages are falling between 6% and 12%,
- most of the percentages are falling between 7% and 9%.

THE MODE IN CASE OF A DISCRETE FREQUENCY DISTRIBUTION:

In case of a discrete frequency distribution, identification of the mode is immediate; one simply finds that value which has the highest frequency.

EXAMPLE:

An airline found the following numbers of passengers in fifty flights of a forty-seated plane

No. of Passengers X	No. of Flights f
28	1
33	1
34	2
35	3
36	5
37	7
38	10
39	13
40	8
Total	**50**

Highest Frequency fm = 13
Occurs against the X value 39
Hence: Mode = x= 39
The mode is obviously 39 passengers and the company should be quite satisfied that a 40 seater is the correct-size aircraft for this particular route.

THE MODE IN CASE OF THE *FREQUENCY DISTRIBUTION* OF A CONTINUOUS VARIABLE

In case of grouped data, the modal group is easily recognizable (the one that has the highest frequency).
At what point within the modal group does the mode lie?
The answer is contained in the following formula:

Mode:

$$\hat{X} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

Where

l = lower class boundary of the modal class,

f_m = frequency of the modal class,

f_1 = frequency of the class preceding the modal class

f_2 = frequency of the class following modal

class

h = length of class interval of the modal class

Going back to the example of EPA mileage ratings, we have:

EPA MILEAGE RATINGS

Mileage Rating	Class Boundaries	No. of Cars
30.0 – 32.9	29.95 – 32.95	2
33.0 – 35.9	32.95 – 35.95	$4 = f_1$
36.0 – 38.9	35.95 – 38.95	$14 = f_m$
39.0 – 41.9	38.95 – 41.95	$8 = f_2$
42.0 – 44.9	41.95 – 44.95	2

It is evident that the third class is the modal class. The mode lies somewhere between 35.95 and 38.95.
In order to apply the formula for the mode, we note that fm = 14, f1 = 4 and f2 = 8.
Hence we obtain:

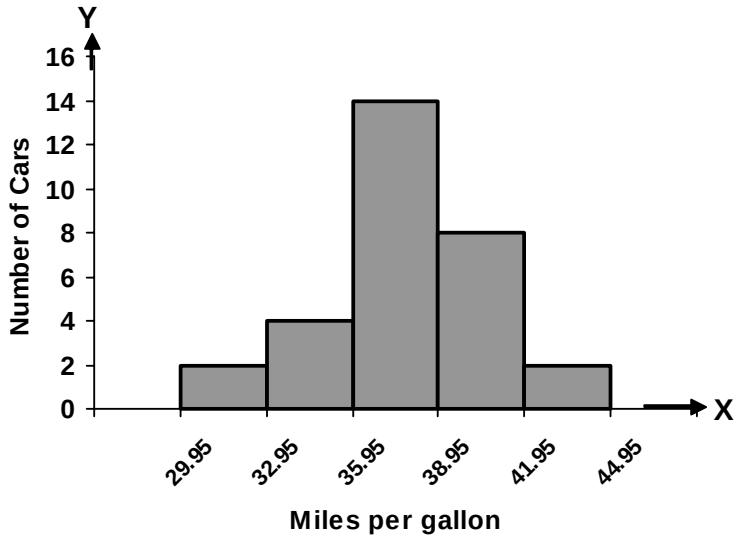
$$\hat{X} = 35.95 + \frac{14 - 4}{(14 - 4) + (14 - 8)} \times 3$$

$$= 35.95 + \frac{10}{10 + 6} \times 3$$

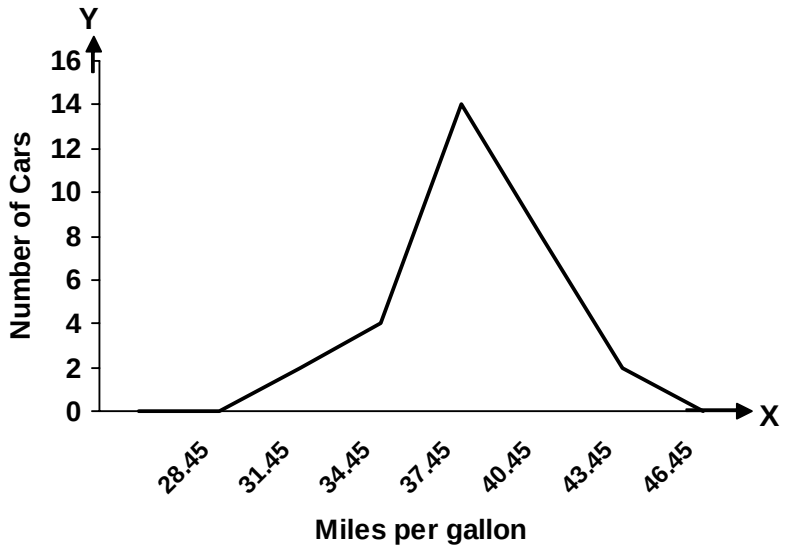
$$= 35.95 + 1.875$$

$$= 37.825$$

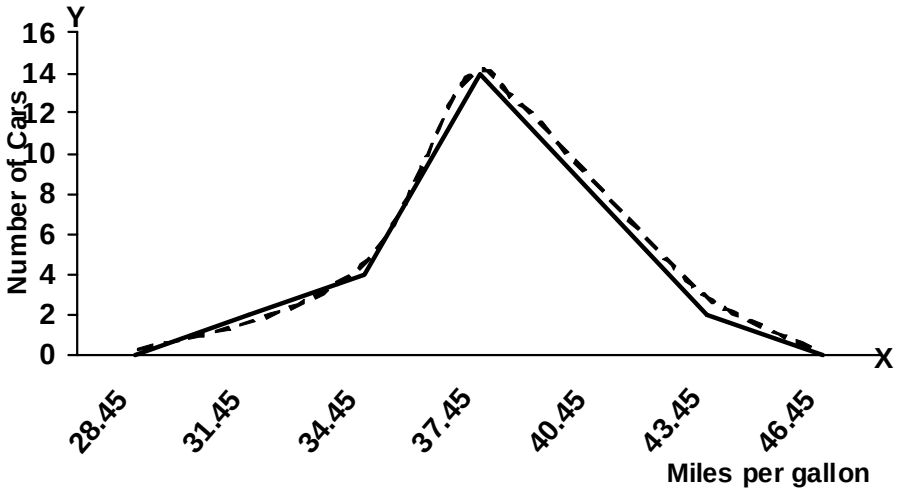
Let us now perceive the mode by considering the graphical representation of our frequency distribution. You will recall that, for the example of EPA Mileage Ratings, the histogram was as shown below:



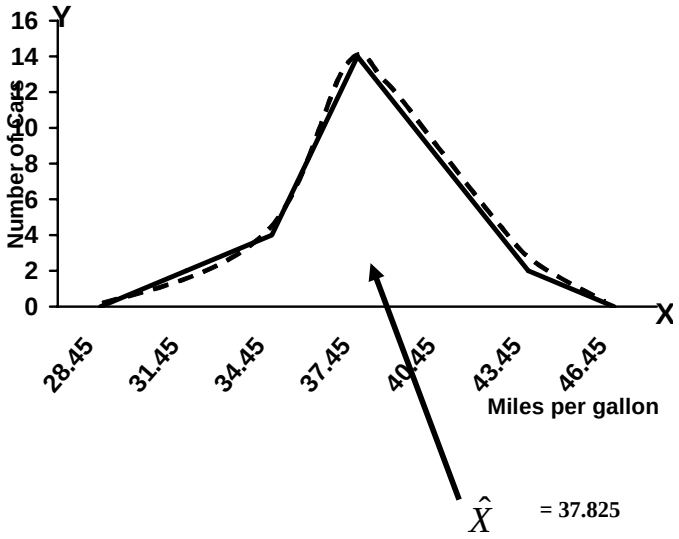
The frequency polygon of the same distribution was:



And the frequency curve was as indicated by the dotted line in the following figure:



In this example, the mode is 37.825, and if we locate this value on the X-axis, we obtain the following picture:



Since, in most of the situations the mode exists somewhere in the middle of our data-values, hence it is thought of as a measure of central tendency.

LECTURE NO. 7

In general, it was noted that, for most of the frequency distributions, the mode lies somewhere in the middle of our frequency distribution, and hence is eligible to be called a measure of central tendency.
The mode has some very desirable properties.

DESIRABLE PROPERTIES OF THE MODE:

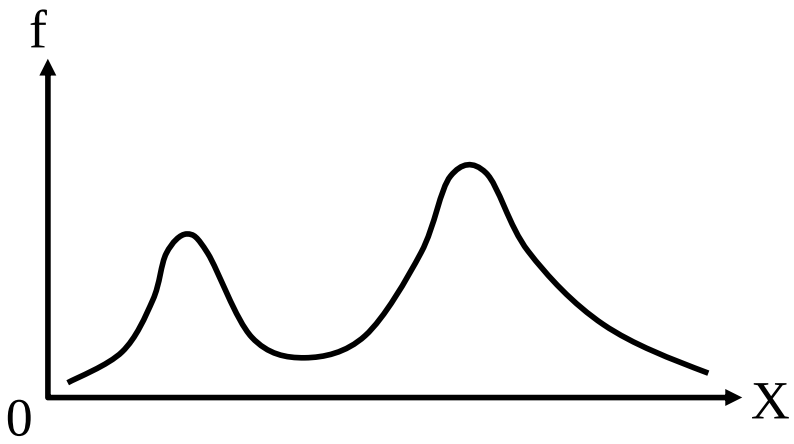
- The mode is easily understood and easily ascertained in case of a discrete frequency distribution.
- It is not affected by a few very high or low values.

The question arises, "When should we use the mode?" The answer to this question is that the mode is a valuable concept in certain situations such as the one described below:

Suppose the manager of a men's clothing store is asked about the average size of hats sold. He will probably think not of the arithmetic or geometric mean size, or indeed the median size. Instead, he will in all likelihood quote that particular size which is sold ***most often.*** This average is of far more use to him as a businessman than the arithmetic mean, geometric mean or the median. The modal size of all clothing is the size which the businessman must stock in the greatest quantity and variety in comparison with other sizes. Indeed, in most inventory (stock level) problems, one needs the mode more often than any other measure of central tendency. It should be noted that in some situations there may be no mode in a simple series where no value occurs more than once.

On the other hand, sometimes a frequency distribution contains two modes in which case it is called a bi-modal distribution as shown below:

THE BI-MODAL FREQUENCY DISTRIBUTION



The next measure of central tendency to be discussed is the arithmetic mean.

THE ARITHMETIC MEAN

The arithmetic mean is the statistician's term for what the layman knows as the average. It can be thought of as that value of the variable series which is *numerically* MOST representative of the whole series. Certainly, this is the most widely used average in statistics. Easiest In addition, it is probably the to calculate.
Its formal definition is:
"The arithmetic mean or simply the mean is a value obtained by dividing the sum of all the observations by their number."

$$\bar{X} = \frac{\text{Sum of all the observations}}{\text{Number of the observations}}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where n represents the number of observations in the sample that has been the ith observation in the sample (i = 1, 2, 3, …, n), and represents the mean of the sample.
For simplicity, the above formula can be written as

$$\bar{X} = \frac{\sum X}{n}$$

In other words, it is not necessary to insert the subscript 'i'.)

EXAMPLE:

Information regarding the receipts of a news agent for seven days of a particular week are given below

Day	Receipt of News Agent
Monday	£ 9.90
Tuesday	£ 7.75
Wednesday	£ 19.50
Thursday	£ 32.75
Friday	£ 63.75
Saturday	£ 75.50
Sunday	£ 50.70
Week Total	£ 259.85

Mean sales per day in this week:
= £ 259.85/7 = £ 37.12 (To the nearest penny).

INTERPRETATION:

The mean, £ 37.12, represents the amount (in pounds sterling) that would have been obtained on each day if the same amount were to be obtained on each day. The above example pertained to the computation of the arithmetic mean in case of ungrouped data i.e. raw data.

Let us now consider the case of data that has been grouped into a frequency distribution. When data pertaining to a continuous variable has been grouped into a frequency distribution, the frequency distribution is used to calculate the approximate values of descriptive measures --- as the identity of the observations is lost.
To calculate the approximate value of the mean, the observations in each class are assumed to be identical with the class midpoint Xi.
The mid-point of every class is known as its class-mark. In other words, the midpoint of a class 'marks' that class. As was just mentioned, the observations in each class are assumed to be identical with the midpoint i.e. the class-mark. (This is based on the assumption that the observations in the group are evenly scattered between the two extremes of the class interval).
As was just mentioned, the observations in each class are assumed to be identical with the midpoint i.e. the class-mark. (This is based on the assumption that the observations in the group are evenly scattered between the two extremes of the class interval).

FREQUENCY DISTRIBUTION

Mid Point X	Frequency f
X_1	f_1
X_2	f_2
X_3	f_3
:	:
:	:
:	:
X_k	f_k

In case of a frequency distribution, the arithmetic mean is defined as:
ARITHMETIC MEAN

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i X_i}{n}$$

For simplicity, the above formula can be written as

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{n}$$ (The subscript 'i' can be dropped.)

Let us understand this point with the help of an example:
Going back to the example of EPA mileage ratings, that we dealt with when discussing the formation of a frequency distribution. The frequency distribution that we obtained was:

EPA MILEAGE RATINGS OF 30 CARS OF A CERTAIN MODEL

Class (Mileage Rating)	Frequency (No. of Cars)
30.0 – 32.9	2
33.0 – 35.9	4
36.0 – 38.9	14
39.0 – 41.9	8
42.0 – 44.9	2
Total	30

The first step is to compute the mid-point of every class.
(You will recall that the concept of the mid-point has already been discussed in an earlier lecture.)

CLASS-MARK (MID-POINT):

The mid-point of each class is obtained by adding the sum of the two limits of the class and dividing by 2.
Hence, in this example, our mid-points are computed in this manner:
30.0 plus 32.9 divided by 2 is equal to 31.45,
33.0 plus 35.9 divided by 2 is equal to 34.45,

And so on.

Class (Mileage Rating)	Class-mark (Midpoint) X
30.0 – 32.9	31.45
33.0 – 35.9	34.45
36.0 – 38.9	37.45
39.0 – 41.9	40.45
42.0 – 44.9	43.45

In order to compute the arithmetic mean, we first need to construct the column of fX, as shown below:

Class-mark (Midpoint) X	Frequency f	fX
31.45	2	62.9
34.45	4	137.8
37.45	14	524.3
40.45	8	323.6
43.45	2	86.9
	30	1135.5

Applying the formula

$$\bar{X} = \frac{\sum fX}{\sum f},$$

We obtain

$$\bar{X} = \frac{1135.5}{30} = 37.85$$

INTERPRETATION:

The average mileage rating of the 30 cars tested by the Environmental Protection Agency is 37.85 – on the average, these cars run 37.85 miles per gallon. An important concept to be discussed at this point is the concept of grouping error.

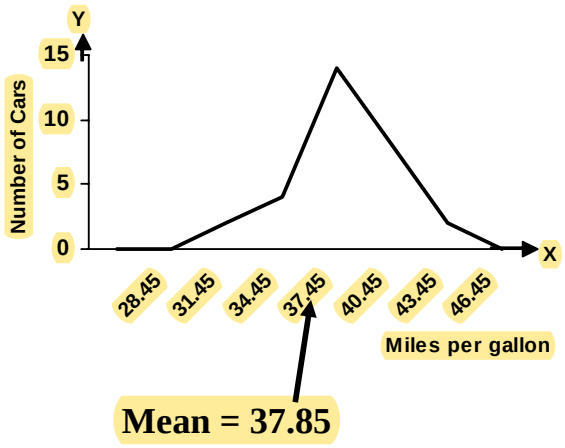
GROUPING ERROR:

"Grouping error" refers to the error that is introduced by the assumption that all the values falling in a class are equal to the mid-point of the class interval. In reality, it is highly improbable to have a class for which all the values lying in that class are equal to the mid-point of that class. This is why the mean that we calculate from a frequency distribution does not give exactly the same answer as what we would get by computing the mean of our raw data.
As indicated earlier, a frequency distribution is used to calculate the approximate values of various descriptive measures.(The word 'approximate' is being used because of the grouping error that was just discussed.) This grouping error arises in the computation of many descriptive measures such as the geometric mean, harmonic mean, mean deviation and standard deviation. But, experience has shown that in the calculation of the arithmetic mean, this error is usually small and never serious. Only a slight difference occurs between the true answer that we would get from the raw data, and the answer that we get from the data that has been grouped in the form of a frequency distribution.
In this example, if we calculate the arithmetic mean directly from the 30 EPA mileage ratings, we obtain:
Arithmetic mean computed from raw data of the EPA mileage ratings:

$$\bar{X} = \frac{363+301+....+339+398}{30}$$

$$= \frac{1134.7}{30} = 37.82$$

The difference between the true value of i.e. 37.82 and the value obtained from the frequency distribution i.e. 37.85 is indeed very slight. The arithmetic mean is predominantly used as a measure of central tendency.
The question is, "Why is it that the arithmetic mean is known as a measure of central tendency?"
The answer to this question is that we have just obtained i.e. 37.85 falls more or less in the centre of our frequency distribution.



As indicated earlier, the arithmetic mean is predominantly used as a measure of central tendency.
It has many desirable properties:

DESIRABLE PROPERTIES OF THE ARITHMETIC MEAN

- Best understood average in statistics.
- Relatively easy to calculate
- Takes into account every value in the series.

But there is one limitation to the use of the arithmetic mean:
As we are aware, every value in a data-set is included in the calculation of the mean, whether the value be high or low. Where there are a few very high or very low values in the series, their effect can be to *drag* the arithmetic mean towards them. This may make the mean unrepresentative.

EXAMPLE:

Example of the Case Where the Arithmetic Mean Is Not a Proper Representative of the Data:
Suppose one walks down the main street of a large city centre and counts the number of floors in each building.
Suppose, the following answers are obtained:
5, 4, 3, 4, 5, 4, 3, 4, 5, 20, 5, 6, 32, 8, 27
The mean number of floors is 9 even though 12 out of 15 of the buildings have 6 floors or less.
The three skyscraper blocks are having a disproportionate effect on the arithmetic mean (Some other average in this case would be more representative). The concept that we just considered was the concept of the simple arithmetic mean.
Let us now discuss the concept of the weighted arithmetic mean.
Consider the following example:

EXAMPLE:

Suppose that in a particular high school, there are:-

100	–	freshmen
80	–	sophomores
70	–	juniors
50	–	seniors

And suppose that on a given day, 15% of freshmen, 5% of sophomores, 10% of juniors, 2% of seniors are *absent.*
The problem is that: What percentage of students is absent for the school as a whole on that particular day?
Now a student is likely to attempt to find the answer by adding the percentages and dividing by 4
i.e.

$$\frac{15+5+10+2}{4}=\frac{32}{4}=8$$

But the fact of the matter is that the above calculation gives a wrong answer. In order to figure out why this is a wrong calculation, consider the following: As we have already noted, 15% of the freshmen are absent on this particular day. Since, in all, there are 100 freshmen in the school, hence the total number of freshmen who are absent is also 15.

But as far as the sophomores are concerned, the total number of them in the school is 80, and if 5% of them are absent on this particular day, this means that the total number of sophomores who are absent is only 4. Proceeding in this manner, we obtain the following table.

Category of Student	Number of Students in the school	Number of Students who are absent
Freshman	100	15
Sophomore	80	4
Junior	70	7
Senior	50	1
TOTAL	300	27

Dividing the total number of students who are absent by the total number of students enrolled in the school, and multiplying by 100, we obtain:

$$\frac{27}{300} \times 100 = 9$$

Thus its very clear that previous result was not correct. This situation leads us to a very important observation, i.e. here our figures pertaining to absenteeism in various categories of students cannot be regarded as having equal weightage.

When we have such a situation, the concept of "weighing" applies i.e. every data value in the data set is assigned a certain weight according to a suitable criterion. In this way, we will have a weighted series of data instead of an un-weighted one. In this example, the number of students enrolled in each category acts as the weight for the number of absences pertaining to that category i.e.

Category of Student	Percentage of Students who are absent X_i	Number of students enrolled in the school (Weights) W_i	W_iX_i (Weighted X_i)
Freshman	15	100	$100 \times 15 = 1500$
Sophomore	5	80	$80 \times 5 = 400$
Junior	10	70	$70 \times 10 = 700$
Senior	2	50	$50 \times 2 = 100$
	Total	$\Sigma W_i = 300$	$\Sigma W_iX_i = 2700$

The formula for the weighted arithmetic mean is:

WEIGHTED MEAN

$$\bar{X}_w = \frac{\sum W_i X_i}{\sum W_i}$$

And, in this example, the weighted mean is equal to:

$$\bar{X}_w = \frac{\sum W_i X_i}{\sum W_i}$$

$$= \frac{2700}{300}$$

$$= 9$$

Thus we note that, in this example, the weighted mean yields exactly the same as the answer that we obtained earlier.

As obvious, the weighing process leads us to a correct answer under the situation where we have data that cannot be regarded as being such that each value should be given equal weightage.

An important point to note here is the *criterion* for assigning weights. Weights can be assigned in a number of ways depending on the situation and the problem domain.

The next measure of central tendency that we will discuss is the median.

Let us understand this concept with the help of an example.

Let us return to the problem of the 'average' number of floors in the buildings at the centre of a city. We saw that the arithmetic mean was distorted towards the few extremely high values in this series and became unrepresentative.

We could more appropriately and easily employ the median as the 'average' in these circumstances.

MEDIAN

The median is the middle value of the series when the variable values are placed in order of magnitude.

The median is defined as a "value which divides a set of data into two halves, one half comprising of observations greater than and the other half smaller than it. More precisely, the median is a value at or below which 50% of the data lie."

The median value can be ascertained by inspection in many series. For instance, in this very example, the data that we obtained was:

EXAMPLE-1

The average number of floors in the buildings at the centre of a city:
5, 4, 3, 4, 5, 4, 3, 4, 5, 20, 5, 6, 32, 8, 27

Arranging these values in ascending order, we obtain
3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 8, 20, 27, 32

Picking up the middle value, we obtain the median equal to 5.

INTERPRETATION

The median number of floors is 5. Out of those 15 buildings, 7 have unto 5 floors and 7 have 5 floors or more. We noticed earlier that the arithmetic mean was distorted toward the few extremely high values in the series and hence became unrepresentative. The median = 5 is much more representative of this series.

Height of buildings (number of floors)	
3	
3	
4	
4	7 lower
4	
5	
5	
5 = median height	
5	
5	
6	
8	7 higher
20	
27	
32	

EXAMPLE 2

Retail price of motor-car (£) (several makes and sizes)	
415	
480	4 above
525	
608	
719	= median price
1,090	
2,059	
4,000	4 above
6,000	

A slight complication arises when there are even numbers of observations in the series, for now there are two middle values.
The expedient of taking the arithmetic mean of the two is adopted as explained below:

EXAMPLE-3

Number of passengers travelling on a bus at six Different times during the day	
4	
9	
14	= median value
18	
23	
47	
Median = $\frac{14+18}{2}$ = 16 passengers	

EXAMPLE -4:

The number of passengers traveling on a bus at six different times during a day is as follows:
5, 14, 47, 34, 18, 23
Find the median.

Solution:
Arranging the values in ascending order, we obtain
5, 14, 18, 23, 34, 47
As before, a slight complication has arisen because of the fact that there are even numbers of observations in the series and, as such, there are two middle values. As before, we take the arithmetic mean of the two middle values.
Hence we obtain:

Median:

$$\tilde{X} = \frac{18 + 23}{2} = 20.$$

A very important point to be noted here is that we must arrange the data in ascending order before searching for the two middle values. All the above examples pertained to raw data. Let us now consider the case of grouped data.
We begin by discussing the case of discrete data grouped into a frequency table.
As stated earlier, a discrete frequency distribution is no more than a concise representation of a simple series pertaining to a discrete variable, so that the same approach as the one discussed just now would seem relevant.

EXAMPLE OF A DISCRETE FREQUENCY DISTRIBUTION

Comprehensive School

Number of pupils per class	Number of Classes
23	1
24	0
25	1
26	3
27	6
28	9
29	8
30	10
31	7
	45

In order to locate the middle value, the best thing is to first of all construct a column of cumulative frequencies:

Comprehensive School

Number of pupils per class X	Number of Classes f	Cumulative Frequency cf
23	1	1
24	0	1
25	1	2
26	3	5
27	6	11
28	9	20
29	8	28
30	10	38
31	7	45
	45	

In this school, there are 45 classes in all, so that we require as the median that class-size below which there are 22 classes and above which also there are 22 classes.
In other words, we must find the 23rd class in an ordered list. We could simply count down noticing that there is 1 class of 23 children, 2 classes with up to 25 children, 5 classes with up to 26 children. Proceeding in this manner, we find that 20 classes contain up to 28 children whereas 28 classes contain up to 29 children. This means that the 23rd class --- the one that we are looking for --- is the one which contains exactly 29 children.

Comprehensive School

Number of pupils per class X	Number of Classes f	Cumulative Frequency cf
23	1	1
24	0	1
25	1	2
26	3	5
27	6	11
28	9	20
29	8	28
30	10	38
31	7	45
	45	

Median number of pupils per class:

$$\tilde{X} = 29$$

This means that 29 is the middle size of the class. In other words, 22 classes are such which contain 29 or less than 29 children, and 22 classes are such which contain 29 or more than 29 children.

LECTURE NO. 8

- Median in case of a frequency distribution of a continuous variable
- Median in case of an open-ended frequency distribution
- Empirical relation between the mean, median and the mode
- Quantiles (quartiles, deciles & percentiles)
- Graphic location of quantiles.

MEDIAN IN CASE OF A FREQUENCY DISTRIBUTION OF A CONTINUOUS VARIABLE:

In case of a frequency distribution, the median is given by the formula

$$\tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

Where
l =lower class boundary of the median class (i.e. that class for which the cumulative frequency is just in excess of n/2).
h=class interval size of the median class
f =frequency of the median class
n=Σf (the total number of observations)
c =cumulative frequency of the class preceding the median class

Note:
This formula is based on the assumption that the observations in each class are evenly distributed between the two class limits.

EXAMPLE:

Going back to the example of the EPA mileage ratings, we have

Mileage Rating	No. of Cars	Class Boundaries	Cumulative Frequency
30.0 – 32.9	2	29.95 – 32.95	2
33.0 – 35.9	4	32.95 – 35.95	6
36.0 – 38.9	14	35.95 – 38.95	20
39.0 – 41.9	8	38.95 – 41.95	28
42.0 – 44.9	2	41.95 – 44.95	30

In this example, n = 30 and n/2 = 15.
Thus the third class is the median class. The median lies somewhere between 35.95 and 38.95. Applying the above formula, we obtain

$$\begin{aligned}\tilde{X} &= 35.95 + \frac{3}{14}(15-6)\\ &= 35.95 + 1.93\\ &= 37.88\\ &\simeq 37.9\end{aligned}$$

INTERPRETATION

This result implies that half of the cars have mileage less than or up to 37.88 miles per gallon whereas the other half of the cars has mileage greater than 37.88 miles per gallon. As discussed earlier, the median is preferable to the arithmetic

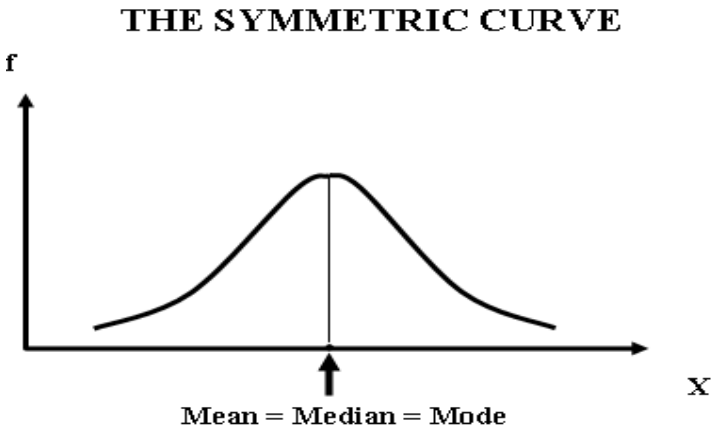
mean when there are a few very high or low figures in a series. It is also exceedingly valuable when one encounters a frequency distribution having open-ended class intervals.
The concept of open-ended frequency distribution can be understood with the help of the following example.

Example:

WAGES OF WORKERS IN A FACTORY	
Monthly Income (in Rupees)	No. of Workers
Less than 2000/-	100
2000/- to 2999/-	300
3000/- to 3999/-	500
4000/- to 4999/-	250
5000/- and above	50
Total	1200

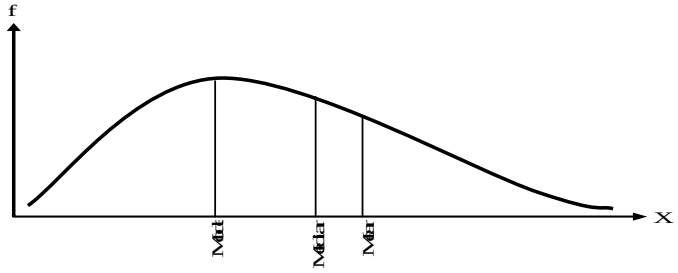
In this example, both the first class and the last class are open-ended classes. This is so because of the fact that we do not have exact figures to begin the first class or to end the last class. The advantage of computing the median in the case of an open-ended frequency distribution is that, except in the unlikely event of the median falling within an open-ended group occurring in the beginning of our frequency distribution, there is no need to estimate the upper or lower boundary. This is so because of the fact that, if the median is falling in an intermediate class, then, obviously, the first class is not being involved in its computation. The next concept that we will discuss is the empirical relation between the mean, median and the mode. This is a concept which is not based on a rigid mathematical formula; rather, it is based on observation. In fact, the word 'empirical' implies 'based on observation'.

This concept relates to the relative positions of the mean, median and the mode in case of a hump-shaped distribution. In a single-peaked frequency distribution, the values of the mean, median and mode coincide if the frequency distribution is absolutely symmetrical.



But in the case of a skewed distribution, the mean, median and mode do not all lie on the same point. They are pulled apart from each other, and the empirical relation explains the way in which this happens. Experience tells us that in a unimodal curve of moderate skewness, the median is usually sandwiched between the mean and the mode.

The second point is that, in the case of many real-life data-sets, it has been observed that the distance between the mode and the median is approximately double of the distance between the median and the mean, as shown below:



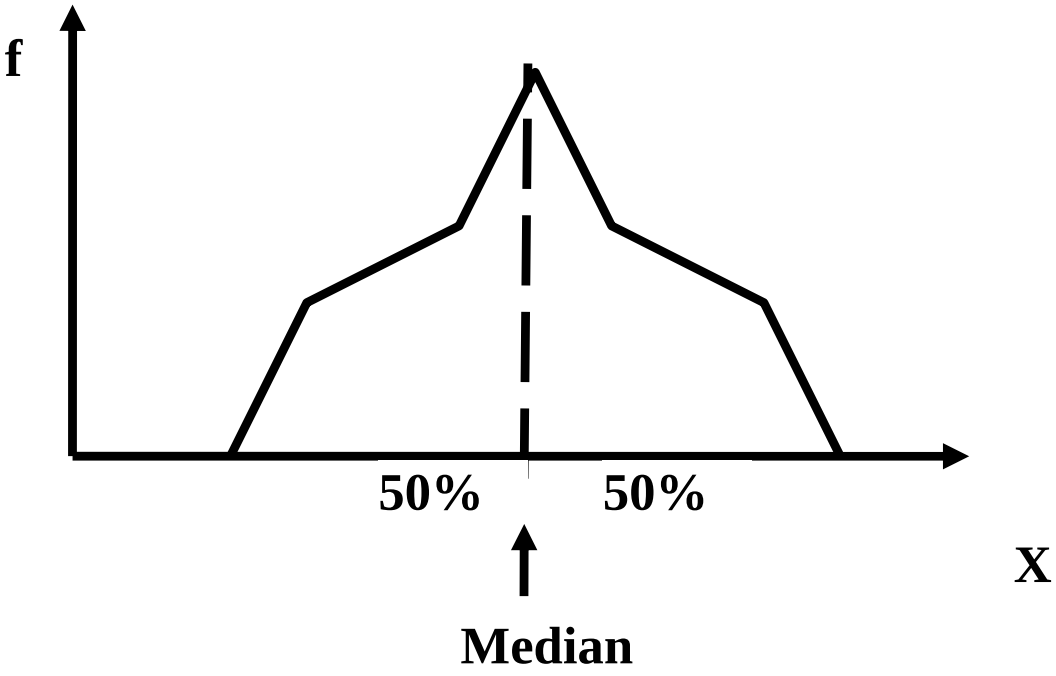
This diagrammatic picture is equivalent to the following algebraic expression:
Median - Mode 2 (Mean - Median) ---- (1)
The above-mentioned point can also be expressed in the following way:

Mean – Mode = 3 (Mean – Median) ---- (2)
Equation (1) as well as equation (2) yields the approximate relation given below:

EMPIRICAL RELATION BETWEEN THE MEAN, MEDIAN AND THE MODE

Mode = 3 Median – 2 Mean
An exactly similar situation holds in case of a moderately negatively skewed distribution.
An important point to note is that this empirical relation does not hold in case of a J-shaped or an extremely skewed distribution.

Let us now extend the concept of partitioning of the frequency distribution by taking up the concept of quantiles (i.e. quartiles, deciles and percentiles).
We have already seen that the median divides the area under the frequency polygon into two equal halves:



A further split to produce quarters, tenths or hundredths of the total area under the frequency polygon is equally possible, and may be extremely useful for analysis. (We are often interested in the highest 10% of some group of values or the middle 50% another.

QUARTILES

The quartiles, together with the median, achieve the division of the total area into four equal parts.
The first, second and third quartiles are given by the formulae:

1. FIRST QUARTILE

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

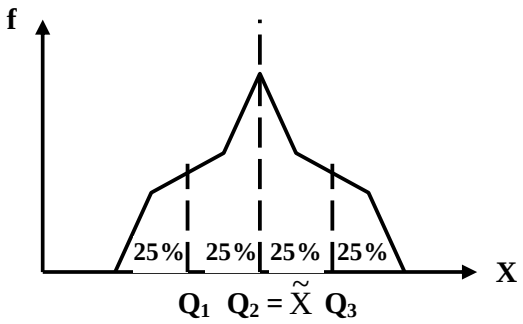
2. SECOND QUARTILE (I.E. MEDIAN)

$$Q_2 = l + \frac{h}{f}\left(\frac{2n}{4} - c\right) = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

3. THIRD QUARTILE

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

It is clear from the formula of the second quartile that the second quartile is the same as the median.



DECILES & PERCENTILES

The deciles and the percentiles give the division of the total area into 10 and 100 equal parts respectively.
The formula for the first decile is

$$D_1 = l + \frac{h}{f}\left(\frac{n}{10} - c\right)$$

The formulae for the subsequent deciles are

$$D_2 = l + \frac{h}{f}\left(\frac{2n}{10} - c\right)$$

$$D_3 = l + \frac{h}{f}\left(\frac{3n}{10} - c\right)$$

and so on.

It is easily seen that the 5th decile is the same quantity as the median.
The formula for the first percentile is

$$P_1 = l + \frac{h}{f}\left(\frac{n}{100} - c\right)$$

The formulae for the subsequent percentiles are

$$P_2 = l + \frac{h}{f}\left(\frac{2n}{100} - c\right)$$

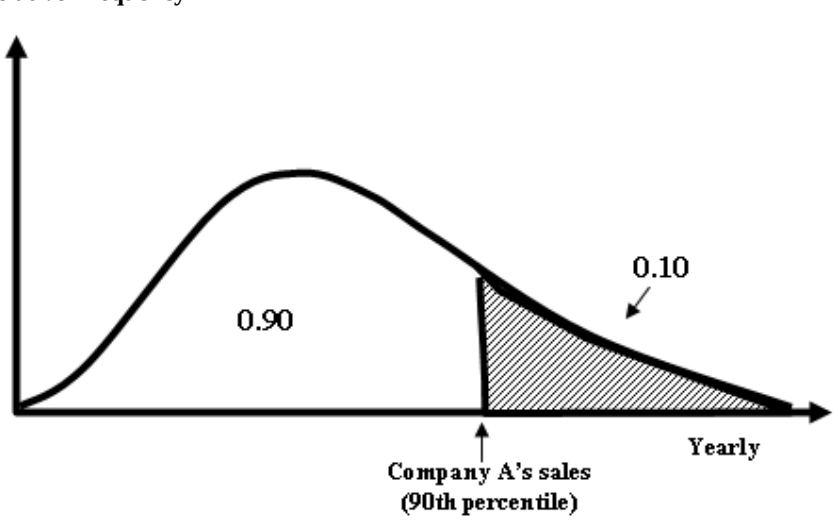
$$P_3 = l + \frac{h}{f}\left(\frac{3n}{100} - c\right)$$

and so on.
Again, it is easily seen that the 50th percentile is the same as the median, the 25th percentile is the same as the 1st quartile, the 75th percentile is the same as the 3rd quartile, the 40th percentile is the same as the 4th decile, and so on.

All these measures i.e. the median, quartiles, deciles and percentiles are collectively called quantiles. The question is, "What is the significance of this concept of partitioning? *Why* is it that we wish to divide our frequency distribution into two, four, ten or hundred parts?"
The answer to the above questions is: In certain situations, we may be interested in describing the *relative* quantitative location of a particular measurement within a data set. Quantiles provide us with an easy way of achieving this. Out of these various quantiles, one of the most frequently used is percentile ranking.
Let us understand this point with the help of an example.

EXAMPLE

If oil company 'A' reports that its yearly sales are at the 90th percentile of all companies in the industry, the implication is that 90% of all oil companies have yearly sales *less* than company A's, and only 10% have yearly sales exceeding company A's,this is demonstrated in the following figure:



It is evident from the above example that the concept of percentile ranking is quite a useful concept, but it should be kept in mind that percentile rankings are of practical value only for large data sets.
It is evident from the above example that the concept of percentile ranking is quite a useful concept, but it should be kept in mind that percentile rankings are of practical value only for large data sets. The next concept that we will discuss is the graphic location of quantiles.
Let us go back to the example of the EPA mileage ratings of 30 cars that was discussed in an earlier lecture.
EXAMPLE

Suppose that the Environmental Protection Agency of a developed country performs extensive tests on all new car models in order to determine their mileage rating. Suppose that the following 30 measurements are obtained by conducting such tests on a particular new car model.

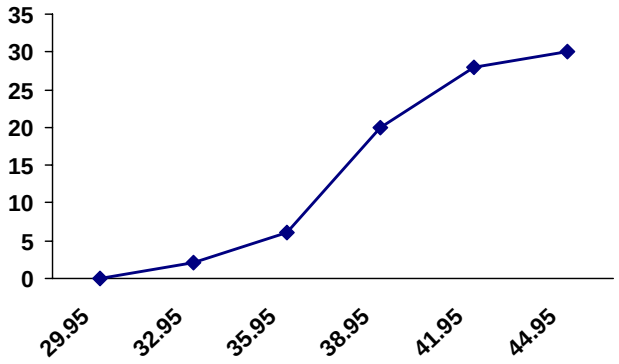
EPA MILEAGE RATINGS ON 30 CARS (MILES PER GALLON)		
36.3	42.1	44.9
30.1	37.5	32.9
40.5	40.0	40.2
36.2	35.6	35.9
38.5	38.8	38.6
36.3	38.4	40.5
41.0	39.0	37.0
37.0	36.7	37.1
37.1	34.8	33.9
39.9	38.1	39.8

When the above data was converted to a frequency distribution, we obtained:

Class Limit	Frequency
30.0 – 32.9	2
33.0 – 35.9	4
36.0 – 38.9	14
39.0 – 41.9	8
42.0 – 44.9	2
	30

Also, we considered the graphical representation of this distribution.
The cumulative frequency polygon of this distribution came out to be as shown in the following figure:

Cumulative Frequency Polygon or OGIVE

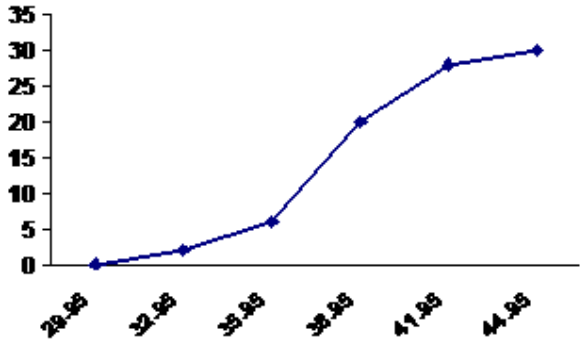


This ogive enables us to find the median and any other quantile that we may be interested in very conveniently. And this process is known as the graphic location of quantiles.
Let us begin with the graphical location of the median:

Because of the fact that the median is that value before which half of the data lies, the first step is to divide the total number of observations n by 2.
In this example:

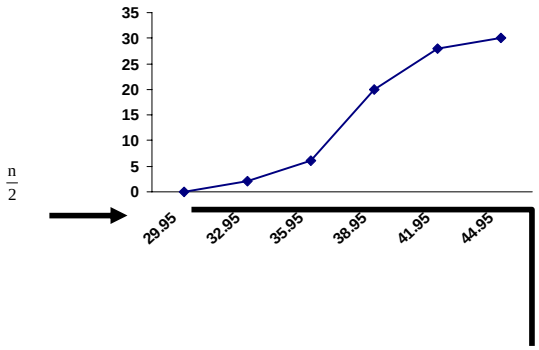
$$\frac{n}{2} = \frac{30}{2} = 15$$

The next step is to locate this number 15 on the y-axis of the cumulative frequency polygon.



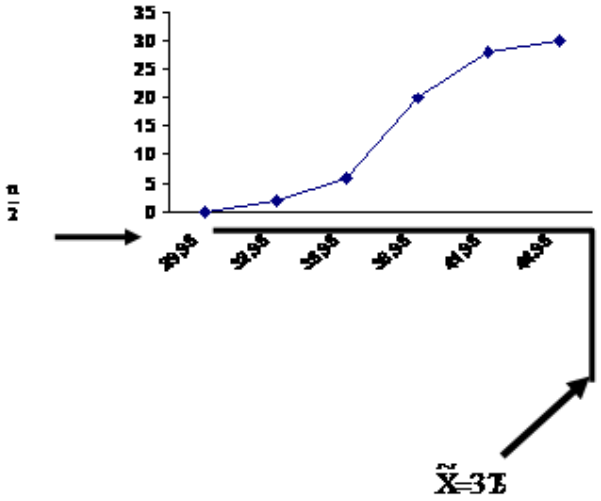
Lastly, we drop a vertical line from the cumulative frequency polygon down to the x-axis.

Cumulative Frequency Polygon or OGIVE



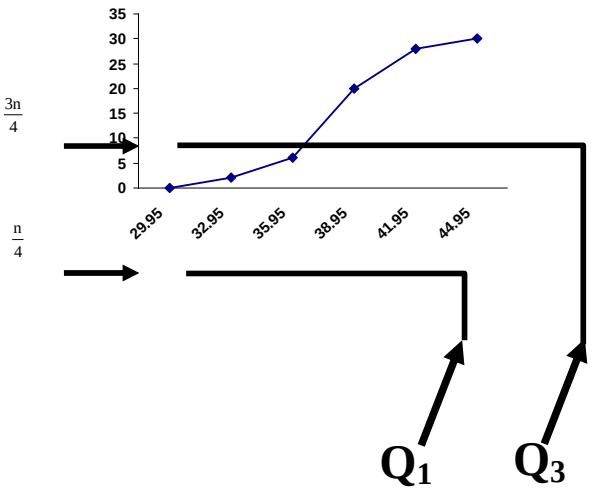
Now, if we read the x-value where our perpendicular touches the x-axis, students, we find that this value is approximately the same as what we obtained from our formula.

Cumulative Frequency Polygon or OGIVE



It is evident from the above example that the cumulative frequency polygon is a very useful device to find the value of the median very quickly. In a similar way, we can locate the quartiles, deciles and percentiles. To obtain the first quartile, the horizontal line will be drawn against the value n/4, and for the third quartile, the horizontal line will be drawn against the value 3n/4.

Cumulative Frequency Polygon or OGIVE



For the deciles, the horizontal lines will be against the values n/10, 2n/10, 3n/10, and so on. And for the percentiles, the horizontal lines will be against the values n/100, 2n/100, 3n/100, and so on.

The graphic location of the quartiles as well as of a few deciles and percentiles for the data-set of the EPA mileage ratings may be taken up as an exercise:

This brings us to the end of our discussion regarding quantiles which are sometimes also known as fractiles --- this terminology because of the fact that they divide the frequency distribution into various parts or fractions.

LECTURE NO. 9

- Geometric mean
- Harmonic mean
- Relation between the arithmetic, geometric and harmonic means
- Some other measures of central tendency

GEOMETRIC MEAN

The geometric mean, G, of a set of n positive values X1, X2,…,Xn is defined as the positive nth root of their product.

$$G = \sqrt[n]{X_1 X_2 ... X_n}$$

(Where $X_i > 0$)

When n is large, the computation of the geometric mean becomes laborious as we have to extract the nth root of the product of all the values.
The arithmetic is simplified by the use of logarithms.

Taking logarithms to the base 10, we get

$$\log G = \frac{1}{n}[\log X_1 + \log X_2 + ... + \log X_n]$$

Hence

$$= \frac{\sum \log X}{n}$$

$$G = \text{anti log}\left[\frac{\sum \log X}{n}\right]$$

EXAMPLE

Find the geometric mean of numbers:

45, 32, 37, 46, 39, 36, 41, 48, 36

Solution:
We need to compute the numerical value of

$$= \sqrt[9]{45 \times 32 \times 37 \times 46 \times 39 \times 36 \times 41 \times 48 \times 36}$$

But, obviously, it is a bit cumbersome to find the ninth root of a quantity. So we make use of logarithms, as shown below:

X	log X
45	1.6532
32	1.5052
37	1.5682
46	1.6628
39	1.5911
36	1.5563
41	1.6128
48	1.6812
36	1.5563
	14.3870

$$\log G = \frac{\sum \log X}{n}$$

$$= \frac{14.3870}{9} = 1.5986$$

$$\text{Hence } G = \text{anti log } 1.5986$$

$$= 39.68$$

The above example pertained to the computation of the geometric mean in case of raw data.
Next, we consider the computation of the geometric mean in the case of grouped data.

GEOMETRICMEAN FOR GROUPED DATA

In case of a frequency distribution having k classes with midpoints X1, X2, …,Xk and the corresponding frequencies f1, f2, …, fk (such that ∑fi = n), the geometric mean is given by

$$G = \sqrt[n]{X_1^{f_1} X_2^{f_2} \ldots X_k^{f_k}}$$

Each value of X thus has to be multiplied by itself f times, and the whole procedure becomes quite a formidable task!
In terms of logarithms, the formula becomes

$$\log G = \frac{1}{n}\left[f_1 \log X_1 + f_2 \log X_2 + \ldots + f_k \log X_k\right]$$

Hence

$$= \frac{\sum f \log X}{n}$$

$$G = \text{antilog}\left[\frac{\sum f \log X}{n}\right]$$

Obviously, the above formula is much easier to handle. Let us now apply it to an example.
Going back to the example of the EPA mileage ratings, we have:

Mileage Rating	No. of Cars	Class-mark (midpoint) X	log X	f log X
30.0 - 32.9	2	31.45	1.4976	2.9952
33.0 - 35.9	4	34.45	1.5372	6.1488
36.0 - 38.9	14	37.45	1.5735	22.0290
39.0 - 41.9	8	40.45	1.6069	12.8552
42.0 - 44.9	2	43.45	1.6380	3.2760
	30			47.3042

$$G = \text{antilog}\ \frac{47.3042}{30}$$

= antilog 1.5768 = 37.74
This means that, if we use the geometric mean to measures the central tendency of this data set, then the central value of the mileage of those 30 cars comes out to be 37.74 miles per gallon.
The question is, “When should we use the geometric mean?”
The answer to this question is that when relative changes in some variable quantity are averaged, we prefer the geometric mean.

EXAMPLE

Suppose it is discovered that a firm’s turnover has increased during 4 years by the following amounts:

Year	Turnover	Percentage Compared With Year Earlier
1958	£ 2,000	–
1959	£ 2,500	125
1960	£ 5,000	200
1961	£ 7,500	150
1962	£ 10,500	140

The yearly increase is shown in a percentage form in the right-hand column i.e. the turnover of 1959 is 125 percent of the turnover of 1958, the turnover of 1960 is 200 percent of the turnover of 1959, and so on. The firm's owner may be interested in knowing his average rate of turnover growth.
If the arithmetic mean is adopted he finds his answer to be:

Arithmetic Mean:

$$\frac{125+200+150+140}{4}$$

$$=153.75$$

i.e. we are concluding that the turnover for any year is 153.75% of the turnover for the previous year. In other words, the turnover in each of the years considered appears to be 53.75 per cent higher than in the previous year.
If *this* percentage is used to calculate the turnover from 1958 to 1962 inclusive, we obtain:

153.75% of £ 2,000 = £ 3,075
153.75% of £ 3,075 = £ 4,728
153.75% of £ 4,728 = £ 7,269
153.75% of £ 7,269 = £ 11,176

Whereas the actual turnover figures were

Year	Turnover
1958	£ 2,000
1959	£ 2,500
1960	£ 5,000
1961	£ 7,500
1962	£ 10,500

It seems that both the individual figures and, more important, the total at the end of the period, are incorrect. Using the arithmetic mean has exaggerated the 'average' annual rate of increase in the turnover of this firm. Obviously, we would like to rectify this false impression. The geometric mean enables us to do so:
Geometric mean of the turnover figures:

$$\sqrt[4]{(125\times 200\times 150\times 140)}$$

$$=\sqrt[4]{525000000}$$

$$=151.37\%$$

Now, if we utilize this particular value to obtain the individual turnover figures, we find that:
151.37% of £2,000 = £3,027
151.37% of £3,027 = £4,583
151.37% of £4,583 = £6,937
151.37% of £6,937 = £10,500
So that the turnover figure of 1962 is exactly the same as what we had in the original data.

INTERPRETATION

If the turnover of this company were to increase annually at a constant rate, then the annual increase would have been 51.37 percent.(On the average, each year's turnover is 51.37% higher than that in the previous year.) The above example clearly indicates the significance of the geometric mean in a situation when relative changes in a variable quantity are to be averaged.

But we should bear in mind that such situations are not encountered too often, and that the occasion to calculate the geometric mean arises less frequently than the arithmetic mean.(The most frequently used measure of central tendency is the arithmetic mean.)
The next measure of central tendency that we will discuss is the harmonic mean.

HARMONIC MEAN

The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of the values.
In case of raw data:

$$H.M.=\frac{n}{\sum\left(\frac{1}{X}\right)}$$

In case of grouped data (data grouped into a frequency distribution):

$$H.M. = \frac{n}{\sum f\left(\frac{1}{X}\right)}$$

(Where X represents the midpoints of the various classes)

EXAMPLE

Suppose a car travels 100 miles with 10 stops, each stop after an interval of 10 miles. Suppose that the speeds at which the car travels these 10 intervals are 30, 35, 40, 40, 45, 40, 50, 55, 55 and 30 miles per hours respectively.
What is the average speed with which the car traveled the total distance of 100 miles?
If we find the arithmetic mean of the 10 speeds, we obtain:
Arithmetic mean of the 10 speeds:

$$= \frac{30+35+....+30}{\frac{420}{10}} = 42 \text{ miles per hour.}$$

But, if we study the problem carefully, we find that the above answer is incorrect.
By definition, the average speed is the speed with which the car would have traveled the 100 mile distance if it had maintained a constant speed throughout the 10 intervals of 10 miles each.

$$\textbf{Average speed} = \frac{\textbf{Total distance travelled}}{\textbf{Total time taken}}$$

Now, *total distance traveled* = 100 miles. *Total time taken* will be computed as shown below:

Interval	Distance	Speed = $\frac{\text{Distance}}{\text{Time}}$	Time = $\frac{\text{Distance}}{\text{Speed}}$
1	10 miles	30 mph	10/30 = 0.3333 hrs
2	10 miles	35 mph	10/35 = 0.2857 hrs
3	10 miles	40 mph	10/40 = 0.2500 hrs
4	10 miles	40 mph	10/40 = 0.2500 hrs
5	10 miles	45 mph	10/45 = 0.2222 hrs
6	10 miles	40 mph	10/40 = 0.2500 hrs
7	10 miles	50 mph	10/50 = 0.2000 hrs
8	10 miles	55 mph	10/55 = 0.1818 hrs
9	10 miles	55 mph	10/55 = 0.1818 hrs
10	10 miles	30 mph	10/30 = 0.333 hrs
Total = 100 miles		Total Time = 2.4881 hrs	

Hence
Average speed = $\frac{100}{2.4881} = 40.2\ mph$
which is not the same as 42 miles per hour.
Let us now try the harmonic mean to find the average speed of the car.

$$H.M. = \frac{n}{\sum\left(\frac{1}{X}\right)}$$

where n is the no. of terms)
We have:

X	1/X
30	1/30 = 0.0333
35	1/35 = 0.0286
40	1/40 = 0.0250
40	1/40 = 0.0250
45	1/45 = 0.0222
40	1/40 = 0.0250
50	1/50 = 0.0200
55	1/55 = 0.0182
55	1/55 = 0.0182
30	1/30 = 0.0333
	$\sum \frac{1}{X} = 0.2488$

$$H.M. = \frac{n}{\sum \frac{1}{X}}$$

$$= \frac{10}{0.2488}$$

$$= 40.2 \text{ mph}$$

Hence it is clear that the harn
gives the totally correct result

The key question is, "When should we compute the harmonic mean of a data set?" The answer to this question will be easy to understand if we consider the following rules:

RULES

- When values are given as x per y where x is constant and y is variable, the Harmonic Mean is the appropriate average to use.
- When values are given as x per y where y is constant and x is variable, the Arithmetic Mean is the appropriate average to use.
- When relative changes in some variable quantity are to be averaged, the geometric mean is the appropriate average to use.

We have already discussed the geometric and the harmonic means. Let us now try to understand Rule No. 1 with the help of an example:

EXAMPLE

If 10 students have obtained the following marks (in a test) out of 20:
13, 11, 9, 9, 6, 5, 19, 17, 12, 9
Then the average marks (by the formula of the arithmetic mean) are:

$$\frac{13+11+9+9+6+5+19+17+12+9}{10}$$

$$= \frac{110}{10} = 11$$

This is equivalent to

$$\frac{\frac{13}{20}+\frac{11}{20}+\frac{9}{20}+\frac{9}{20}+\frac{6}{20}+\frac{5}{20}+\frac{19}{20}+\frac{17}{20}+\frac{12}{20}+\frac{9}{20}}{10}$$

$$= \frac{\frac{110}{20}}{10} = \frac{110}{10 \times 20} = \frac{11}{20}$$

(i.e. the average marks of this group of students are 11 out of 20).In the above example, the point to be noted was that all the marks were expressible as x per y where the denominator y was constant i.e. equal to 20, and hence, it was appropriate to compute the arithmetic mean.
Let us now consider a mathematical relationship exists between these three measures of central tendency.

RELATION BETWEEN ARITHMETIC, GEOMETRIC AND HARMONIC MEANS

Arithmetic Mean ≥ Geometric Mean ≥Harmonic Mean
We have considered the five most well-known measures of central tendency i.e. arithmetic mean, median, mode, geometric mean and harmonic mean. It is interesting to note that there are some other measures of central tendency as well. Two of these are the mid range, and the mid quartile range.
Let us consider these one by one:

MID-RANGE

If there are n observations with x0 and xm as their smallest and largest observations respectively, then their mid-range is defined as

$$\text{mid} - \text{range} = \frac{x_0 + x_m}{2}$$

It is obvious that if we add the smallest value with the largest, and divide by 2, we will get a value which is more or less in the middle of the data-set.

MID-QUARTILE RANGE

If x1, x2… xn are n observations with Q1 and Q3 as their first and third quartiles respectively, then their mid-quartile range is defined as

$$\text{mid} - \text{quartile range} = \frac{Q_1 + Q_3}{2}$$

Similar to the case of the mid-range, if we take the arithmetic mean of the upper and lower quartiles, we will obtain a value that is somewhere in the middle of the data-set. The mid-quartile range is also known as the mid-hinge.
Let us now revise briefly the core concept of central tendency: Masses of data are usually expressed in the form of frequency tables so that it becomes easy to comprehend the data. Usually, a statistician would like to go a step ahead and to compute a number that will represent the data in some definite way.
Any such single number that represents a whole set of data is called **'Average'**.
Technically speaking, there are many kinds of averages (i.e. there are several ways to compute them). These quantities that represent the data-set are called "measures of central tendency".

LECTURE NO. 10

- Concept of dispersion
- Absolute and relative measures of dispersion
- Range
- Coefficient of dispersion
- Quartile deviation
- Coefficient of quartile deviation

Let us begin the concept of DISPERSION.
Just as variable series differ with respect to their location on the horizontal axis (having different 'average' values); similarly, they differ in terms of the amount of variability which they exhibit. Let us understand this point with the help of an example:

EXAMPLE

In a technical college, it may well be the case that the ages of a group of first-year students are quite consistent, e.g. 17, 18, 18, 19, 18, 19, 19, 18, 17, 18 and 18 years.
A class of evening students undertaking a course of study in their spare time may show just the opposite situation, e.g. 35, 23, 19, 48, 32, 24, 29, 37, 58, 18, 21 and 30.
It is very clear from this example that the variation that exists between the various values of a data-set is of substantial importance. We obviously need to be aware of the amount of variability present in a data-set if we are to come to useful conclusions about the situation under review. This is perhaps best seen from studying the two frequency distributions given below.

EXAMPLE

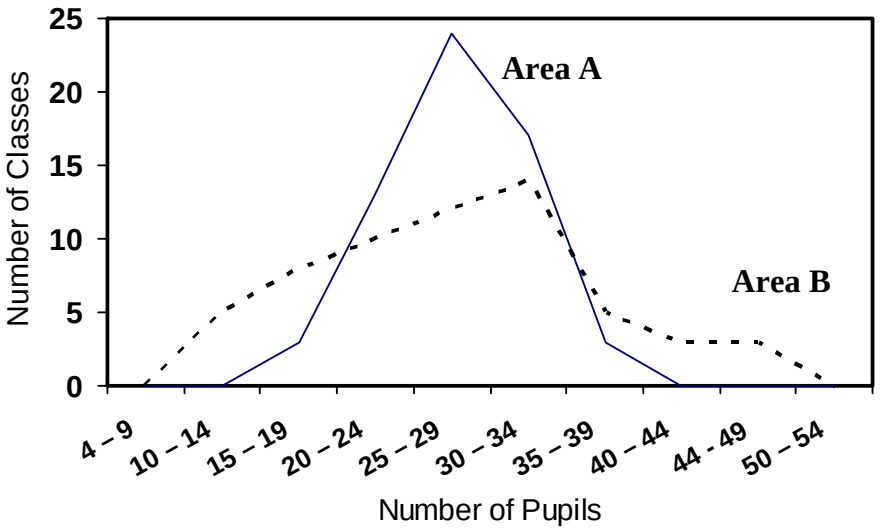
The sizes of the classes in two comprehensive schools in different areas are as follows:

Number of Pupils	Number of Classes	
	Area A	Area B
10 – 14	0	5
15 – 19	3	8
20 – 24	13	10
25 – 29	24	12
30 – 34	17	14
35 – 39	3	5
40 – 44	0	3
45 - 49	0	3
	60	60

If the arithmetic mean size of class is calculated, we discover that the answer is identical: 27.33 pupils in both areas.
Average class-size of each school

$$\overline{X} = 27.33$$

Even though these two distributions share a common average, it can readily be seen that they are entirely DIFFERENT. And the graphs of the two distributions (given below) clearly indicate this fact.



The question which must be posed and answered is 'In what way can these two situations be distinguished?' We need a measure of variability or DISPERSION to accompany the relevant measure of position or 'average' used.
The word 'relevant' is important here for we shall find one measure of dispersion which expresses the scatter of values round the arithmetic mean, another the scatter of values round the median, and so forth. Each measure of dispersion is associated with a particular 'average'.

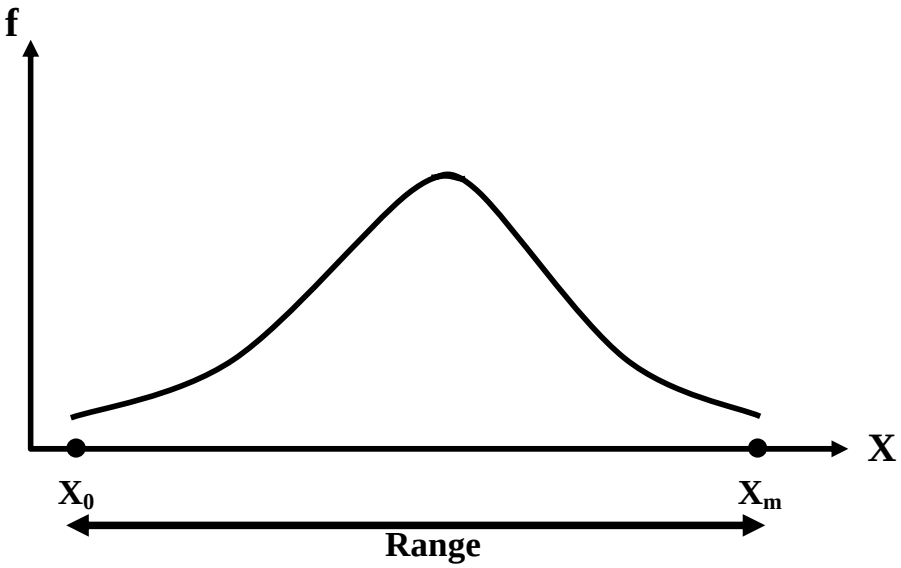
ABSOLUTE VERSUS RELATIVE MEASURES OF DISPERSION

There are two types of measurements of dispersion: absolute and relative.
An absolute measure of dispersion is one that measures the dispersion in terms of the same units or in the square of units, as the units of the data.
For example, if the units of the data are rupees, meters, kilograms, etc., the units of the measures of dispersion will also be rupees, meters, kilograms, etc.
On the other hand, relative measure of dispersion is one that is expressed in the form of a ratio, co-efficient of percentage and is independent of the units of measurement.
A *relative* measure of dispersion is useful for comparison of data of different nature. A measure of central tendency together with a measure of dispersion gives an adequate description of data. We will be discussing FOUR measures of dispersion i.e. the range, the quartile deviation, the mean deviation, and the standard deviation.

RANGE

The range is defined as the difference between the two extreme values of a data-set, i.e. R = Xm – X0 where Xm represents the highest value and X0 the lowest.
Evidently, the calculation of the range is a simple question of MENTAL arithmetic.

The simplicity of the concept does not necessarily invalidate it, but in general it gives no idea of the DISTRIBUTION of the observations between the two ends of the series. For this reason it is used principally as a supplementary aid in the description of variable data, in conjunction with other measures of dispersion. When the data are grouped into a frequency distribution, the range is estimated by finding the difference between the upper boundary of the highest class and the lower boundary of the lowest class.
We now consider the graphical representation of the range:



Obviously, the greater the difference between the largest and the smallest values, the greater will be the range. As stated earlier, the range is a simple concept and is easy to compute. However, because of the fact that it is computed from only the two extreme values in a data-set, it has two serious disadvantages.

- It *ignores* all the INFORMATION available from the intermediate observations.
- It might give a MISLEADING picture of the spread in the data.

From THIS point of view, it is an unsatisfactory measure of dispersion. However, it is APPROPRIATELY used in statistical quality control charts of manufactured products, daily temperatures, stock prices, etc. It is interesting to note that the range can also be viewed in the following way.
It is twice of the arithmetic mean of the deviations of the smallest and largest values round the mid-range i.e.

$$\frac{(\text{Midrange} - X_0) + (X_m - \text{Midrange})}{2}$$

$$= \frac{\text{Midrange} - X_0 + X_m - \text{Midr}}{2}$$

$$= \frac{X_m - X_0}{2}$$

Because of what has been just explained, the range can be regarded as that measure of dispersion which is associated with the mid-range. As such, the range may be employed to indicate dispersion when the mid-range has been adopted as the most appropriate average.
The range is an *absolute* measure of dispersion. Its *relative* measure is known as the CO-EFFICIENT OF DISPERSION, and is defined by the relation given below:
\

COEFFICIENT OF DISPERSION

$$= \frac{\frac{1}{2}(\text{Range})}{\text{Mid} - \text{Range}}$$

$$= \frac{\frac{X_m - X_0}{2}}{\frac{X_m + X_0}{2}} = \frac{X_m - X_0}{X_m + X_0}$$

This is a pure (i.e. dimensionless) number and is used for the purposes of COMPARISON. (This is so because a pure number can be compared with another pure number.)

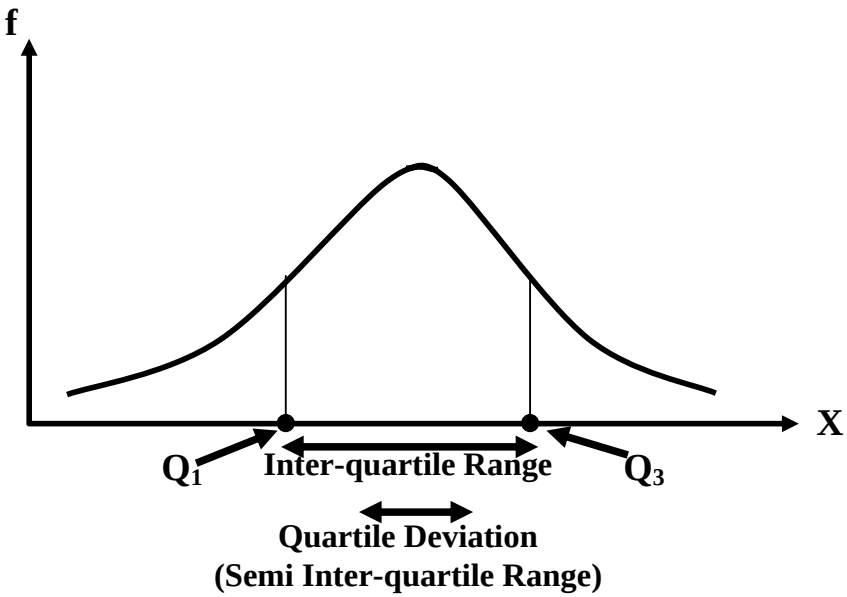
For example, if the coefficient of dispersion for one data-set comes out to be 0.6 whereas the coefficient of dispersion for another data-set comes out to be 0.4, then it is obvious that there is greater amount of dispersion in the first data-set as compared with the second.

QUARTILE DEVIATION

The quartile deviation is defined as half of the difference between the third and first quartiles i.e.

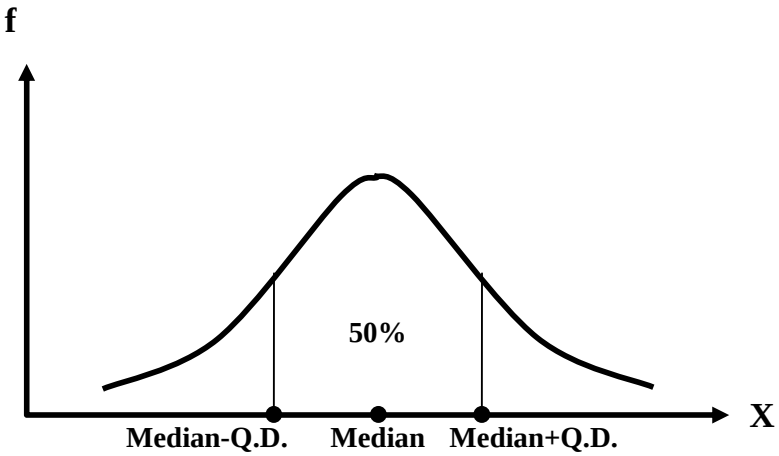
$$Q.D. = \frac{Q_3 - Q_1}{2}$$

It is also known as **semi-interquartile range.** Let us now consider the graphical representation of the quartile deviation:



Although simple to compute, it is NOT an extremely satisfactory measure of dispersion because it takes into account the spread of only two values of the variable round the median, and this gives no idea of the rest of the dispersion within the distribution.

The quartile deviation has an attractive feature that the range "Median + Q.D." contains approximately 50% of the data. This is illustrated in the figure given below:



Let us now apply the concept of quartile deviation to the following example:

EXAMPLE

The shareholding structure of two companies is given below:

	Company X	Company Y
1st quartile	60 shares	165 shares
Median	185 shares	185 shares
3rd quartile	270 shares	210 shares

The quartile deviation for company X is

$$\frac{270-60}{2}=105 \quad \text{Shares}$$

For company Y, it is

$$\frac{210-165}{2}=22 \text{Shares}$$

A comparison of the above two results indicate that there is a considerable concentration of shareholders about the MEDIAN number of shares in company Y, whereas in company X, there does not exist this kind of a concentration around the median. (In company X, there is approximately the SAME numbers of small, medium and large shareholders.)

From the above example, it is obvious that the larger the quartile deviation, the greater is the scatter of values within the series. The quartile deviation is superior to range as it is not affected by extremely large or small observations. It is simple to understand and easy to calculate.

The mean deviation can also be viewed in another way: It is the arithmetic mean of the deviations of the first and third quartiles round the median i.e.

$$\frac{(M-Q_1)+(Q_3-M)}{2}$$
$$=\frac{M-Q_1+Q_3-M}{2}$$
$$=\frac{Q_3-Q_1}{2}$$

Because of what has been just explained, the quartile deviation is regarded as that measure of dispersion which is associated with the median. As such, the quartile deviation should always be employed to indicate dispersion when the median has been adopted as the most appropriate average.

The quartile deviation is also an *absolute* measure of dispersion. Its *relative* measure called the CO-EFFICIENT OF QUARTILE DEVIATION or of Semi-Inter-quartile Range, is defined by the relation:

COEFFICIENT OF QUARTILE DEVIATION

$$=\frac{Quartile\ \ Deviation}{Mid-Quartile\ \ Range}$$
$$=\frac{\frac{Q_3-Q_1}{2}}{\frac{Q_3+Q_1}{2}}=\frac{Q_3-Q_1}{Q_3+Q_1},$$

The Coefficient of Quartile Deviation is a pure number and is used for *COMPARING* the variation in two or more sets of data.

The next two measures of dispersion to be discussed are the Mean Deviation and the Standard Deviation. In this regard, the first thing to note is that, whereas the range as well as the quartile deviation are two such measures of dispersion which are NOT based on all the values, the mean deviation and the standard deviation are two such measures of dispersion that involve each and every data-value in their computation.

The range measures the dispersion of the data-set around the *mid-range*, whereas the quartile deviation measures the dispersion of the data-set around the *median*.

How are we to decide upon the amount of dispersion round the *arithmetic mean*?

It would seem reasonable to compute the *DISTANCE* of each observed value in the series from the arithmetic mean of the series.

But the problem is that the sum of the deviations of the values from the mean is ZERO! (No matter what the amount of dispersion in a data-set is, this quantity will *always* be zero, and hence it cannot be used to measure the dispersion in the data-set.)

Then, the question arises, 'HOW will we be able to measure the dispersion present in our data-set?' In an attempt to answer this question, we might look at the numerical differences between the mean and the data values WITHOUT considering whether these are positive or negative. By ignoring the sign of the deviations we will achieve a NON-ZERO sum, and averaging these absolute differences, again, we obtain a non-zero quantity which can be used as a measure of dispersion. (The larger this quantity, the greater is the dispersion in the data-set).

This quantity is known as the **MEAN DEVIATION.**

Let us denote these absolute differences

by 'modulus of d'

or 'mod d'. Then, the mean deviation is given by

MEAN DEVIATION

$$M.D. = \frac{\sum |d|}{n}$$

As the absolute deviations of the observations from their mean are being averaged, therefore the complete name of this measure is Mean Absolute Deviation --- but generally, it is simply called "Mean Deviation". In the next lecture, this concept will be discussed in detail. (The case of raw data as well as the case of grouped data will be considered.)Next, we will discuss the most important and the most widely used measure of dispersion i.e. the Standard Deviation.

LECTURE NO. 11

- Mean Deviation
- Standard Deviation and Variance
- Coefficient of variation

First, we will discuss it for the case of raw data, and then we will go on to the case of a frequency distribution. The first thing to note is that, whereas the range as well as the quartile deviation are two such measures of dispersion which are NOT based on all the values, the mean deviation and the standard deviation are two such measures of dispersion that involve each and every data-value in their computation.

You must have noted that the range was measuring the dispersion of the data-set around the mid-range, whereas the quartile deviation was measuring the dispersion of the data-set around the median.

How are we to decide upon the amount of dispersion round the arithmetic mean? It would seem reasonable to compute the DISTANCE of each observed value in the series from the arithmetic mean of the series.

Let us do this for a simple data-set shown below:

THE NUMBER OF FATALITIES IN MOTORWAY ACCIDENTS IN ONE WEEK

Day	Number of fatalities X
Sunday	4
Monday	6
Tuesday	2
Wednesday	0
Thursday	3
Friday	5
Saturday	8
Total	28

Let us do this for a simple data-set shown below:

THE NUMBER OF FATALITIES IN MOTORWAY ACCIDENTS IN ONE WEEK

Day	Number of fatalities X
Sunday	4
Monday	6
Tuesday	2
Wednesday	0
Thursday	3
Friday	5
Saturday	8
Total	28

The arithmetic mean number of fatalities per day is

$$\bar{X} = \frac{\sum X}{n} = \frac{28}{7} = 4$$

In order to determine the distances of the data-values from the mean, we subtract our value of the arithmetic mean from each daily figure, and this gives us the deviations that occur in the third column of the table below

Day	Number of fatalities X	$X - \bar{X}$
Sunday	4	0
Monday	6	+ 2
Tuesday	2	– 2
Wednesday	0	– 4
Thursday	3	– 1
Friday	5	+ 1
Saturday	8	+ 4
TOTAL	28	0

The deviations are negative when the daily figure is less than the mean (4 accidents) and positive when the figure is higher than the mean. It does seem, however, that our efforts for computing the dispersion of this data set have been in vain, for we find that the total amount of dispersion obtained by summing the $(x - \bar{x})$ column comes out to be zero! In fact, this should be no surprise, for it is a basic property of the arithmetic means that: The sum of the deviations of the values from the mean is zero. The question arises:

How will we measure the dispersion that is actually present in our data-set?

Our problem might at first sight seem irresolvable, for by this criterion it appears that no series has any dispersion. Yet we know that this is absolutely incorrect, and we must think of some other way of handling this situation. Surely, we might look at the numerical difference between the mean and the daily fatality figures without considering whether these are positive or negative. Let us denote these absolute differences by 'modulus of d' or 'mod d'.

This is evident from the third column of the table below

X	$X - \bar{X} = d$	$\lvert d \rvert$
4	0	0
6	2	2
2	–2	2
0	–4	4
3	–1	1
5	1	1
8	4	4
	Total	14

By ignoring the sign of the deviations we have achieved a non-zero sum in our second column. Averaging these absolute differences, we obtain a measure of dispersion known as the mean deviation.

In other words, the mean deviation is given by the formula:

MEAN DEVIATION

$$M.D. = \frac{\sum |d_i|}{n}$$

As we are averaging the absolute deviations of the observations from their mean, therefore the complete name of this measure is mean absolute deviation --- but generally we just say "mean deviation". Applying this formula in our example, we find that, the mean deviation of the number of fatalities is

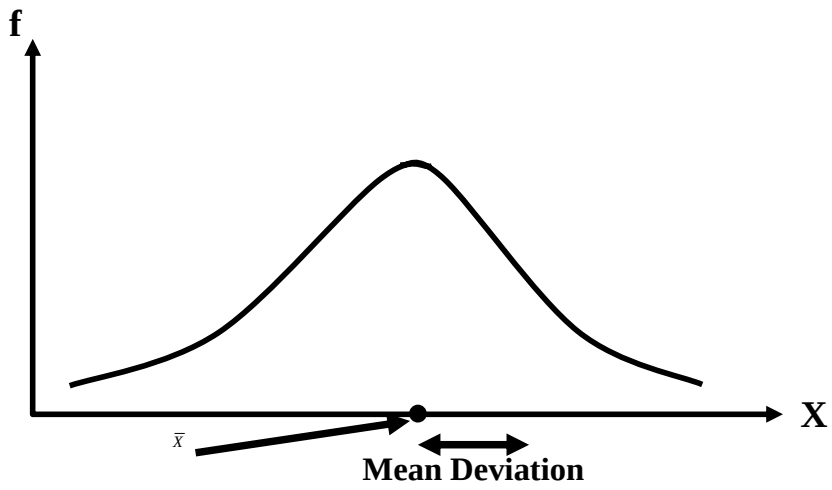
$$M.D. = \frac{14}{7} = 2.$$

The formula that we have just considered is valid in the case of raw data. In case of grouped data i.e. a frequency distribution, the formula becomes

MEAN DEVIATION FOR GROUPED DATA

$$M.D. = \frac{\sum f_i \left| x_i - \bar{x} \right|}{n} = \frac{\sum f_i \left| d_i \right|}{n}$$

As far as the graphical representation of the mean deviation is concerned, it can be depicted by a horizontal line segment drawn below the X-axis on the graph of the frequency distribution, as shown below



The approach which we have adopted in the concept of the mean deviation is both quick and simple. But the problem is that we introduce a kind of artificiality in its calculation by ignoring the algebraic signs of the deviations.
In problems involving descriptions and comparisons alone, the mean deviation can validly be applied; but because the negative signs have been discarded, further theoretical development or application of the concept is impossible.

Mean deviation is an absolute measure of dispersion. Its relative measure, known as the co-efficient of mean deviation, is obtained by dividing the mean deviation by the average used in the calculation of deviations i.e. the arithmetic mean. Thus

CO-EFFICIENT OF M.D

Sometimes, the mean deviation is computed by averaging the absolute deviations of the data-values from the median i.e.

$$\text{Mean deviation} = \frac{\sum \left| x - \tilde{x} \right|}{n}$$

And when will we have a situation when we will be using the median instead of the mean? As discussed earlier, the median will be more appropriate than the mean in those cases where our data-set contains a few very high or very low values. In such a situation, the coefficient of mean deviation is given by:

Co-efficient of M.D:

$$= \frac{M.D.}{\text{Median}} \qquad = \frac{M.D.}{\text{Mean}}$$

Let us now consider the ***standard deviation*** --- that statistic which is the most important and the most widely used measure of dispersion.
The point that made earlier that from the mathematical point of view, it is not very preferable to take the absolute values of the deviations, *This problem is overcome by computing the standard deviation.*
In order to compute the standard deviation, rather than taking the absolute values of the deviations, we square the deviations.

Averaging these squared deviations, we obtain a statistic that is known as the variance.

VARIANCE

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

Let us compute this quantity for the data of the above example.
Our X-values were:

X
4
6
2
0
3
5
8

Taking the deviations of the X-values from their mean, and then squaring these deviations, we obtain:

X	$(x - \bar{x})$	$(x - \bar{x})^2$
4	0	0
6	+ 2	4
2	– 2	4
0	– 4	16
3	– 1	1
5	+ 1	1
8	+ 4	16
		42

Obviously, both (– 2)2 and (2)2 equal 4, both (– 4)2 and (4)2 equal 16, and both (– 1)2 and (1)2 = 1
Hence $\sum(x - \bar{x})^2 = 42$ is now positive, and this positive value has been achieved without 'bending' the rules of mathematics. Averaging these squared deviations, the variance is given by:

Variance:

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

$$= \frac{42}{7} = 6$$

The variance is frequently employed in statistical work, but it should be noted that the figure achieved is in 'squared' units of measurement.
In the example that we have just considered, the variance has come out to be "6 squared fatalities", which does not seem to make much sense! In order to obtain an answer which is in the original unit of measurement, we take the positive square root of the variance. The result is known as the standard deviation.

STANDARD DEVIATION

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

Hence, in this example, our standard deviation has come out to be 2.45 fatalities.

In computing the standard deviation (or variance) it can be tedious to first ascertain the arithmetic mean of a series, then subtract it from each value of the variable in the series, and finally to square each deviation and then sum. It is very much more straight-forward to use the short cut formula given below:

SHORT CUT FORMULA FOR THE STANDARD DEVIATION

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

In order to apply the short cut formula, we require only the aggregate of the series ($\sum x$) and the aggregate of the squares of the individual values in the series ($\sum x2$).
In other words, only two columns of figures are called for. The number of individual calculations is also considerably reduced, as seen below:

	X	X^2
	4	16
	6	36
	2	4
	0	0
	3	9
	5	25
	8	64
Total	28	154

Therefore

$$S = \sqrt{\left\{\frac{154}{7} - \left(\frac{28}{7}\right)^2\right\}} = \sqrt{(22 - 16)}$$

$$= \sqrt{6} = 2.45 \text{ fatalities}$$

The formulae that we have just discussed are valid in case of raw data. In case of grouped data i.e. a frequency distribution, each squared deviation round the mean must be multiplied by the appropriate frequency figure i.e.

STANDARD DEVIATION IN CASE OF GROUPED DATA

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

And the short cut formula in case of a frequency distribution is:

SHORT CUT FORMULA OF THE STANDARD DEVIATION IN CASE OF GROUPED DATA

$$S = \sqrt{\left\{\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2\right\}}$$

Which is again preferred from the computational standpoint
For example, the standard deviation life of a batch of electric light bulbs would be calculated as follows:

EXAMPLE

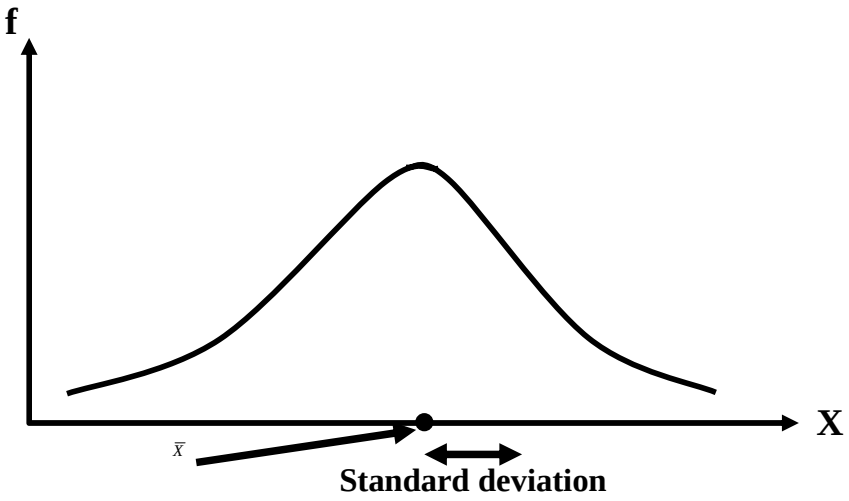
Life (in Hundreds of Hours)	No. of Bulbs f	Mid-point x	fx	fx^2
0 – 5	4	2.5	10.0	25.0
5 – 10	9	7.5	67.5	506.25
10 – 20	38	15.0	570.0	8550.0
20 – 40	33	30.0	990.0	29700.0
40 and over	16	50.0	800.0	40000.0
	100		2437.5	78781.25

Therefore, standard deviation:

$$S = \sqrt{\left\{\frac{78781.25}{100} - \left(\frac{2437.5}{100}\right)^2\right\}}$$

=13.9hundredhours
= 1390 hours

As far as the graphical representation of the standard deviation is concerned, a horizontal line segment is drawn below the X-axis on the graph of the frequency distribution --- just as in the case of the mean deviation.



The standard deviation is an absolute measure of dispersion. Its relative measure called coefficient of standard deviation is defined as:

COEFFICIENT OF S.D

$$= \frac{\text{Standard Deviation}}{\text{Mean}}$$

And, multiplying this quantity by 100, we obtain a very important and well-known measure called the coefficient of variation.

COEFFICIENT OF VARIATION

$$C.V. = \frac{S}{\overline{X}} \times 100$$

As mentioned earlier, the standard deviation is expressed in absolute terms and is given in the same unit of measurement as the variable itself.

There are occasions, however, when this absolute measure of dispersion is inadequate and a relative form becomes preferable. For example, if a comparison between the variability of distributions with different variables is required, or when we need to compare the dispersion of distributions with the same variable but with very different arithmetic means. To illustrate the usefulness of the coefficient of variation, let us consider the following two examples.

EXAMPLE-1

Suppose that, in a particular year, the mean weekly earnings of skilled factory workers in one particular country were $ 19.50 with a standard deviation of $ 4, while for its neighboring country the figures were Rs. 75 and Rs. 28 respectively.

From these figures, it is not immediately apparent which country has the GREATER VARIABILITY in earnings. The coefficient of variation quickly provides the answer:

For country No. 1:

$$\frac{4}{19.5} \times 100 = 20.5 \text{ per cent,}$$

And for country No. 2:

$$\frac{28}{75} \times 100 = 37.3 \text{ per cent.}$$

From these calculations, it is immediately obvious that the spread of earnings in country No. 2 is greater than that in country No. 1, and the reasons for this could then be sought.

EXAMPLE-2:

The crop yield from 20 acre plots of wheat-land cultivated by ordinary methods averages 35 bushels with a standard deviation of 10 bushels. The yield from similar land treated with a new fertilizer averages 58 bushels, also with a standard deviation of 10 bushels. At first glance, the yield variability may seem to be the same, but in fact it has improved (i.e. decreased) in view of the higher average to which it relates.

Again, the coefficient of variation shows this very clearly:

Untreated land:

$$\frac{10}{35} \times 100 = 28.57 \text{ per cent}$$

Treated land:

$$\frac{10}{58} \times 100 = 17.24 \text{ per cent}$$

The coefficient of variation for the untreated land has come out to be 28.57 percent, whereas the coefficient of variation for the treated land is only 17.24 percent.

LECTURE NO. 12

- Chebychev's Inequality
- The Empirical Rule
- The Five-Number Summary

In the last lecture, we discussed the concept of standard deviation in quite a lot of detail.
It is an extremely important concept, and it is very important that we appreciate and understand its role in statistical analysis. We've seen that if we are comparing the variability of two samples selected from a population, the sample with the larger standard deviation is the more variable of the two.
Thus, we know how to interpret the standard deviation on a relative or comparative basis, but we haven't considered how it provides a measure of variability for a single sample.
To understand how the standard deviation provides a measure of variability of a data set, consider a specific data set and answer the following questions:

Question-1

'How many measurements are within 1 standard deviation of the mean?'

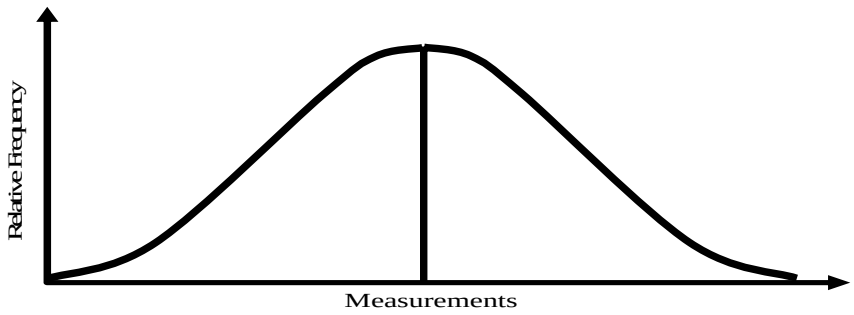
Question-2

'How many measurements are within 2 standard deviations?'
and so on.
For any specific data set, we can answer these questions by counting the number of measurements in each of the intervals. However, if we are interested in obtaining a general answer to these questions the problem is a bit more difficult. We will discuss to you two sets of answers to the questions of how many measurements fall within 1, 2, and 3 standard deviations of the mean. General answer to these questions the problem is a bit more difficult. The first, which applies to any set of data, is derived from a theorem proved by Russian mathematician, P.L. Chebychev (1821-1894).The second, which applies to mound-shaped, symmetric distributions of data, is based upon empirical evidence that has accumulated over the years. And this set of answers is valid and applicable even if our distribution is slightly skewed, Let us begin with the Chebychev's theorem.
Chebychev's Rule applies to any data set, regardless of the shape of the frequency distribution of the data.

CHEBYCHEV'S THEOREM

For any number k greater than 1, at least $1 - 1/k^2$ of the data-values fall within k standard deviations of the mean, i.e., within the interval $(\bar{X} - kS, \bar{X} + kS)$
This means that:
a) At least 1-1/22 = 3/4 will fall within 2 standard deviations of the mean, i.e. within the interval $(\bar{X} - 2S, \bar{X} + 2S)$.
b) At least 1-1/32=8/9 of the data-values will fall within 3 standard deviations of the mean, i.e. within the interval $(\bar{X} - 3S, \bar{X} + 3S)$
Because of the fact that Chebychev's theorem requires k to be greater than 1, therefore no useful information is provided by this theorem on the fraction of measurements that fall within 1 standard deviation of the mean, i.e. within the interval $(\bar{X}-S, \bar{X}+S)$.

Next, let us consider the ***Empirical Rule mentioned above.***
This is a rule of thumb that applies to data sets with frequency distributions that are mound-shaped and symmetric, as follows:



According to this empirical rule:

- Approximately 68% of the measurements will fall within 1 standard deviation of the mean, i.e. within the interval ($\bar{X} - S$, $\bar{X} + S$)
- Approximately 95% of the measurements will fall within 2 standard deviations of the mean, i.e. within the interval ($\bar{X} - 2S$, $\bar{X} + 2S$).
- Approximately 100% (practically all) of the measurements will fall within 3 standard deviations of the mean, i.e. within the interval ($\bar{X} - 3S$, $\bar{X} + 3S$).

Let us understand this point with the help of an example:

EXAMPLE

The 50 companies' percentages of revenues spent on R&D (i.e. Research and Development) are:

13.5	9.5	8.2	6.5	8.4	8.1	6.9	7.5	10.5	13.5
7.2	7.1	9.0	9.9	8.2	13.2	9.2	6.9	9.6	7.7
9.7	7.5	7.2	5.9	6.6	11.1	8.8	5.2	10.6	8.2
11.3	5.6	10.1	8.0	8.5	11.7	7.1	7.7	9.4	6.0
8.0	7.4	10.5	7.8	7.9	6.5	6.9	6.5	6.8	9.5

Calculate the proportions of these measurements that lie within the intervals $\bar{X} \pm S$, $\bar{X} \pm 2S$, and $\bar{X} \pm 3S$, and compare the results with the theoretical values. The mean and standard deviation of these data come out to be 8.49 and 1.98, respectively.

Mean:

$\bar{X} = 8.49$

Standard deviation:

$S = 1.98$

Hence

$(\bar{X} - S, \bar{X} + S)$

$= (8.49 - 1.98, 8.49 + 1.98)$

$= (6.51, 10.47)$

A check of the measurement reveals that 34 of the 50 measurements, or 68%, fall between 6.51and 10.47.

Similarly, the interval

$(\bar{X} - 2S, \bar{X} + 2S)$

$= (8.49 - 3.96, 8.49 + 3.96)$

$= (4.53, 12.45)$

Contains 47 of the 50 measurements, i.e. 94% of the data-values

Finally, the 3-standard deviation interval around $\bar{X}$, i.e. $(\bar{X} - 3S, \bar{X} + 3S)$

$= (8.49 - 5.94, 8.49 + 5.94)$

$= (2.55, 14.43)$ contains all, or 100%, of the measurements.

In spite of the fact that the distribution of these data is skewed to the right, the percentages of data-values falling within 1, 2, and 3 standard deviations of the mean are remarkably close to the theoretical values (68%, 95%, and 100%) given by the Empirical Rule.

The fact of the matter is that, unless the distribution is extremely skewed, the mound-shaped approximations will be reasonably accurate. Of course, no matter what the shape of the distribution, Chebychev's Rule, assures that at least 75% and at least 89% (8/9) of the measurements will lie within 2 and 3 standard deviations of the mean, respectively.

In this example, 94% of the values are lying inside the interval $\bar{X} + 2S$, and this percentage IS greater than 75%.

Similarly,100% of the values are lying inside the interval $\bar{X} + 3S$, and this percentage IS greater than 89%.

But, before we discuss all these new concepts, let us revise the concept of the Chebychev's Inequality. In the last lecture, we noted that when all the values in a set of data are located near their mean, they exhibit a small amount of variation or dispersion.

And those sets of data in which some values are located far from their mean have a large amount of dispersion. Expressing these relationships in terms of the standard deviation, which measures dispersion, we can say that when the values of a set of data are concentrated near their mean, the standard deviation is small. And when the values of a set of data are scattered widely about the mean, the standard deviation is large. In exactly the same way, if the standard deviation computed from a set of data is large, the values from which it is computed are dispersed widely about their mean. A useful rule that illustrates the relationship between dispersion and standard deviation is given by *Chebychev's theorem*, named after the Russian mathematician P.L. Chebychev (1821-1894). This theorem enables us to calculate for any set of data the minimum proportion of values that can be expected to lie within a specified number of standard deviations of the mean.

The theorem tells us that at least 75% of the values in a set of data can be expected to fall within two standard deviations of the mean, at least 89% (8/9) within three standard deviations of the mean, and at least 94% (15/16) within four standard deviations of the mean.
In general, Chebychev's theorem may be stated as follows:

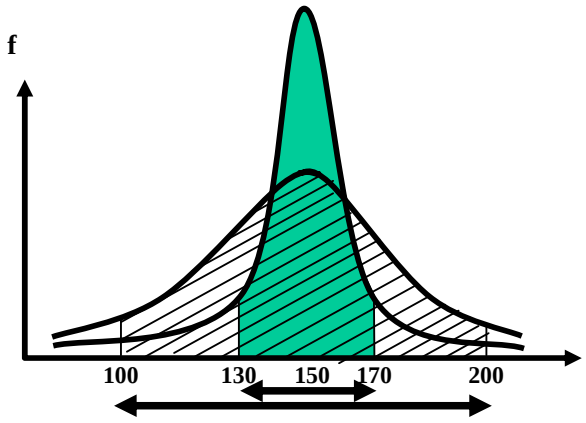
CHEBYCHEV'S THEOREM

Given a set of n observations x1, x2, x3… xn on the variable X, the probability is at least (1 – 1/k2) that X will take on a value within k standard deviations of the mean of the set of observations (where k > 1).
Chebychev's theorem is applicable to any set of observations, so we can use it for either samples or populations. Let us now see how we can suppose that a set of data has a mean of 150 and a standard deviation of 25. Putting k = 2 in the Chebychev's theorem, at least 1 – 1/ (2)2 = 75% of the data-values will take on a value within two standard deviations of the mean.
Apply it in practice.
Since the standard deviation is 25, hence 2(25) = 50, and at least 75% of the data-values will take on a value between 150 – 50 = 100 and 150 + 50 = 200. Consequently, we can say that we can expect at least 75% of the values to be between 100 and 200.By similar calculations we find that we can expect at least 89% to be between 75 and 225, and at least 96% to be between 25 and 275.
(The last statement has been made by putting k = 5 in the formula 1 - 1/k2)
Suppose that another set of data has the same mean as before, i.e. 150, but a standard deviation of 10. Applying Chebychev's theorem, for this set of data we can expect at least 75% of the values to be between 130 and 170, at least 89% to be between 120 and 180, and at least 96% to be between 100 and 200.
The above results are summarized in the following table:

PERCENTAGE OF DATA	FOR DATA-SET NO. 1	FOR DATA-SET NO. 2
At least 75 %	Lies Between 100 & 200	Lies Between 130 & 170
At least 89 %	Lies Between 75 & 225	Lies Between 120 & 180
At least 96 %	Lies Between 25 & 275	Lies Between 100 & 200

Thus the intervals computed for the latter set of data are all narrower than those for the former.
For two symmetric, hump-shaped distributions having the same mean, this point is depicted in the following diagram:

THE SYMMETRIC CURVE



Therefore, we see that for a set of data with a small standard deviation, a larger proportion of the values will be concentrated near the mean than for a set of data with a large standard deviation.
A limitation of the Chebychev's theorem is that it gives no information at all about the probability of observing a value within one standard deviation of the mean, since 1 – 1/k2 = 0 when k = 1. Also, it should be noted that the Chebychev's theorem provides weak information for our variable of interest. For many random variables, the probability of observing a value within 2 standard deviations of the mean is far greater than 1 – 1/22 = 0.75.
In this way, the Chebychev's theorem and the Empirical Rule play an important role in understanding the nature and importance of the standard deviation as a measure of dispersion.

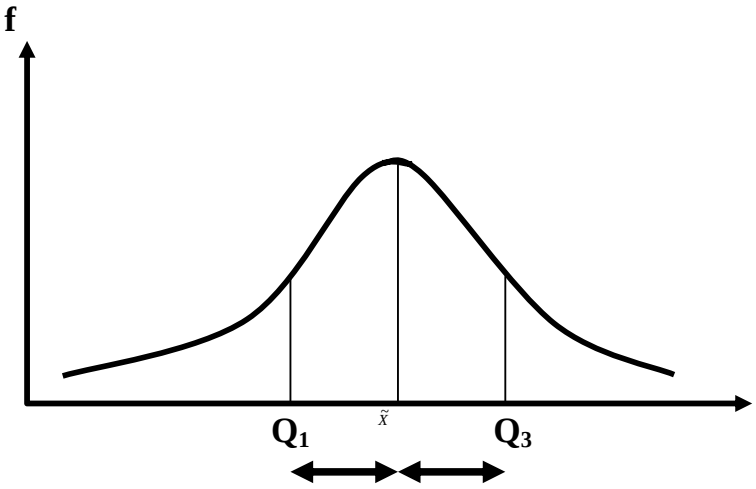
The next topic of today's lecture is the five-number summary. (Now that we have studied the three major properties of numerical data (i.e. central tendency, variation, and shape), it is important that we identify and describe the major features of the data in a summarized format.)
One approach to this "exploratory data analysis" is to develop a five-number summary.

FIVE-NUMBER SUMMARY

A five-number summary consists of X0,Q1, Median, Q3, and Xm ; It provides us quite a good idea about the shape of the distribution. If the data were perfectly symmetrical, the following would be true:

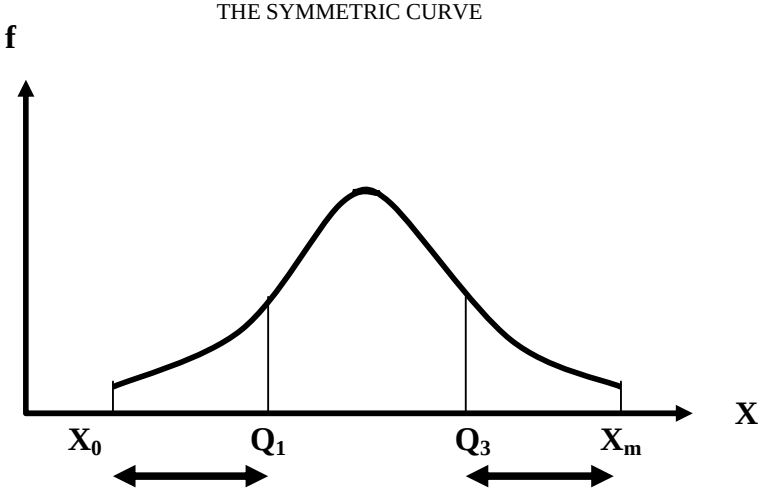
1. The distance from Q1 to the median would be equal to the distance from the median to Q3:

THE SYMMETRIC CURVE



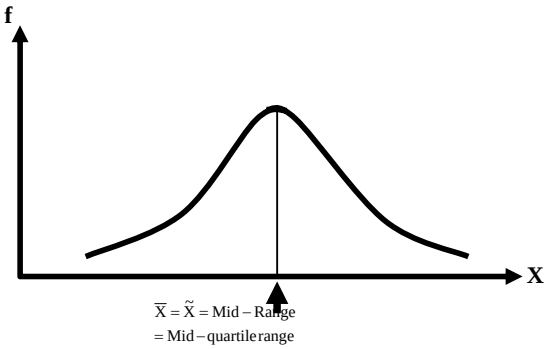
3. The distance from X_0 to Q_1 would be equal to the distance from Q_3 to $X_{m.}$

THE SYMMETRIC CURVE



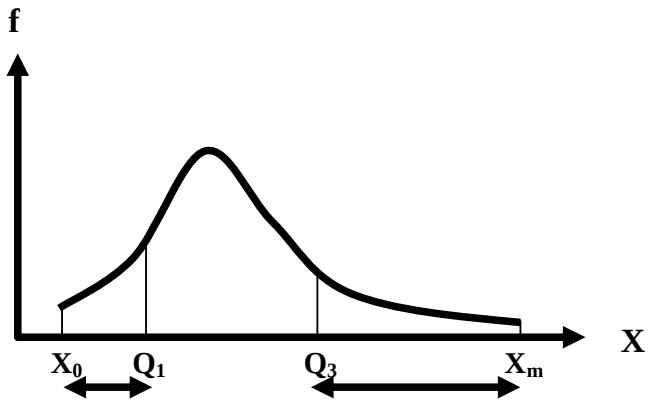
3. The median, the mid-quartile range, and the midrange would all be equal. All these measures would also be equal to the arithmetic mean of the data:

THE SYMMETRIC CURVE



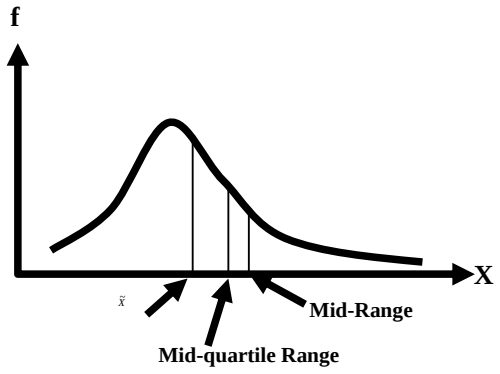
On the other hand, for non-symmetrical distributions, the following would be true:
1. In right-skewed distributions the distance from Q_3 to X_m greatly exceeds the distance from X_0 to Q_1.

THE POSITIVELY SKEWED CURVE



2. in right-skewed distributions,
median < mid-quartile range < midrange:

THE POSITIVELY SKEWED CURVE



Similarly, in left-skewed distributions, the distance from X0 to Q1 greatly exceeds the distance from Q3 to Xm.Also, in left-skewed distributions, midrange < mid-quartile range < median.
Let us try to understand this concept with the help of an example

EXAMPLE

Suppose that a study is being conducted regarding the annual costs incurred by students attending public versus private colleges and universities in the United States of America. In particular, suppose, for exploratory purposes, our sample consists of 10 Universities whose athletic programs are members of the 'Big Ten' Conference. The annual costs incurred for tuition fees, room, and board at 10 schools belonging to Big Ten Conference are given as follows:

Name of University	Annual Costs (in $000)
Indiana University	15.6
Michigan State University	17.0
Ohio State University	15.2
Pennsylvania State University	16.4
Purdue University	15.2
University of Illinois	15.4
University of Iowa	13.0
University of Michigan	23.1
University of Minnesota	14.3
University of Wisconsin	14.9

If we wish to state the five-number summary for these data, the first step will be to arrange our data-set in ascending order:
Ordered Array:

$X_0 = 13.0$	14.3	14.9	15.2	15.2	15.4	15.6	16.4	17.0	$X_m = 23.1$

And if we carry out the relevant computations, we find that:

- The median for this data comes out to be 15.30 thousand dollars.
- The first quartile comes out to be 14.90 thousand dollars, and
- The third quartile comes out to be 16.40 thousand dollars.

Therefore, the five-number summary for this data-set is:
The Five-Number Summary:

X_0	Q_1	$\tilde{X}$	Q_3	X_m
13.0	**14.9**	**15.3**	**16.4**	**23.1**

If we apply the rules that I am conveyed to you a short while ago, it is clear that the annual cost data for our sample are right-skewed. We come to this conclusion because of two reasons:

- The distance from Q3 to Xm (i.e., 6.7) greatly exceeds the distance from X0 to Q1 (i.e., 1.9).
- If we compare the median (which is 15.3), the mid-quartile range (which is 15.65), and the midrange (which is 18.05), we observe that the median < the mid-quartile range < the midrange.

Both these points clearly indicate that our distribution is positively skewed.
The gist of the above discussion is that the five-number summary is a simple yet effective way of determining the shape of our frequency distribution --- without actually drawing the graph of the frequency distribution.

LECTURE NO. 13

- Box and Whisker Plot
- Pearson's Coefficient of Skewness

Prior to discussing the THE BOX-AND-WHISKER PLOT, let us review the concept of THE FIVE-NUMBER SUMMARY. As indicated in the last lecture, once we have studied the three major properties of numerical data (i.e. central tendency, variation, and shape), it is important that we identify and describe the major features of the data in a SUMMARIZED format. One way of doing this is to develop a five-number summary.

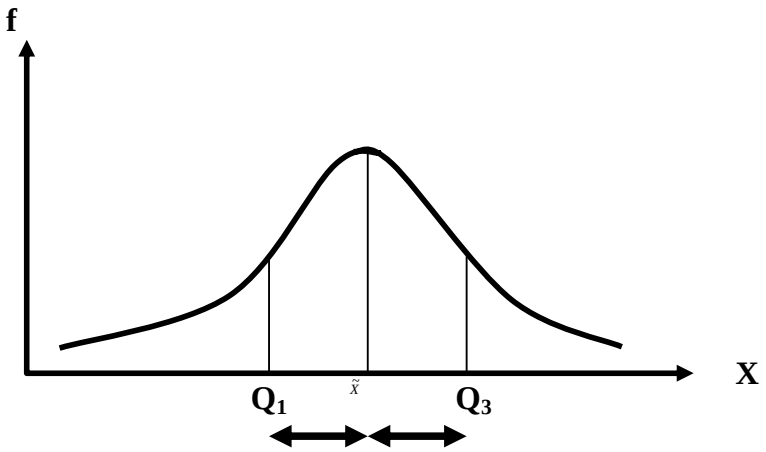
FIVE-NUMBER SUMMARY

A five-number summary consists of X_0, Q_1, Median, Q_3; Xm.It provides us a better idea as to the SHAPE of the distribution, as explained below:

If the data were perfectly symmetrical, the following would be true:

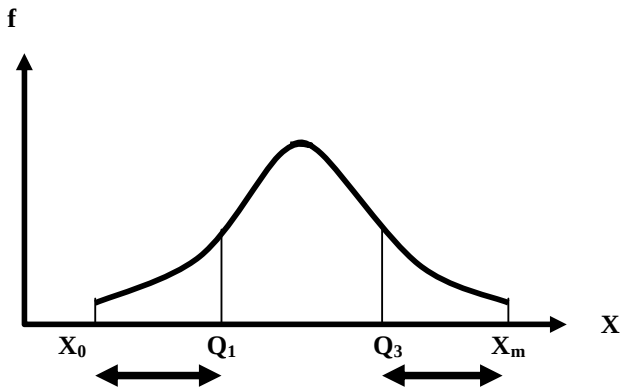
1. The distance from Q1 to the median would be equal to the distance from the median to Q3, as shown below:

THE SYMMETRIC CURVE



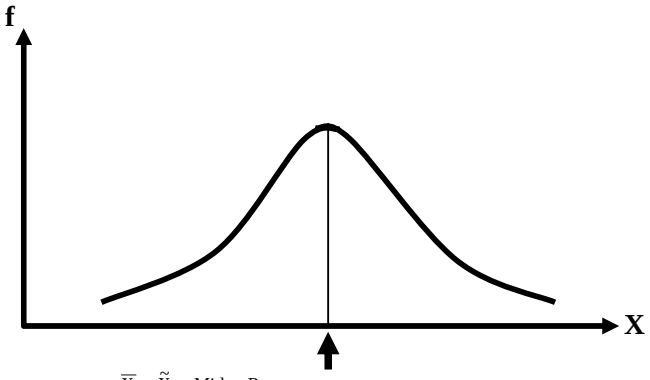
2. The distance from X0 to Q1 would be equal to the distance from Q3 to Xm, as shown below:

THE SYMMETRIC CURVE



3. The median, the mid-quartile range, and the midrange would ALL be equal.
These measures would also be equal to the arithmetic mean of the data, as shown below:

THE SYMMETRIC CURVE

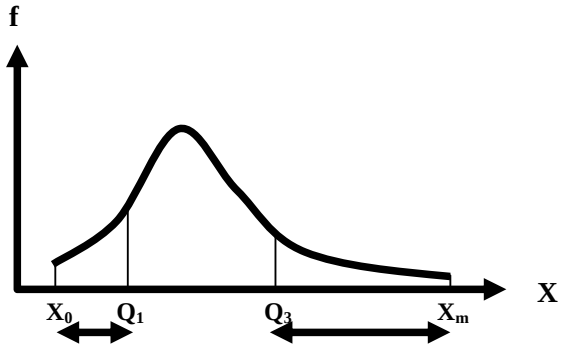


$\overline{X} = \tilde{X}$ = Mid – Range
= Mid – quartile range

On the other hand, for non-symmetrical distributions, the following would be true:

1. In right-skewed (positively-skewed) distributions the distance from Q3 to Xm greatly EXCEEDS the distance from X0 to Q1, as shown below:

THE POSITIVELY SKEWED CURVE

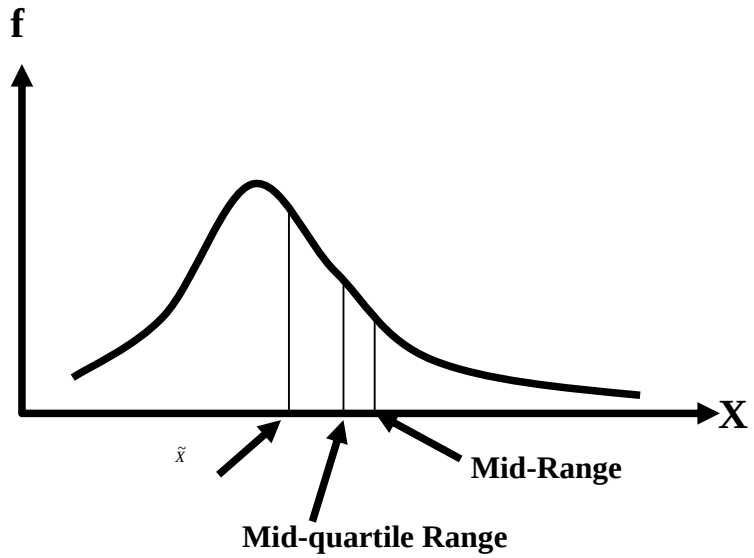


2. In right-skewed distributions,

median < mid-quartile range < midrange

This is indicated in the following figure:

THE POSITIVELY SKEWED CURVE



Similarly, in left-skewed distributions, the distance from X0 to Q1 greatly exceeds the distance from Q3 to Xm.Also, in left-skewed distributions, midrange < mid-quartile range < median.
Let us try to understand this concept with the help of an example:

EXAMPLE

Suppose that a study is being conducted regarding the annual costs incurred by students attending public versus private colleges and universities in the United States of America. In particular, suppose, for exploratory purposes, our sample consists of 10 Universities whose athletic programs are members of the 'Big Ten' Conference? The annual costs incurred for tuition fees, room, and board at 10 schools belonging to Big Ten Conference are given in the following table; state the five-number summary for these data.

Annual Costs Incurred on Tuition Fees, etc.

Name of University	Annual Costs (in $000)
Indiana University	15.6
Michigan State University	17.0
Ohio State University	15.2
Pennsylvania State University	16.4
Purdue University	15.2
University of Illinois	15.4
University of Iowa	13.0
University of Michigan	23.1
University of Minnesota	14.3
University of Wisconsin	14.9

SOLUTION:
For our sample, the ordered array is

X_0 = 13.0	14.3	14.9	15.2	15.2	15.4	15.6	16.4	17.0	X_m = 23.1

The median for this data comes out to be 15.30 thousand dollars. The first quartile comes out to be 14.90 thousand dollars, and the third quartile comes out to be 16.40 thousand dollars.
Therefore, the five-number summary is:

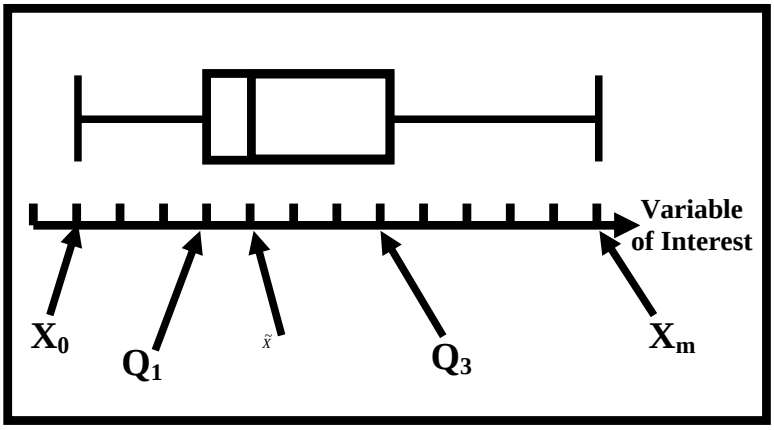
X_0	Q_1	$\tilde{X}$	Q_3	X_m
13.0	14.9	15.3	16.4	23.1

We may now use the five-number summary to study the *shape* of this distribution:
We notice that
1. The distance from Q3 to Xm (i.e., 6.7) greatly exceeds the distance from X0 to Q1 (i.e., 1.9).
2. If we compare the median (which is 15.3), the mid-quartile range (which is 15.65), and the midrange (which is 18.05), we observe that the median < the mid-quartile range < the midrange.
Hence, from the preceding rules, it is clear that the annual cost data for our sample are *right-skewed*. The gist of the above discussion is that the five-number summary is a SIMPLE yet effective way of determining the shape of our frequency distribution --- WITHOUT actually drawing the graph of the frequency distribution. The concept of the five number summary is directly linked with the concept of the box and whisker plot:

BOX AND WHISKER PLOT

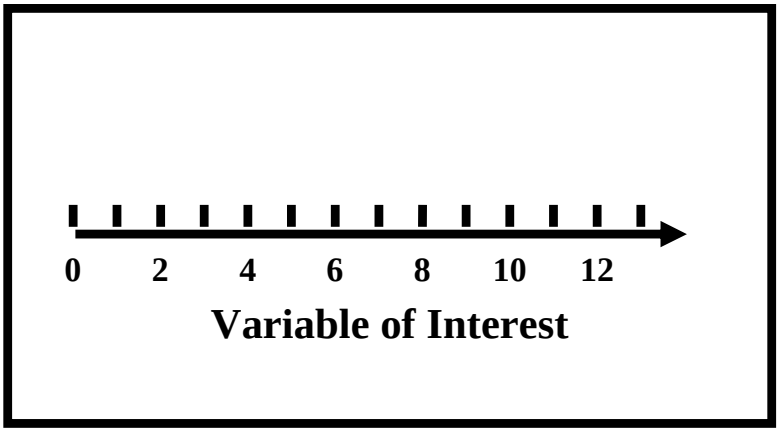
In its simplest form, a box-and-whisker plot provides a graphical representation of the data through its five-number summary.

Box and Whisker Plot

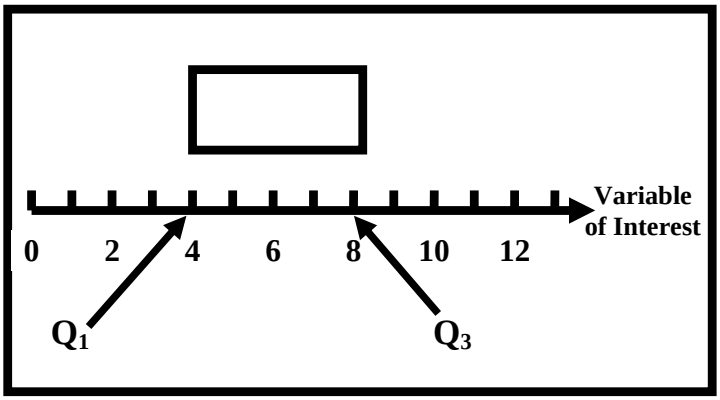


To construct a box-and-whisker plot, we proceed as follows:
Steps involved in the construction of the Box and Whisker Plot:

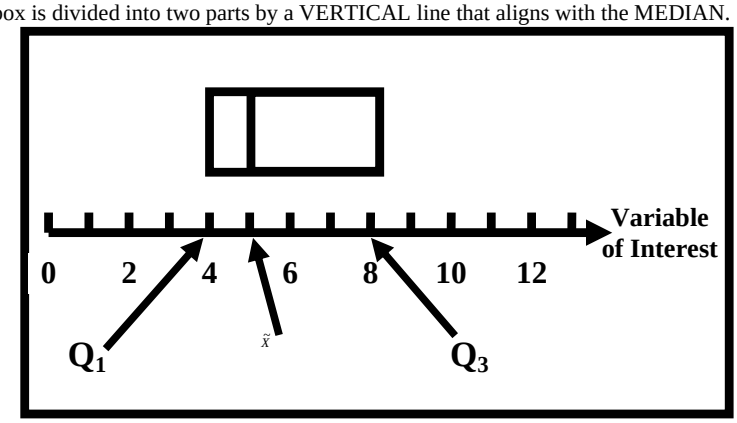
1. The variable of interest in represented on the horizontal axis.



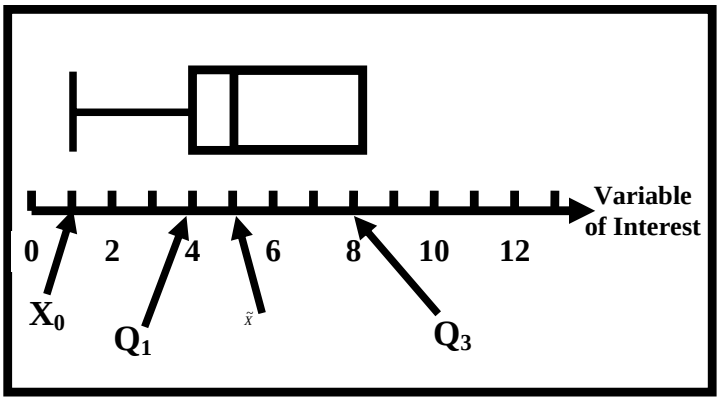
2. A BOX is drawn in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile Q1 and the right end of the box is aligned with the third quartile Q3.



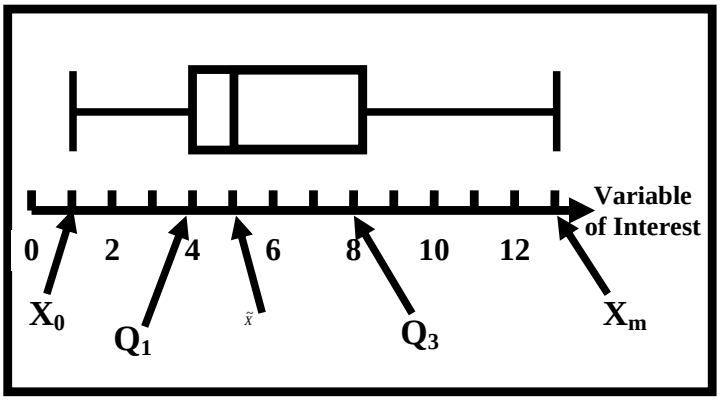
3. The box is divided into two parts by a VERTICAL line that aligns with the MEDIAN.



4. A line, called a whisker, is extended from the LEFT end of the box to a point that aligns with X0, the smallest measurement in the data set.



5. Another line, or whisker, is extended from the RIGHT end of the box to a point that aligns with the LARGEST measurement in the data set.



Let us understand the construction of the box-and-whisker plot with reference to an example:

EXAMPLE

The following table shows the downtime, in hours, recorded for 30 machines owned by a large manufacturing company. The period of time covered was the same for all machines.

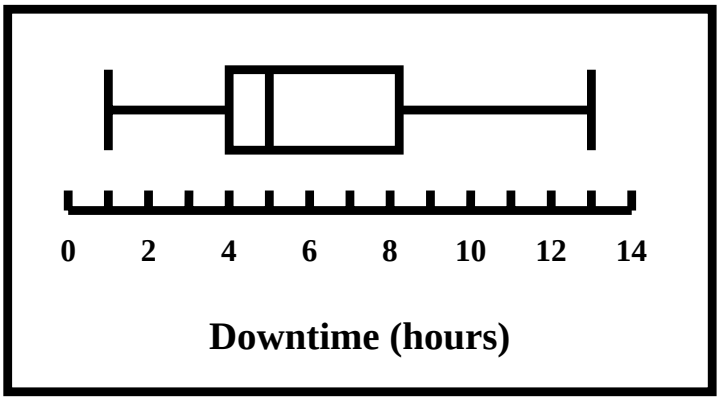
DOWNTIME IN HOURS OF 30 MACHINES

4	4	1	4	1	4
6	10	5	5	8	2
1	6	10	1	13	5
8	4	3	9	4	9
1	4	4	11	8	9

In order to construct a box-and-whisker plot for these data, we proceed as follows:
First of all, we determine the two extreme values in our data-set:
The smallest and largest values are X0 = 1 and Xm = 13, respectively.
As far as the computation of the quartiles is concerned, we note that, in this example, we are dealing with raw data.

The first quartile is the (30 + 1)/4 = 7.75th ordered measurement and is equal to 4.
The median is the (30 + 1)/2 = 15.5th measurement, or 5,and
The third quartile is the 3(30 + 1)/4 = 23.25th ordered measurement, which is 8.25.
As a result, we obtain the following box and whisker plot:

Box and Whisker Plot



INTERPRETATION OF THE BOX AND WHISKER PLOT

With regard to the *interpretation* of the Box and Whisker Plot, it should be noted that, by looking at a box-and-whisker plot, one can quickly form an impression regarding the amount of SPREAD, location of CONCENTRATION, and SYMMETRY of our data set.
A glance at the box and whisker plot of the example that we just considered reveals that:

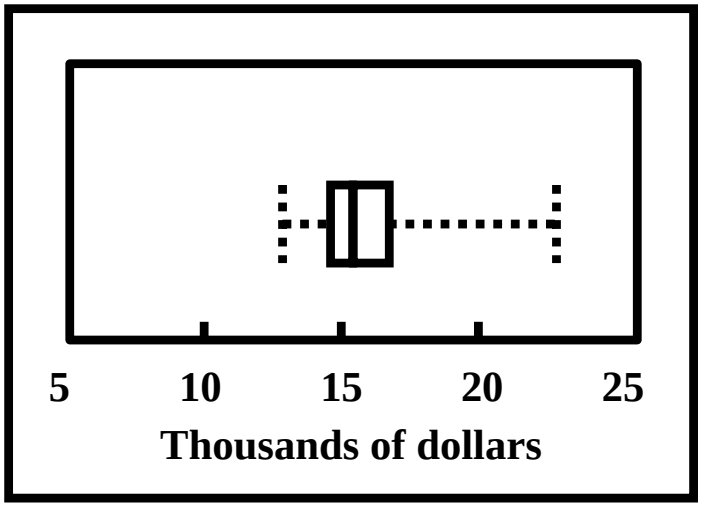
- 50% of the measurements are between 4 and 8.25.
- The median is 5, and the range is 12. and, most importantly:
- Since the median line is closer to the *left* end of the box, hence the data are SKEWED to the RIGHT.(The fundamental point is that in a perfectly symmetrical data set, the median line will be EXACTLY HALFWAY between the two ends of the box, and in a data set that is skewed to the LEFT, the median line will be CLOSER TO THE RIGHT END of the box.)

Let us consolidate all the above ideas by going back to the example of the Big Ten Universities in which the annual costs incurred for tuition fees, room, and board at 10 schools belonging to Big Ten Conference were given as follows:

Name of University	Annual Costs (in $000)
Indiana University	15.6
Michigan State University	17.0
Ohio State University	15.2
Pennsylvania State University	16.4
Purdue University	15.2
University of Illinois	15.4
University of Iowa	13.0
University of Michigan	23.1
University of Minnesota	14.3
University of Wisconsin	14.9

As stated earlier, the Five-Number Summary of this data-set is :
For this data, the Box and Whisker Plot is of the form given below:

Box and Whisker Plot



As indicated earlier, the vertical line drawn within the box represents the location of the median value in the data; the vertical line at the LEFT side of the box represents the location of Q1, and the vertical line at the RIGHT side of the box represents the location of Q3. Therefore, the BOX contains the middle 50% of the observations in the distribution. The lower 25% of the data are represented by the whisker that connects the *left* side of the box to the location of the *smallest* value, X0, and the upper 25% of the data are represented by the whisker connecting the *right* side of the box to Xm.

INTERPRETATION OF THE BOX AND WHISKER PLOT

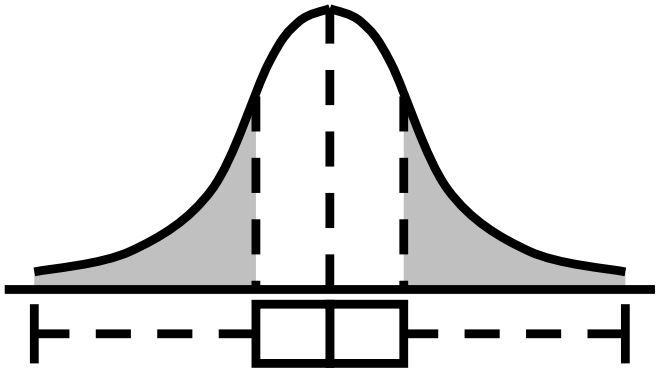
We note that (1) the vertical median line is CLOSER to the left side of the box, and (2) the left side whisker length is clearly SMALLER than the right side whisker length .Because of these observations, we The gist of the above discussion is that if the median line is at a greater distance from the left side of the box as compared with its distance from the right side of the box, our distribution will be skewed to the left.

In this situation, the whisker appearing on the left side of the box and whisker plot will be longer than the whisker of the right side. Conclude that the data-set of the annual costs is RIGHT-skewed.

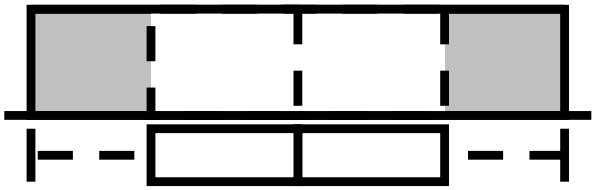
The gist of the above discussion is that if the median line is at a greater distance from the left side of the box as compared with its distance from the right side of the box, our distribution will be skewed to the left. In this situation, the whisker appearing on the left side of the box and whisker plot will be longer than the whisker of the right side. The Box and Whisker Plot comes under the realm of "exploratory data analysis" (EDA) which is a relatively new area of statistics. The following figures provide a comparison between the Box and Whisker Plot and the traditional procedures such as the frequency polygon and the frequency curve with reference to the SKEWNESS present in the data-set.

Four different types of hypothetical distributions are depicted through their box-and-whisker plots and corresponding frequency curves.

1) When a data set is perfectly symmetrical, as is the case in the following two figures, the mean, median, midrange, and mid-quartile range will be the SAME:



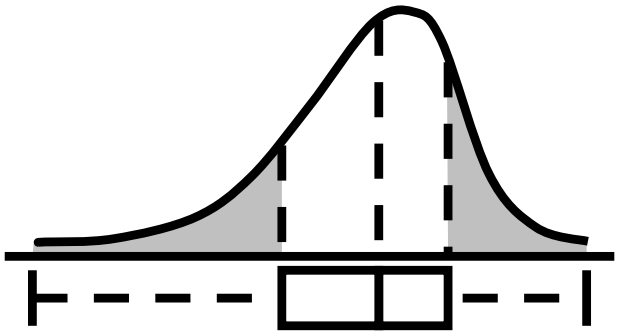
(a) Bell-shaped distribution



(b) Rectangular distribution

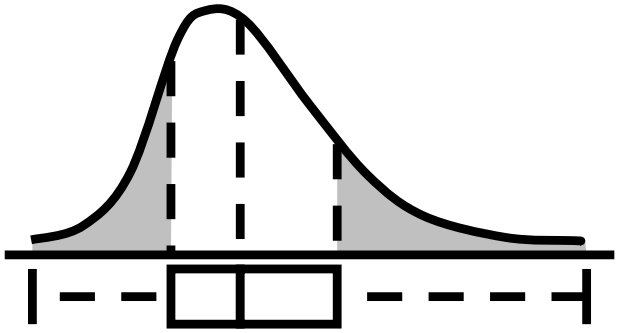
In ADDITION, the length of the left whisker will be equal to the length of the right whisker, and the median line will divide the box in HALF. (In *practice*, it is unlikely that we will observe a data set that is perfectly symmetrical. However, we should be able to state that our data set is *approximately* symmetrical *if* the lengths of the two whiskers are *almost* equal and the *median* line *almost* divides the box in HALF.)

2) When our data set is LEFT-skewed as in the following figure, the few small observations pull the midrange and mean toward the LEFT tail:



Left-skewed distribution

For this LEFT-skewed distribution, we observe that the skewed nature of the data set indicates that there is a HEAVY CLUSTERING of observations at the HIGH END of the scale (i.e., the RIGHT side).
75% of all data values are found between the left edge of the box (Q1) and the end of the right whisker (Xm). Therefore, the LONG left whisker contains the distribution of only the smallest 25% of the observations, demonstrating the distortion from symmetry in this data set.
3) If the data set is RIGHT-skewed as shown in the following figure, the few large observations PULL the midrange and mean toward the right tail.



Right-skewed distribution

For the right-skewed data set, the concentration of data points is on the LOW end of the scale (i.e., the left side of the box-and-whisker plot). Here, 75% of all data values are found between the beginning of the left whisker (X0) and the RIGHT edge of the box (Q3), and the remaining 25% of the observations are DISPERSED ALONG the LONG right whisker at the upper end of the scale. This brings us to the end of the discussion of the five number summary and the box and whisker plot.
Next, we discuss *another* way of determining the skewness of the data-set and that is the **PEARSON'S**

COEFFICIENT OF SKEWNESS

In this connection, the first thing to note is that, by providing information about the location of a series and the dispersion within that series it might appear that we have achieved a PERFECTLY adequate overall description of the data. But, the fact of the matter is that, it is quite possible that two series are decidedly dissimilar and yet have exactly the same arithmetic mean AND standard deviation: Let us understand this point with the help of an example:

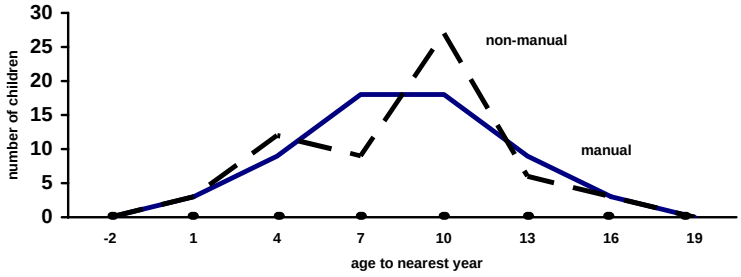
EXAMPLE:

Age of Onset of Nervous Asthma in Children (to Nearest Year)	Children of Manual Workers	Children of Non-Manual Workers
0 – 2	3	3
3 – 5	9	12
6 – 8	18	9
9 – 11	18	27
12 – 14	9	6
15 – 17	3	3
	60	60

In order to compute the mean and standard deviation for each distribution, we carry out the following calculations:

Age of Onset of Nervous Asthma in Children (to Nearest Year)		Children of Manual Workers			Children of Non-Manual Workers		
Age Group	X	f_1	f_1X	f_1X^2	f_2	f_2X	f_2X^2
0 – 2	1	3	3	3	3	3	3
3 – 5	4	9	36	144	12	48	192
6 – 8	7	18	126	882	9	63	441
9 – 11	10	18	180	1800	27	270	2700
12 – 14	13	9	117	1521	6	78	1014
15 – 17	16	3	48	768	3	48	768
	51	60	510	5118	60	510	5118

We find that, for each of the two distributions, the mean is 8.5 years and the standard deviation is 3.61 years.
The frequency polygons of the two distributions are as follows:



By inspecting these, it can be seen that one distribution is symmetrical while the other is quite different. The distinguishing feature here is the degree of asymmetry or *SKEWNESS* in the two polygons. In order to *measure* the skewness in our distribution, we compute the **PEARSON's COEFFICIENT OF SKEWNESS** which is defined as:
Pearson's Coefficient of Skewness:

$$\frac{mean - \mathrm{mod}\, e}{s\,\mathrm{tan}\, dard\ \ deviation}$$

Applying the *empirical relation* between the mean, median and the mode, the Pearson's Coefficient of Skewness is given by:

Pearson's Coefficient of Skewness

$$= \frac{3\,(\mathrm{mean} - \mathrm{median})}{\mathrm{s\,tan\,dard\ \ deviation}}$$

For a *symmetrical* distribution the coefficient will always be ZERO, for a distribution skewed to the RIGHT the answer will always be positive, and for one skewed to the LEFT the answer will always be negative.
Let us now calculate this coefficient for the example of the children of the manual and non-manual workers.
Sample statistics pertaining to the ages of these children are as follows:

	Children of Manual Workers	Children of Non-Manual Workers
Mean	8.50 years	8.50 years
Standard deviation	3.61 years	3.61 years
Median	8.50 years	9.16 years
Q_1	6.00 years	5.50 years
Q_3	11.00 years	10.83 years
Quartile deviation	2.50 years	2.66 years

The Pearson's Coefficient of Skewness is calculated for each of the two categories of children, as shown below:
Pearson's Coefficient of Skewness (Modified):

$$\frac{3(\text{Ages of Children of Manual Workers} - 8.50)}{3.61} = 0$$

$$\frac{3(\text{Ages of Children of Non-Manual Workers} - 9.16)}{3.61} = -0.55$$

For the data pertaining to children of manual workers, the coefficient is zero, whereas, for the children of non-manual workers, the coefficient has turned out to be a negative number. This indicates that the distribution of the ages of the children of the manual workers is symmetric whereas the distribution of the ages of the children of the non-manual workers is negatively skewed.
The students are encouraged to draw the frequency polygon and the frequency curve for each of the two distributions, and to compare the results that have just been obtained with the *shapes* of the two distributions.

LECTURE NO. 14

- Bowley's coefficient of skewness
- The Concept of Kurtosis
- Percentile Coefficient of Kurtosis
- Moments & Moment Ratios
- Sheppard's Corrections
- The Role of Moments in Describing Frequency Distributions

You will recall that the Pearson's coefficient of skewness is defined as (mean - mode)/standard deviation, and if we apply the empirical relation between the mean, median and the mode, then the coefficient is given by:

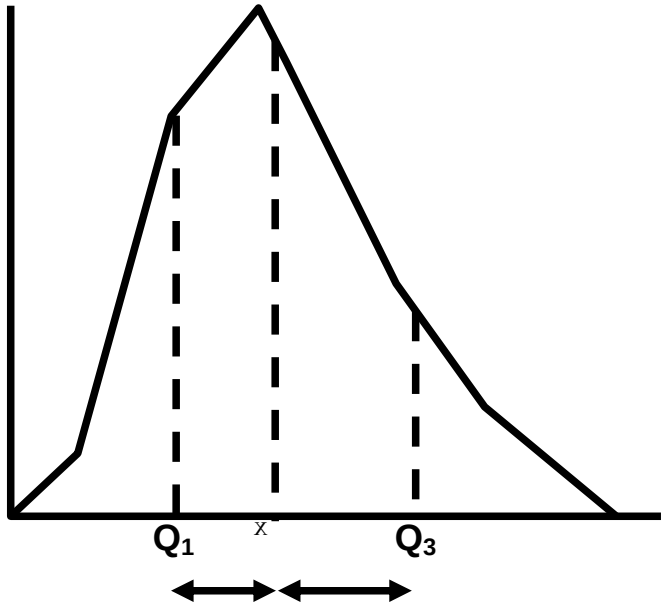
PEARSON'S COEFFICIENT OF SKEWNESS:

$$= \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

As you can see, this coefficient involves the calculation of the mean as well as the standard deviation. Actually, the numerator is divided by the standard deviation in order to obtain a pure number. If the analysis of a data-set is being undertaken using the median and quartiles alone, then we use a measure called Bowley's coefficient of skewness.

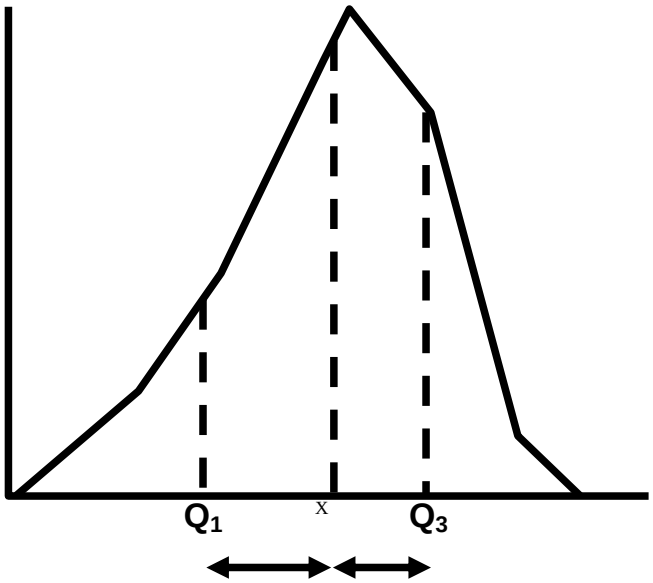
The advantage of this particular formula is that it requires NO KNOWLEDGE of the MEAN or STANDARD DEVIATION. In an asymmetrical distribution, the quartiles will NOT be equidistant from the median, and the AMOUNT by which each one deviates will give an indication of skewness. Where the distribution is positively skewed, Q1 will be closer to the median than Q3.In other words, the distance between Q3 and the median will be greater than the distance between the median and Q1.

POSITIVE SKEWNESS



And hence, if we subtract the distance median - Q1 from the distance Q3 - median, we will obtain a positive answer.
In case of a positively skewed distribution:
(Q3 - median) - (Median - Q1) > 0
i.e. Q1 + Q3 - 2 median > 0
The opposite is true for skewness to the left

NEGATIVE SKEWNESS



In this case:

(Q3 - median) - (Median - Q1) < 0 i.e.

Q1 + Q3 - 2 median < 0

The gist of the above discussion is that in case of a positively skewed distribution, the quantity

$$\mathbf{Q1 + Q3 -}\ 2\tilde{X}$$

will be positive, whereas in case of a negatively distribution, this quantity will be negative.

A RELATIVE measure of skewness is obtained by dividing

$$\mathbf{Q1 + Q3 -}\ 2\tilde{X}$$

by the inter-quartile range i.e. Q3 - Q1, so that Bowley's coefficient of skewness is given by:

Bowley's coefficient of Skewness

$$= \frac{\left(Q_1 + Q_3 - 2\tilde{X}\right)}{Q_3 - Q_1}$$

It is a pure (unit less) number, and its value lies between 0 and ± 1.

For a positively skewed distribution, this coefficient will turn out to be positive, and for a negatively skewed distribution this coefficient will come out to be negative. Let us apply this concept to the example regarding the ages of children of the manual and non-manual workers that we considered in the last lecture.

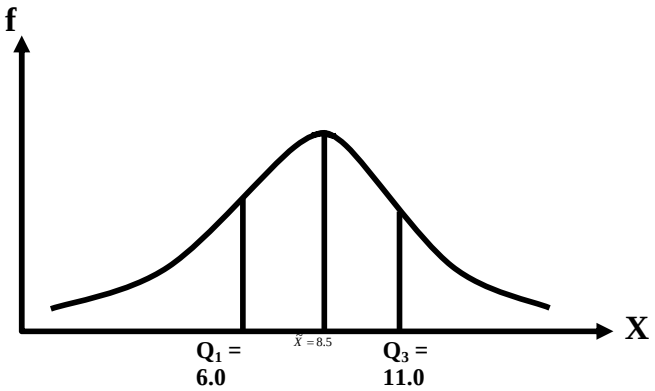
Age of Onset of Nervous Asthma in Children (to Nearest Year)	Children of Manual Workers	Children of Non-Manual Workers
0 – 2	3	3
3 – 5	9	12
6 – 8	18	9
9 – 11	18	27
12 – 14	9	6
15 – 17	3	3
	60	60

EXAMPLE:
Sample statistics pertaining to ages of children of manual and non-manual workers:

	Children of Manual Workers	Children of Non-Manual Workers
Mean	8.50 years	8.50 years
Standard deviation	3.61 years	3.61 years
Median	8.50 years	9.16 years
Q_1	6.00 years	5.50 years
Q_3	11.00 years	10.83 years
Quartile deviation	2.50 years	2.66 years

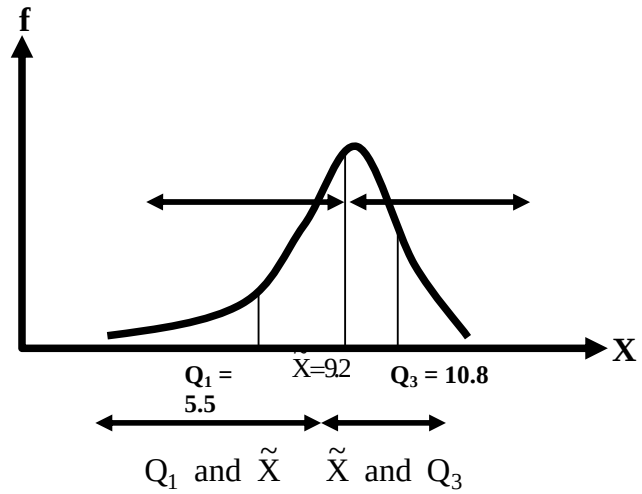
The statistics pertaining to children of manual workers yield the following PICTURE:

Ages of Children of Manual Workers



On the other hand, the statistics pertaining to children of non-manual workers yield the following PICTURE:

Ages of Children of Non-Manual Workers



The diagram pertaining to children of non-manual workers clearly shows that the distance between

is much greater than the distance between
which happens whenever we are dealing with a negatively skewed distribution. If we compute the Bowley's coefficient of skewness for each of these two data-sets, we obtain:

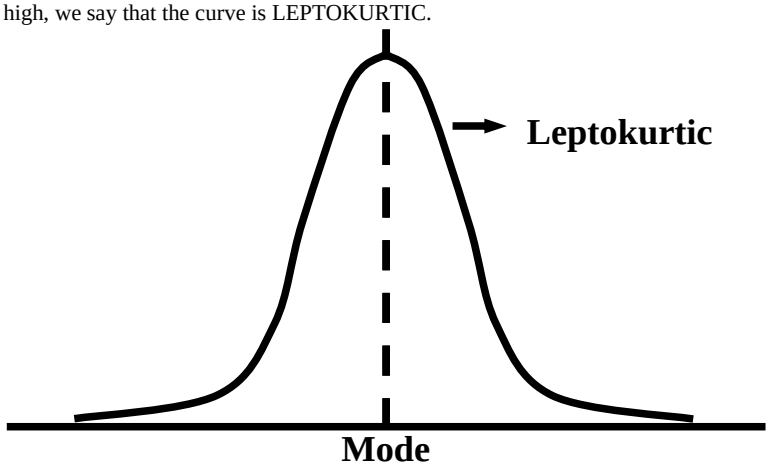
Bowley's Coefficient of Skewness

Ages of Children of Manual Workers	**Ages of Children of Non-Manual Workers**
$= \frac{11.00 + 6.00 - 2 \times 8.50}{2.50}$	$\frac{10.83 + 5.50 - 2 \times 9.16}{10.83 - 5.50}$
$= 0$	$= -0.37$

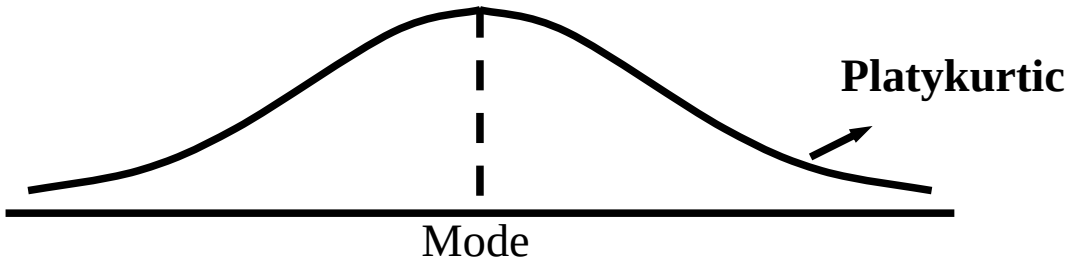
As you have noticed, for the children of the manual workers, the Bowley's coefficient has come out to be zero, whereas for the children of the non-manual workers, the coefficient has come out to be negative. This indicates that the distribution of the ages of the children of manual workers is symmetrical whereas the distribution of the ages of the children of the non-manual workers IS negatively skewed --- EXACTLY the same conclusion that we obtained when we computed the Pearson's coefficient of skewness.

KURTOSIS

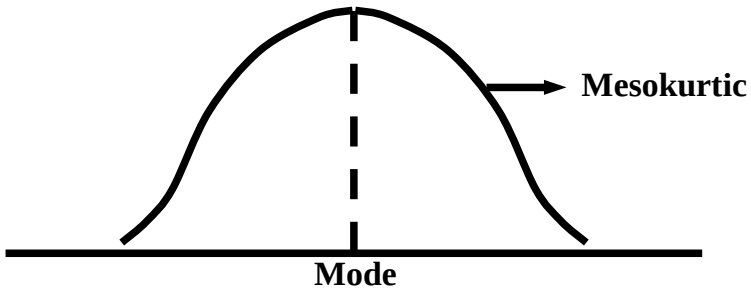
The term kurtosis was introduced by Karl Pearson. This word literally means 'the amount of hump', and is used to represent the degree of PEAKEDNESS or flatness of a unimodal frequency curve.
When the values of a variable are closely BUNCHED round the mode in such a way that the peak of the curve becomes relatively high, we say that the curve is LEPTOKURTIC.



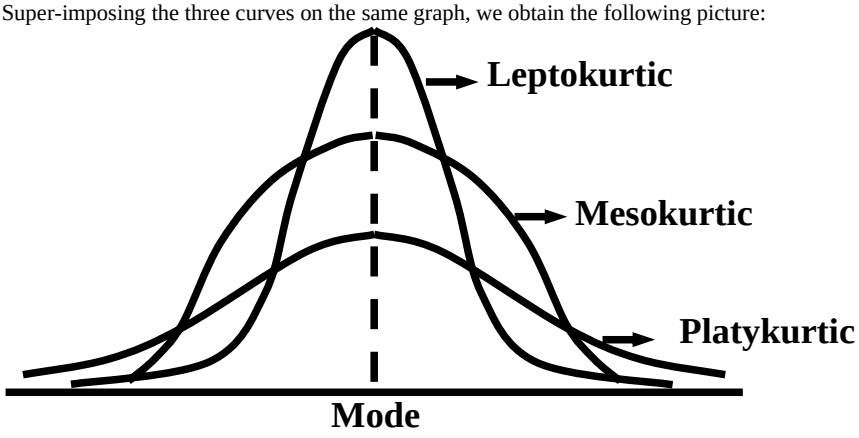
On the other hand, if the curve is flat-topped, we say that the curve is **PLATYKURTIC:**



The NORMAL curve is a curve which is neither very peaked nor very flat, and hence it is taken as A BASIS FOR COMPARISON. The normal curve itself is called **MESOKURTIC.**
I will discuss with you the normal in detail when we discuss continuous probability distributions.
At the moment, just think of the symmetric hump shaped curve shown below:



Super-imposing the three curves on the same graph, we obtain the following picture:



The tallest one is called leptokurtic, the intermediate one is called mesokurtic, and the flat one is called platykurtic.
The question arises, "How will we MEASURE the degree of peakedness or kurtosis of a data-set?"
A MEASURE of kurtosis based on quartiles and percentiles is

$$K = \frac{Q.D.}{P_{90} - P_{10}},$$

This is known as the *PERCENTILE COEFFICIENT OF KURTOSIS.*
It has been shown that K for a normal distribution is 0.263 and that it lies between 0 and 0.50.
In case of a leptokurtic distribution, the percentile coefficient of kurtosis comes out to be LESS THAN 0.263, and in the case of a platykurtic distribution, the percentile coefficient of kurtosis comes out to be GREATER THAN 0.263.The next concept that I am going to discuss with you is the concept of moments --- a MATHEMATICAL concept, and a very important concept in statistics.

.

MOMENTS

A *moment* designates the power to which deviations are raised before averaging them.
For example, the quantity

$$\frac{1}{n}\sum\left(x_i - \bar{x}\right)^1 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)$$

is called the first sample moment about the mean, and is denoted by m1.
Similarly, the quantity

$$\frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$$

is called the second sample moment about the mean, and is denoted by m2.In general, the rth moment about the mean is: the arithmetic mean of the *rth* power of the deviations of the observations from the mean. In symbols, this means that

$$m_r = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^r \qquad \text{for sample data.}$$

Moments about the mean are also called the central moments or the mean moments.
In a similar way, moments about an arbitrary origin, say α, are defined by the relation

$$m'_r = \frac{1}{n}\sum\left(x_i - \alpha\right)^r \quad \text{for sample data}$$

For r = 1, we have

$$m_1 = \frac{1}{n}\sum\left(x_i - \bar{x}\right) = \frac{\sum x_i}{n} - \bar{x} = \bar{x} - \bar{x} = 0,$$

and

$$m_1' = \frac{1}{n}\sum\left(x_i - \alpha\right) = \frac{\sum x_i}{n} - \alpha = \bar{x} - \alpha.$$

Putting r = 2 in the relation for mean moments, we see that

$$m_2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$$

which is exactly the same as the sample variance.
If we take the positive square root of this quantity, we obtain the standard deviation.
In the formula,

$$m'_r = \frac{1}{n}\sum\left(x_i - \alpha\right)^r$$

if we put $\alpha = 0$, we obtain

$$m'_r = \frac{1}{n}\sum x_i^{\,r}$$

and this is called the rth moment about zero, or the rth moment about the origin.
Let us now consolidate the idea of moments by considering an example.

EXAMPLE

Calculate the first four moments about the mean for the following set of examination marks: 45, 32, 37, 46, 39, 36, 41, 48 & 36.
For convenience, the observed values are written in an increasing sequence. The necessary calculations appear in the table below:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
32	– 8	64	– 512	4096
36	– 4	16	– 64	256
36	– 4	16	– 64	256
37	– 3	9	– 27	81
39	– 1	1	– 1	1
41	1	1	1	1
45	5	25	125	625
46	6	36	216	1296
48	8	64	512	4096
360	0	232	186	10708

Now $$\bar{x} = \frac{\sum x_i}{n} = \frac{360}{9} = 40 \qquad \text{marks.}$$

Therefore

$$m_1 = \frac{\sum (x_i - \bar{x})}{n} = 0$$

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \ (\text{marks})^2$$

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n} = \frac{186}{9} = 20.67 \ (\text{marks})^3$$

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n} = \frac{10708}{9} = 1189.78 \ (\text{marks})^4$$

All the formulae that I have discussed until now pertain to the case of raw data.
How will we compute the various moments in the case of grouped data?

MOMENTS IN THE CASE OF GROUPED DATA

When the sample data are grouped into a frequency distribution having k classes with midpoints x1, x2, …, xk and the corresponding frequencies f1, f2, …,fk , ($\sum fi = n$), the rth sample moments are given by

$$m_r = \frac{1}{n} \sum f_i (x_i - \bar{x})^r \text{ , and}$$

$$m'_r = \frac{1}{n} \sum f_i (x_i - \alpha)^r .$$

In the calculation of moments from a grouped frequency distribution, an error is introduced by the assumption that the frequencies associated with a class are located at the MIDPOINT of the class interval. You remember the concept of grouping error that I discussed with you in an earlier lecture? Our moments therefore need corrections.
These corrections were introduced by W.F. Sheppard, and hence they are known as **SHEPPARD'S CORRECTIONS:**
Sheppard's Corrections for Grouping Error:
It has been shown by W.F. Sheppard that, if the frequency distribution (i) is continuous and (ii) tails off to zero at each end, the corrected moments are as given below:

$$m_2 \text{ (corrected)} = m_2 \text{ (uncorrected)} - \frac{h^2}{12};$$

$$m_3 \text{ (corrected)} = m_3 \text{ (uncorrected)};$$

$$m_4 \text{ (corrected)} = m_4 \text{ (uncorrected)} - \frac{h^2}{2} . \ m_2 \text{ (uncorrected)} + \frac{7}{240} . \ h^4;$$

where **h** denotes the uniform class-interval.
The important point to note here is that these corrections are NOT applicable to highly skewed distributions and distributions having unequal class-intervals. I am now going to discuss with you certain mathematical RELATIONSHIPS that exist between the moments about the mean and the moments about an arbitrary origin.
The reason for doing so is that, in many situations, it is easier to calculate the moments in the first instance, about an arbitrary origin. They are then transformed to the mean-moments using the relationships that I am now going to convey to you.
The equations are:

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2;$$

$$m_3 = m'_3 - 3 \ m'_2 \ m'_1 + 2(m'_1)^3, \text{ and}$$

$$m_4 = m'_4 - 4 \ m'_3 \ m'_1 + 6 m'_2 (m'_1)^2 - 3(m'_1)^4$$

In this course, I will not be discussing the mathematical derivation of these relationships. You are welcome to study the mathematics behind these formulae if you are interested. (The derivation is available in your own text book.)But I would like to give you two tips for remembering these formulae:

- In each of these relations, the sum of the coefficients of various terms on the right hand side equals zero and
- Each term on the right is of the same dimension as the term on the left.

Let us now apply these concepts to an example:

EXAMPLE

Compute the first four moments for the following distribution of marks after applying Sheppard's corrections:

Marks out of 20	5	6	7	8	9	10	11	12	13	14	15
No. of Students	1	2	5	10	20	51	22	11	5	3	1

If we wish to compute the first four moments about the mean by the direct method, first of all, we will have to compute mean itself. The mean of this particular data-set comes out to be 10.06.
But, 10.06 is not a very convenient number to work with!
This is so because when we construct the columns of $X - \overline{X}$, $\left(X - \overline{X}\right)^2$ etc.,

we will have a lot many decimals. An alternative way of computing the moments is to take a convenient number as the arbitrary origin and to compute the moments about this number. Later, we utilize the relationships between the moments about the mean and the moments about the arbitrary origin in order to find the moments about the mean.
In this example, we may select 10 as the arbitrary origin, which is the X-value corresponding to the highest frequency 51, and construct the column of D which is the same as X-10. Next, we compute the columns of fD, fD2, fD3, and so on.

Earnings in Rs.(x_i)	No. of Men f_i	D_i $(x_i - 10)$	f_iD_i	$f_iD_i^2$	$f_iD_i^3$	$f_iD_i^4$
5	1	– 5	– 5	25	– 125	625
6	2	– 4	– 8	32	– 128	512
7	5	– 3	– 15	45	– 135	405
8	10	– 2	– 20	40	– 80	160
9	20	– 1	– 20	20	– 20	20
10	51	0	0	0	0	0
11	22	1	22	22	22	22
12	11	2	22	44	88	176
13	5	3	15	45	135	405
14	3	4	12	48	192	768
15	1	5	5	25	125	625
Sum	131	..	8	346	74	3718
Sum ÷ n	1	..	0.06 $=m'_1$	2.64 $=m'_2$	0.56 $=m'_3$	28.38 $=m'_4$

Moments about the mean are:

m1 = 0
m2 = m´2 – (m ´1)2 = 2.64 – (0.06)2 = 2.64

m3 = m ´3 – 3m ´2m´1 + 2 (m ´1)3
= 0.56 – 3(2.64) (0.06) + 2(0.06)3
= 0.08

m4 = m ´4 – 4m ´3m ´1 + 6m ´2 (m ´1)2 – 3(m ´1)4
= 28.38 – 4.(0.56) (0.06) + 6(2.64) (0.06)2 – 3(0.06)4

= 28.30

Applying Sheppard's corrections, we have

m2 (corrected) = m2 (uncorrected) – = 2.64 – 0.08 = 2.56,

m3 (corrected) = m3 (uncorrected) = 0.08,

m4 (corrected) = m4 (uncorrected)
– . m2 (uncorrected) +

= 28.30 – 1.32 + 0.03 = 27.01

I have discussed with you in quite a lot of detail the concept of moments.
The question arises, "Why is it that we are going through all these lengthy calculations? What is the significance of computing moments? "You will obtain the answer to this question when I discuss with you the concept of moment ratios. There are certain ratios in which both the numerators and the denominators are moments. The most common of these moment-ratios are denoted by b1 and b2, and defined by the relations:

MOMENT RATIOS:

$$b_1 = \frac{(m_3)^2}{(m_2)^3} \text{ and } b_2 = \frac{m_4}{(m_2)^2}$$

(in the case of sample data)
They are independent of origin and units of measurement, i.e. they are pure numbers.
b1 is used to measure the skewness of our distribution, and b2 is used to measure the kurtosis of the distribution.

INTERPRETATION OF b1

For symmetrical distributions, b1 is equal to zero. Hence, for any data-set, b1 comes out to be zero, we can conclude that our distribution is symmetric. It should be noted that the measure which will indicate the direction of skewness is the third moment round the mean.

If our distribution is positively skewed, m3 will be positive, and if our distribution is negatively skewed, m3 will be negative.b1 will turn out to be positive in both situations because it is given by

$$b_1 = \frac{(m_3)^2}{(m_2)^3}$$

(Since m3 is being squared, b1 will be positive regardless of the sign of m3.)

INTERPRETATION OF b2

For the normal distribution, b2 = 3.
For a leptokurtic distribution, b2 > 3, and for a platykurtic distribution, b2 < 3
You have noted that the third and fourth moments about the mean provide information about the skewness and the kurtosis of our data-set. This is so because m3 occurs in the numerator of b1 and m4 occurs in the numerator of b2.
What about the dispersion and the centre of our data-set? Do you not remember that the second moment about the mean is exactly the same thing as the variance, the positive square root of which is the standard deviation --- the most important measure of dispersion? What about the centre of the distribution? You will be interested to note that the first moment about zero is NONE OTHER than the arithmetic mean!
This is so because $\frac{1}{n}\sum(x_i - 0)^1$ is equal to $\frac{1}{n}\sum x_i$

--- none other than the arithmetic mean! In this way, the first four moments play a KEY role in describing frequency distributions.

LECTURE NO. 15

On numerous occasions, our interest lies not in just one single variable but in two, three, four or more variables. For example, if we talk about the yield of a crop, we realize that the yield of any crop depends on a variety of factors --- the fertility of the soil, the type of fertilizer used, the amount of rainfall, and so on.

- Simple Linear Regression
- Standard Error of Estimate
- Correlation

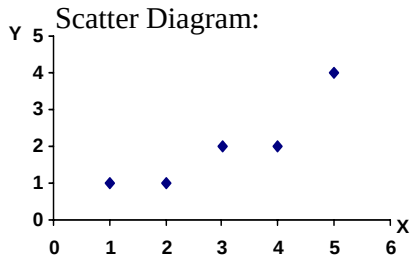
Let me begin the discussion of the bivariate situation by picking up an example.

EXAMPLE:

An important concern for any pharmaceutical company producing drugs is to determine how a particular drug will affect one's perception or general awareness. Suppose one such company wants to establish a relationship between the PERCENTAGE of a drug in the blood-stream and the LENGTH OF TIME it takes to respond to a stimulus.
Suppose the company administers this drug on 5 subjects and obtains the following information:

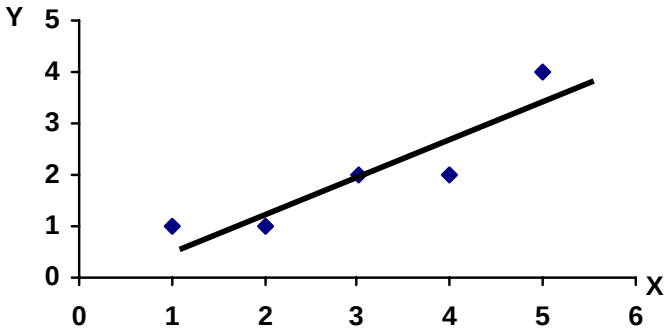
Subject	Percentage of drug	Reaction Time (milli-seconds)
	X	Y
A	1	1
B	2	1
C	3	2
D	4	2
E	5	4

In this example, the reaction time to the stimulus will DEPEND on the amount of drug in the blood-stream. As you must know, the dependent variable is denoted by Y, and the independent variable is denoted by X.In this example, the reaction time will be denoted by Y, and the percentage of drug in the blood stream by X. Going back to the example that we were just considering, it is obvious that we are interested in determining the nature of the relationship between the amount of drug in the blood stream and the time it takes to react to a stimulus.
In order to ascertain the nature of the relationship between these two variables, the first step is to draw a **SCATTER DIAGRAM** --- which is a simple graph of the X-values against the Y-values depicted on the graph paper in the form of points.
In this example, the scatter diagram is as follows:



As you can see, there is an upward trend in the scatter diagram i.e. it is clear that as X increases, Y also increases. Of course, the points are not all falling on a straight line, but if we look carefully, we find an overall linear pattern as shown below:

Scatter Diagram:



It will be very RARE in the field of behavioral or social sciences to find two sets of data which are related perfectly by a straight line: it is more likely that only a general linear pattern or tendency will be apparent.
WHY is it that we will not get an exact linear relationship?
Let me explain this to you with the help of an example:
Suppose one is studying the relationship between the research and development expenditure and the profit margin on products of a number of firms. While it may be generally true to state that the two will increase together, it is INEVITABLE that some firms' profit margin will be higher than others with the SAME *R and D* expenditure, and vice versa. The reasons for this may be that the conditions under which the various firms are operating may be very different. The goods being produced, the firm's share of the market, the efficiency of the firm etc. will ALL play a part in determining the individual results.

A linear relationship between two variables is a SURPRISINGLY common occurrence, and even where a refined non-linear curve might prove slightly superior, the SIMPLER form will often be quite adequate in the context of the problem under consideration. Having plotted the n pairs of values in the form of a scatter diagram, IF an overall linear pattern emerges, then the object of regression is to superimpose on this pattern the general relationship between y and x in the linear form which will REMOVE the effect of outside factors. I am sure that you are aware of the equation of a straight line.
Do you not remember the equation **Y = mX + c,** where Y represents the slope of the line, and c represent the Y-intercept?
This equation can also be stated as Y = c + mX, and if we rename c and m as 'a' and 'b', the equation becomes **Y = a + bX.**

EQUATION OF A STRAIGHT LINE

$Y = a + bX$

Where

- Y represents the dependent variable
- X represents the independent variable
- a represents the Y-intercept

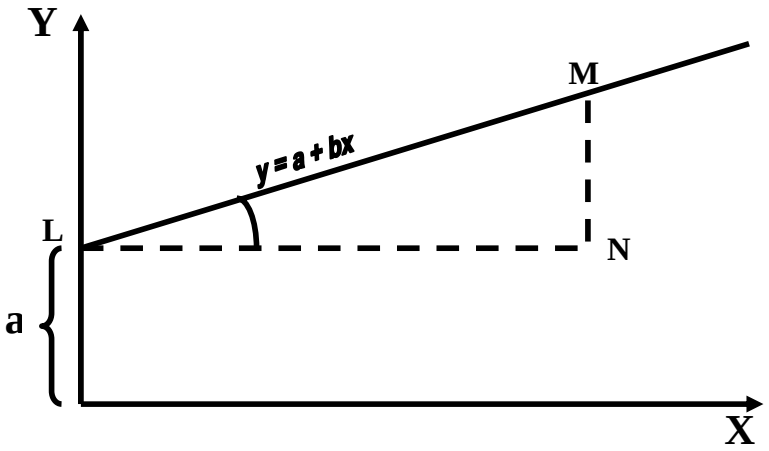
(i.e. the value of Y when X is equal to zero)

- b represents the slope of the line

(i.e. the value of the tan θ, where θ represents the angle between the line and the horizontal axis)

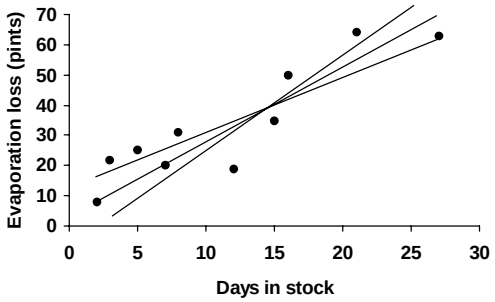
Interpretation of 'a' and 'b':

$$b = \tan\theta = \frac{MN}{NL}$$



A very important point to note is that MANY lines can be drawn through the same scatter diagram.

THE LINEAR PATTERN:



Even with the greatest care and skill, a line drawn between the points with a ruler will be highly SUBJECTIVE, and different individuals will arrive at different lines.

The *real* objective is to find the line of BEST fit. For this, we use a method known as THE METHOD OF LEAST SQUARES. The line of best fit obtained by the method of least squares is called the REGRESSION LINE of Y on X.

And, this whole process is known as simple linear regression. A very important point to note here is that, from the MATHEMATICAL standpoint, simple linear regression requires that X is a NON-RANDOM variable, whereas Y is a RANDOM variable. For example, consider the case of agricultural experiments. If we conduct an experiment to determine the optimal amount of a particular fertilizer to obtain the maximum yield of a certain crop, then the amount of fertilizer is a non-random variable whereas the yield is a random variable. This is so because the amount of fertilizer is in our OWN control. But, the yield is a random variable because it is NOT in our control. In connection with determining the line of BEST fit, the first point is that.

If we use the 'FREE-hand' method of curve-fitting in order to represent the relationship between X and Y as portrayed by the scatter diagram, one tends, consciously or subconsciously, to draw the straight line such that there is EQUAL numbers of points located on either side of the line.

What is more, the eye will automatically try to judge and EQUATE the total distances between the points above and below the line.

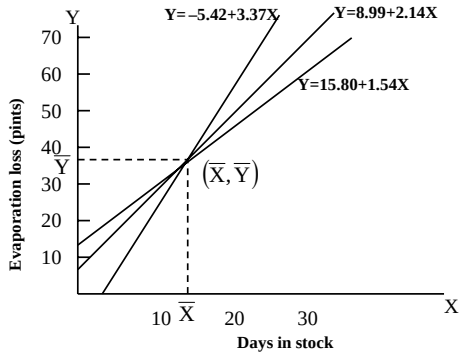
This is a recognition of the fact that the line of best fit must be an 'AVERAGE' line in the true sense.

You will recall that the sum of the deviations round the ARITHMETIC MEAN of a data-set is always equal to zero i.e. the positive and negative deviations CANCEL each other.

Similarly, POSITIVE and NEGATIVE deviations round a line of BEST fit must CANCEL out.

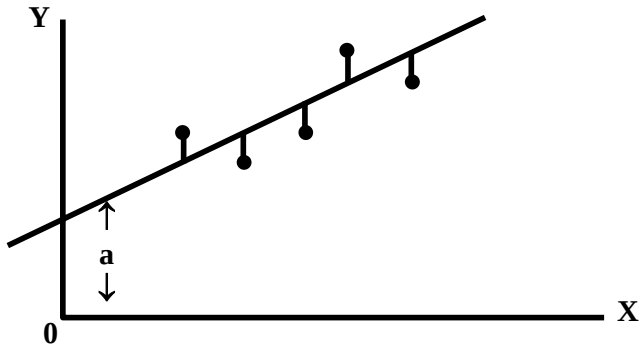
This is the first of the conditions or requirements for an optimal line.
But, the point to understand is that there are an INFINITELY large number of straight lines which will satisfy this condition.
Any line that passes through the point $(\overline{X}, \overline{Y})$ will satisfy this condition, and as shown in the following figure, numerous lines can pass through the point $(\overline{X}, \overline{Y})$

Three equations where $\sum(Y - \hat{Y}) = 0$:



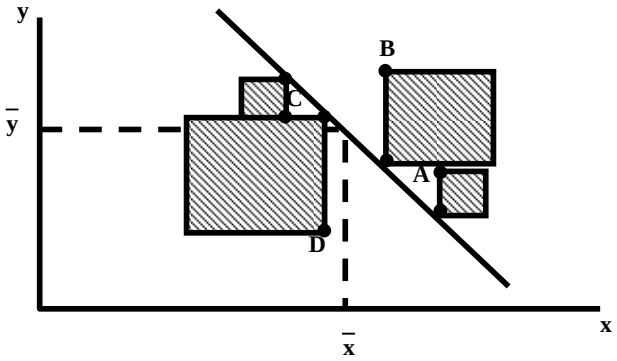
.

For each of the three lines that you see, the SUM of the VERTICAL deviations between the data-points and the line is ZERO. These deviations are depicted by the following diagram:

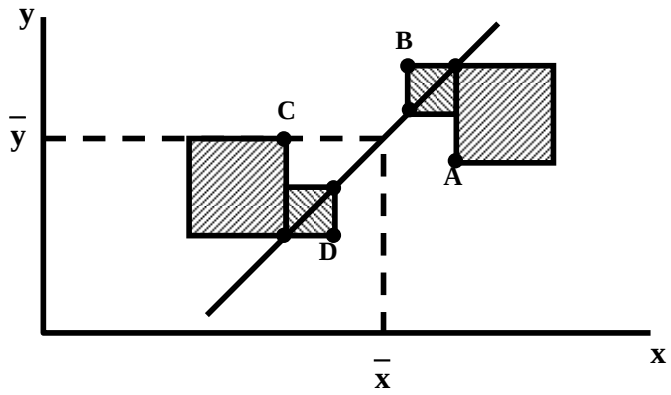


How will we calculate these vertical deviations?
The values of Y obtained from the line are denoted by .
And the deviations of the actual Y-values from the corresponding Y-values obtained from the line are obtained by subtracting from Y.
In all the cases --- as long as our line passes through the point $(\overline{X}, \overline{Y})$ --- , we find that the sum of the deviations of the *actual* Y-values from the corresponding Y-values obtained from the *line* is zero.
Hence, it appears that we need some SECOND criterion for establishing a *unique* position for the BEST-fitting line.
Our interest is NOT simply in achieving a non-zero sum: it is the MAGNITUDE of the sum which is our main concern.
In the two figures that follow, the DIFFERENCES in the sums of squared deviations for two different lines passing through the SAME scatter diagram are clearly portrayed. In position 1, the shaded areas are relatively large, but AS the line is rotated around the point $(\overline{X}, \overline{Y})$ in a clockwise direction to position 2, the areas become smaller.

Position 1



Position 2



There would seem to be some UNIQUE position at which the sum of the square deviations is at a MINIMUM. *This* is the position of *least* squares.
If we can ascertain the location of THIS particular straight line in terms of the constants a and b of the linear equation Y = a + bX, then we have found the line of BEST fit.
The rationale is that the SMALLER the sum of the squared deviations round the mean, the LESS dispersed are the data points around the fitted line.

THE PRINCIPAL OF LEAST SQUARES

According to the principal of least squares, the best-fitting line to a set of points is the one for which the sum of the squares of the vertical distances between the points and the line is minimum.
The line Y = a + bX is the one that best fits the given set of points according to the principal of least squares.
And, this best fitting line is obtained by solving simultaneously two equations which are known as the normal equations.

NORMAL EQUATIONS

$$\left.\begin{aligned} \sum Y &= na + b\sum X \\ \sum XY &= a\sum X + b\sum X^2 \end{aligned}\right\}$$

In connection with these two equations, two points should be noted:
1) I will not be discussing the mathematical derivation of these equations.
2) The word "normal" here has nothing to do with the well-known normal distribution.
For any bivariate data-set, obviously we will have available to us two columns, a column of X and a column of Y.
Hence, obviously, we will be in a position to compute sums like ∑X, ∑Y, ∑XY, and so on.

Hence, the only unknown quantities in the two normal equations are a and b, as shown below:

NORMAL EQUATIONS

$$\left.\begin{aligned} \sum Y &= na + b\sum X \\ \sum XY &= a\sum X + b\sum X^2 \end{aligned}\right\}$$

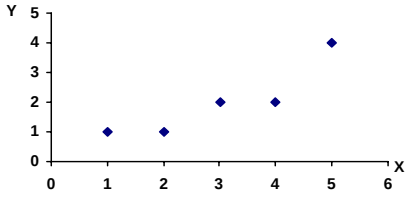
Hence, when we solve the two normal equations simultaneously, we will obtain the values of a and b, and these are EXACTLY the two quantities that we need in order to obtain the BEST-fitting line.
Let me explain this whole concept to you with the help of the same example that I picked up in the beginning of today's lecture:

EXAMPLE

An important concern for any pharmaceutical company producing drugs is to determine how a particular drug will affect one's perception or general awareness. Suppose one such company wants to establish a relationship between the PERCENTAGE of a drug in the blood-stream and the LENGTH OF TIME it takes to respond to a stimulus. Suppose the company administers this drug on 5 subjects and obtains the following information:

Subject	Percentage of drug	Reaction Time (milli-seconds)
	X	Y
A	1	1
B	2	1
C	3	2
D	4	2
E	5	4

Scatter Diagram:



In order to find a and b, we need to solve the two normal equations, and for this purpose, we will carry out computations as shown below:

X	Y	X^2	XY
1	1	1	1
2	1	4	2
3	2	9	6
4	2	16	8
5	4	25	20
15	10	55	37

$$10 = 5a + 15\,b \qquad \text{— } 1$$
$$37 = 15a + 55b \qquad \text{— } 2$$
$$-7 = -10b$$
$$b = \frac{-7}{-10} = 0.7$$
$$10 = 5a + 15b$$
$$\therefore \quad 10 - 15b = 5a$$
$$\therefore \quad 10 - 15(0.7) = 5a$$
$$\therefore \quad 10 - 10.5 = 5a$$
$$\therefore \quad -0.5 = 5a \qquad \therefore \quad a = \frac{-0.5}{5} = -0.1$$

Hence our straight line is given by

$$Y = -0.1 + 0.7\,X$$

A hat is placed on top of the Y so as to differentiate the Y values obtained from the line from the ones that pertain to the actual data-points.

The question is, "What is the advantage of fitting this line?"

The answer to this question is that this line can be used to ESTIMATE the value of the dependent variable corresponding to some particular value of the independent variable. In this example, suppose that we are interested in finding out what will be the reaction time of a person who has 4.33% of the drug in his blood stream? The answer will be obtained by putting X=4.33 in the equation that we just obtained.

Our regression line is

$$= -0.1 + 0.7\,X$$

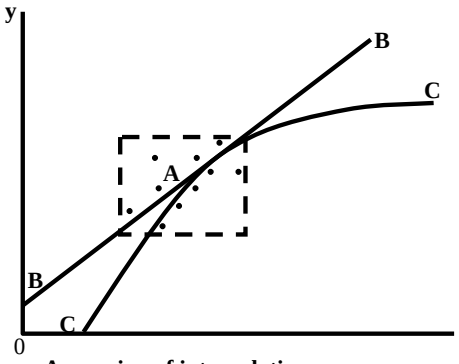
Putting X = 4.33, we obtain

$$= -0.1 + 0.7\,(4.33)$$
$$= -0.1 + 3.031 = 2.931$$

Hence we conclude that it can be expected that a person having 4.33% of the drug in his blood stream will take 2.9 milli-seconds to react to the stimulus. A point to be noted here is that this procedure of estimating the value of the dependent variable should not be used for extrapolation. Extrapolation means the making of estimates or predictions outside the range of the experimental data, and in some situations, this can be very unwise.

Let me explain this point with the help of the following diagram:

The extrapolation trap



A = region of interpolation
B = regions of extrapolation
C = true relationship in regions of extrapolation

While a set of observations may show a good linear relationship between the variables, there is NEVER any *guarantee* that the SAME linear form is present over THOSE ranges of the variable NOT under consideration. I would now like to convey to you another point: All the discussion that I have done until now assumes that Y is the dependent variable and X is the independent variable, and therefore, we are regressing Y on X. But, in some situations, we may be interested in just the OPPOSITE --- i.e. we may wish to regress X on Y. In this situation, all we have to do is to interchange the roles of X and Y. I would like to encourage you to work on this on your own, and to establish the normal equations that will

be required in this situation. You may be thinking, "Why should we go through this hassle? Can't we use the equation that we have just fitted i.e. Y = a + bX to estimate X from Y?"
It is important to note that this is not the case. If we are confronted with a situation where we require to predict Y from X AND X from Y, then *two* DISTINCT equations need to be found. The regressions of Y on X and X on Y for the same bivariate data are NOT identical.
The next concept that I am going to discuss with you is the STANDARD ERROR OF ESTIMATE.

STANDARD ERROR OF ESTIMATE

The observed values of (X, Y) do not all fall on the regression line but they scatter away from it. The degree of scatter of the observed values about the regression line is measured by what is called the standard deviation of regression or the standard error of estimate of Y on X.
For sample data, the standard error of estimate is obtained from the formula

$$s_{y.x} = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n-2}}$$

where Y denotes an observed values, denotes the corresponding values obtained from the least-squares line and n denotes the sample size.
The formula that I just conveyed to you is a bit cumbersome to apply because, in order to apply it, we first need to compute corresponding to all our X-values.
Alternative formula for Sy.x
The standard error of estimate can be more conveniently computed from the alternative formula

$$s_{y.x} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

INTERPRETATION OF Sy.x

The range within which sy.x lies is given by 0 < sy.x < sy (where sy denotes the standard deviation of the y values). sy.x will be zero when all the observed points fall on the regression line (denoting perfect relationship between the two variables). sy.x will be equal to sy when there is no relationship between the two variables. Hence the closer sy.x is to zero (the further away it is from sy), the closer the points are to the line, and the more RELIABLE is our line for purposes of prediction.
Let us apply this concept to the example of the amount of drug in the blood stream and the time taken to react to a stimulus:

X	Y	X^2	Y^2	XY
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

a = -0.1
b = 0.7

$$Sy.x = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

$$= \sqrt{\frac{26 - (-0.1)(10) - (0.7)(37)}{5-2}}$$

$$= \sqrt{\frac{26 + 1 - 25.9}{3}}$$

$$= \sqrt{\frac{1.1}{3}} = \sqrt{0.3667} = 0.61$$

$$\text{Also, } s_y = \sqrt{\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2}$$

$$= \sqrt{\frac{26}{5} - \left(\frac{10}{5}\right)^2} = \sqrt{5.2 - 4}$$

$$= \sqrt{1.2} \qquad = 1.10$$

INTERPRETATION

Sy.x is NOT very small compared with sy, hence our least-squares line.

$$\hat{Y} = -0.1 + 0.7\,X$$

is probably NOT very reliable for purposes of prediction. As it explained a short while ago, the smaller our Standard Error of Estimate, the closer the data-points will be to the line --- i.e. the more REPRESENTATIVE our line will be of the data-points --- and, hence, the more RELIABLE our line will be for estimation purposes.
The next concept that I am going to discuss with you is the concept of CORRELATION.
It is a concept that is very closely linked with the concept of linear regression.

CORRELATION

is a measure of the strength or the degree of relationship between two RANDOM variables.

A numerical measure of the strength of the linear relationship between two random variables X and Y is known as Pearson's Product-Moment Coefficient of Correlation.

PEARSON'S COEFFICIENT OF CORRELATION

where, covariance of X and Y is defined as $r = \dfrac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$

$$Cov(X,Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}$$

This formula is a bit cumbersome to apply. Therefore, we may use the following short cut formula:

SHORT CUT FORMULA FOR THE PEARSON'S COEFFICIENT OF CORRELATION

$$r = \frac{\sum XY - \left(\sum X\right)\left(\sum Y\right)/n}{\sqrt{\left[\sum X^2 - \left(\sum X\right)^2/n\right]\left[\sum Y^2 - \left(\sum Y\right)^2/n\right]}}$$

It should be noted that
r is a pure number that lies between –1 and 1 i.e.

$-1 < r < 1$

Actually, the mathematical expressions that you have just seen is a combination of three different mathematical expressions:

Case 1:

Positive correlation: $0 < r < 1$

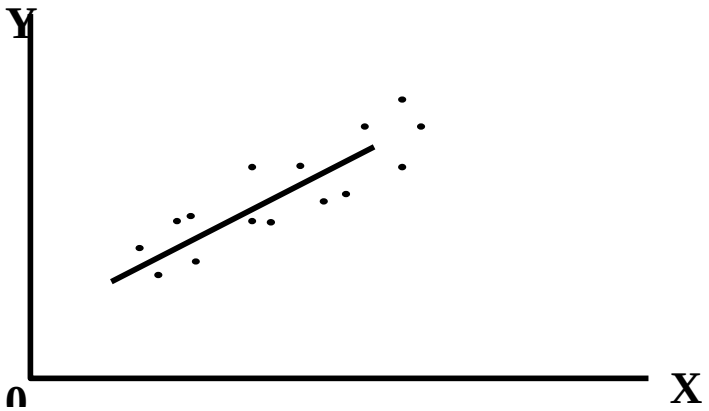
Case 2:

No correlation: $r = 0$

Case 3:

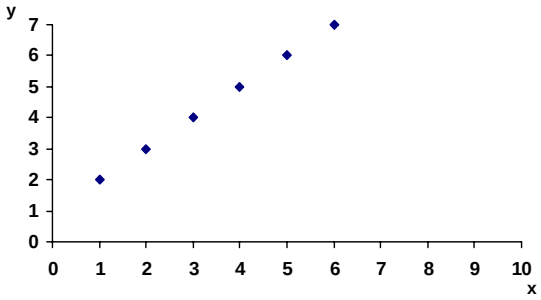
Negative correlation: $-1 < r < 0$

Case 1:
In case of a positive linear relationship, r lies between 0 and 1.



In this case, the closer the points are to the UPWARD-going line, the STRONGER is the positive linear relationship, and the closer r is to 1.

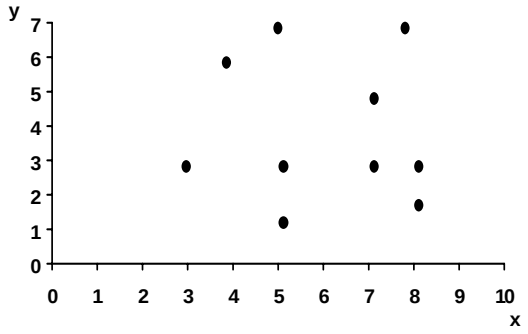
Perfect **Positive Linear Correlation (r = 1):**



In this case, the closer the points are to the DOWNWARD-going line, the stronger is the linear relationship, and the closer r is to –1.

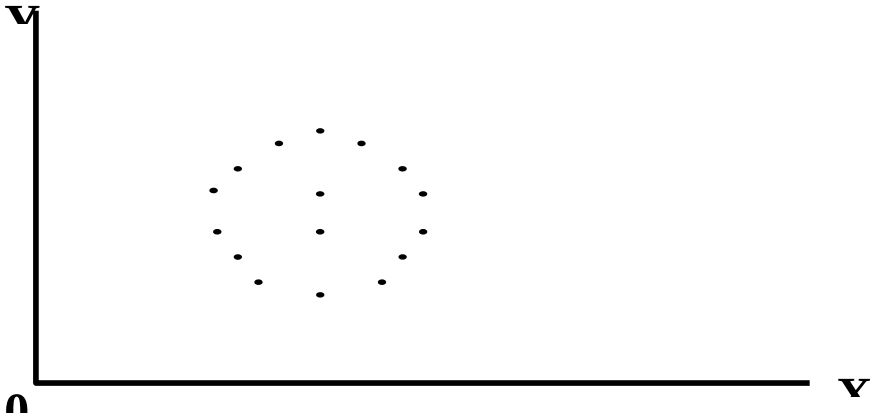
Case 3:

In a situation where neither an upward linear trend nor a downward linear trend can be visualized, r ~ 0



Here, the bivariate data seem to be *completely* random.

The *extreme* of dissociation (*zero* correlation (r = 0)):



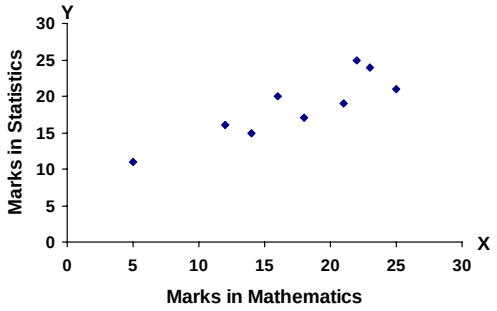
In such a situation, X and Y are said to be *uncorrelated*.

EXAMPLE

Suppose that the principal of a college wants to know if there exists any correlation between grades in Mathematics and grades in Statistics. Suppose that he selects a random sample of 9 students out of all those who take this combination of subjects. The following information is obtained:

Student	Marks in Mathematics (Total Marks: 25)	Marks in Statistics (Total Marks: 25)
	X	Y
A	5	11
B	12	16
C	14	15
D	16	20
E	18	17
F	21	19
G	22	25
H	23	24
I	25	21

SCATTER DIAGRAM



In order to compute the correlation coefficient, we carry out the following computations,

X	Y	X^2	Y^2	XY
5	11	25	121	55
12	16	144	256	192
14	15	196	225	210
16	20	256	400	320
18	17	324	289	306
21	19	441	361	399
22	25	484	625	550
23	24	529	576	552
25	21	625	441	525
156	168	3024	3294	3109

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{\left[\sum X^2 - (\sum X)^2/n\right]\left[\sum Y^2 - (\sum Y)^2/n\right]}}$$

$$= \frac{3109 - (156)(168)/9}{\sqrt{\left[3024 - (156)^2/9\right]\left[3294 - (168)^2/9\right]}}$$

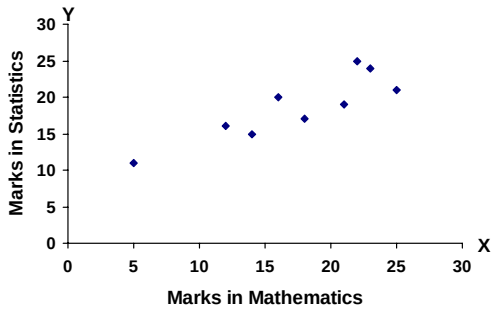
$$= \frac{3109 - 2912}{\sqrt{[3024 - 2704][3294 - 3136]}}$$

$$= \frac{197}{\sqrt{320 \times 158}} = \frac{197}{224.86} = 0.88$$

INTERPRETATION

There exists a strong *positive* linear correlation between marks in Mathematics and marks in Statistics for these 9 students who have been taken into consideration.
The conclusion that we have just drawn i.e. strong positive linear correlation --- this conclusion is supported by the scatter diagram.

SCATTER DIAGRAM



As you can see in the scatter diagram, the data-points appear to follow a linear pattern quite strongly.
In today's lecture, I have discussed with you the concept of regression and correlation. Although I have conveyed to you a number of interesting concepts, believe me, this is only the BEGINNING of a very vast and important area of Statistics. You can study this concept further, and, if possible, to study a little bit about MULTIPLE regression and correlation as well --- the situation when we try to study the relationship between three or more variables.
This brings us to the end of the FIRST part of this course i.e. Descriptive Statistics.
This brings us to the end of the FIRST part of this course i.e. Descriptive Statistics.

LECTURE NO. 16

- Set Theory
- Counting Rules:
- The Rule of Multiplication

"SET"

A set is any well-defined collection or list of distinct objects, e.g. a group of students, the books in a library, the integers between 1 and 100, all human beings on the earth, etc. The term well-defined here means that any object must be classified as either belonging or not belonging to the set under consideration, and the term distinct implies that each object must appear only once. The objects that are in a set, are called members or elements of that set. Sets are usually denoted by capital letters such as A, B, C, Y, etc., while their elements are represented by small letters such as, a, b, c, y, etc.

Elements are enclosed by parentheses to represent a set.

For example:

EXAMPLES OF SETS:

A = {a, b, c, d} or
B = {1, 2, 3, 7}

The Number of a set A, written as n(A), is defined as the number of elements in A.

If x is an element of a set A, we write $x \in A$ which is read as "x belongs to A" or x is in A. If x does not belong to A, i.e. x is not an element of A, we write $x \notin A$.

A set that has no elements is called an empty or a null set and is denoted by the symbol ϕ. (It must be noted that {0} is not an empty set as it contains an element 0.)

If a set contains only one element, it is called a unit set or a singleton set.

It is also important to note the difference between an element "x" and a unit set {x}.

A set may be specified in two ways:

1. We may give a list of all the elements of a set (the "Roster" method),

e.g.

A = {1, 3, 5, 7, 9, 11} ;
B = {a book, a city, a clock, a teacher};

2. We may state a rule that enables us to determine whether or not a given object is a member of the set(the "Rule" method or the "Set Builder" method),

e.g.

A = {x | x is an odd number and x < 12} meaning that A is a set of all elements x such that x is an odd number and x is less than 12. (The vertical line is read as "such that".). *An important point to note is that:* The repetition or the order in which the elements of a set occur, does not change the nature of the set. The size of a set is given by the number of elements present in it. This number may be finite or infinite. Thus a set is finite when it contains a finite number of elements; otherwise it is an infinite set.

The Empty set is regarded as a Finite set.

EXAMPLES OF FINITE SETS

i) A = {1, 2, 3……., 99, 100};

ii) B = {x | x is a month of the year};

iii) C = {x | x is a printing mistake in a book};

iv) D = {x | x is a living citizen of Pakistan};

Examples of infinite sets:

i) A = {x | x is an even integer};

ii) B = {x | x is a real number between 0 and 1 inclusive},
i.e. B = (x | x 0 < x < 1}

iii) C = {x | x is a point on a line};

iv) D = {x | x is a sentence in a English language}; etc

SUBSETS

A set that consists of some elements of another set, is called a subset of that set.
For example, if B is a subset of A, then every member of set B is also a member of set A.
If B is a subset of A, we write:
$B \subset A$ or equivalently:
$A \supset B$
'B is a subset of A' is also read as 'B is contained in A',
or 'A contains B'.

EXAMPLE

If A = {1, 2, 3, 4, 5, 10}
and B {1, 3, 5}
then $B \subset A$,
i.e. B is contained in A.
It should be noted that any set is always regarded a subset of itself.
and an empty set ϕ is considered to be a subset of every set.
Two sets A and B are Equal or Identical, if and only if they contain exactly the same elements.
In other words, A = B if and only if $A \subset B$ and $B \subset A$.

PROPER SUBSET

If a set B contains some but not all of the elements of another set A, while A contains each element of B, i.e. if
$B \subset A$ and $B \neq A$
then the set B is defined to be a proper subset of A.

Universal Set:

The original set of which all the sets we talk about, are subsets, is called the universal set (or the space) and is generally denoted by S or Ω.
The universal set thus contains all possible elements under consideration.
A set S with n elements will produce 2n subsets, including S and ϕ.

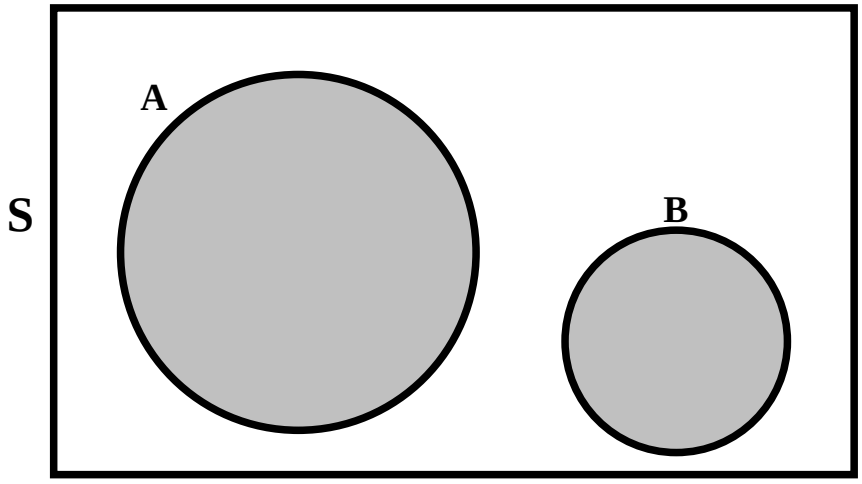
EXAMPLE;

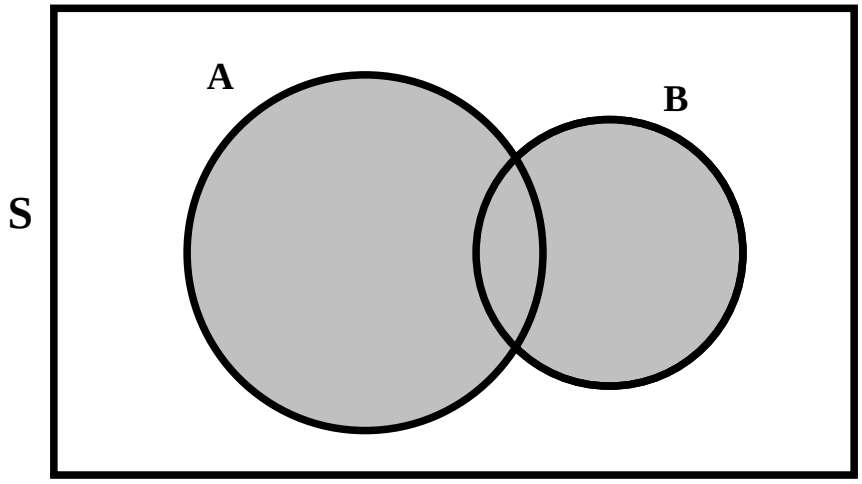
Consider the set A = {1, 2, 3}.
All possible subsets of this set are:
ϕ, {1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3} and {1, 2, 3}
Hence, there are 23 = 8 subsets of the set A.

VENN DIAGRAM

A diagram that is understood to represent sets by circular regions, parts of circular regions or their complements with respect to a rectangle representing the space S is called a Venn diagram, named after the English logician John Venn (1834-1923).
The Venn diagrams are used to represent sets and subsets in a pictorial way and to verify the relationship among sets and subsets.
A Simple Venn diagram:

Disjoint Sets



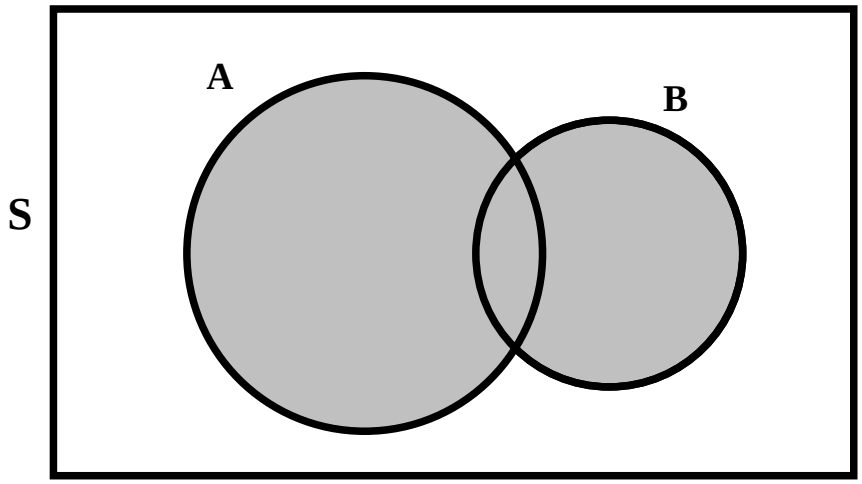


OPERATIONS ON SETS

Let the sets A and B be the subsets of some universal set S. Then these sets may be combined and operated on in various ways to form new sets which are also subsets of S.
The basic operations are union, intersection, difference and complementation.

UNION OF SETS

The union or sum of two sets A and B, denoted by $A \cup B$, and read as "A union B", means the set of all elements that belong to at least one of the sets A and B, that is
$A \cup B = \{ x \mid x \in A \text{ or } x \in B\}$
By means of a Venn Diagram, $A \cup B$ is shown by the shaded area as below:



$A \cup B$ is shaded

EXAMPLE

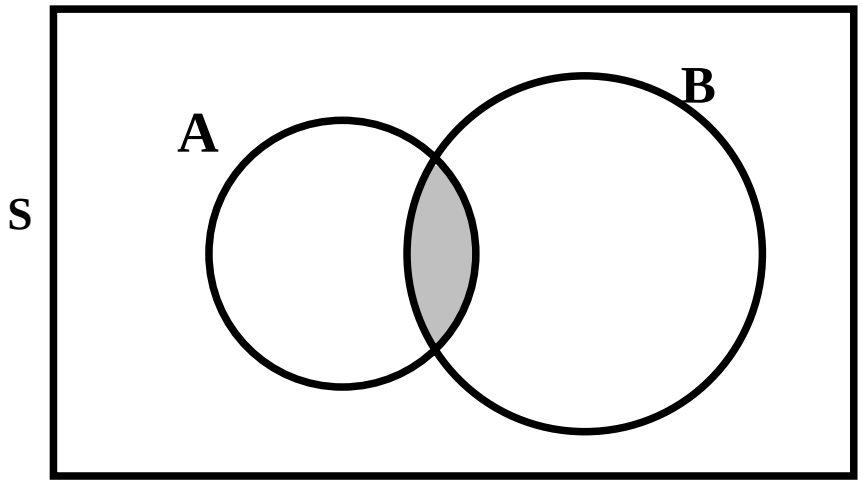
Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$
Then $A \cup B = \{1, 2, 3, 4, 5, 6\}$

INTERSECTION OF SETS

The intersection of two sets A and B, denoted by $A \cap B$, and read as "A intersection B", means that the set of all elements that belong to both A and B; that is

$A \cap B = \{x \mid x \in \text{and } x \in B\}$.

Diagrammatically, $A \cap B$ is shown by the shaded area as below:



$A \cap B$ is shaded

EXAMPLE

Let A = {1, 2, 3, 4} and B = {3, 4, 5, 6}

Then $A \cap B = \{3, 4\}$

The operations of union and intersection that have been defined for two sets may conveniently be extended to any finite number of sets.

DISJOINT SETS

Two sets A and B are defined to be disjoint or mutually exclusive or non-overlapping when they have no elements in common, i.e. when their intersection is an empty set

i.e. $A \cap B = \phi$.

On the other hand, two sets A and B are said to be conjoint when the have at least one element in common.

SET DIFFERENCE

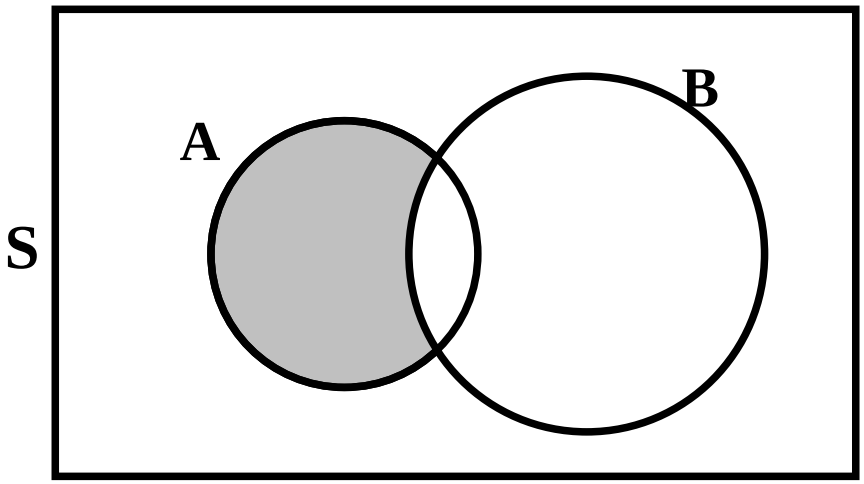
The difference of two sets A and B, denoted by A – B or by $A - (A \cap B)$, is the set of all elements of A which do not belong to B.

Symbolically,

$A - B = \{x \mid x \in A \text{ and } x \notin B\}$

It is to be pointed out that in general $A - B \neq B - A$.

The shaded area of the following Venn diagram shows the difference A – B:



Difference A – B is shaded

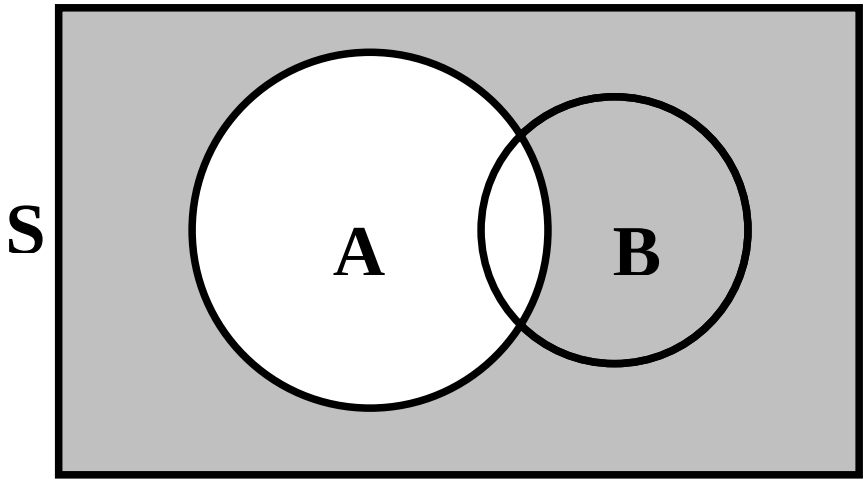
It is to be noted that A – B and B are disjoint sets. If A and B are disjoint, then the difference A – B coincides with the set A.

COMPLEMENTATION

The particular difference S – A, that is, the set of all those elements of S which do not belong to A, is called the complement of A and is denoted by $\bar{A}$ or by Ac.
In symbols:

$$\bar{A} = \{x \mid x \in S \text{ and } s \notin A\}$$

The complement of S is the empty set ϕ.
The complement of A is shown by the shaded portion in the following Venn diagram.



$\bar{A}$ is shaded

It should be noted that A – B and $A \cap \bar{B}$, where $\bar{B}$ is the complement of set B, are the same set. Next, we consider the Algebra of Sets. The algebra of sets provides us with laws which can be used to solve many problems in probability calculations.
Let A, B and C be any subsets of the universal set S. Then, we have:

1. Commutative laws:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

2. Associative laws:

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

3. Distributive laws

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

4. Idempotent laws

$A \cup A = A$
$A \cap A = A$

5. Identity laws

$A \cup S = S$,
$A \cap S = A$,
$A \cup \phi = A$, and
$A \cap \phi = \phi$.

6. Complementation laws

$A \cup \bar{A} = S$,
$A \cap \bar{A} = \phi$,

$(\bar{A}) = A$,

$\bar{S} = \phi$, and
$\bar{\phi} = S$

7. De Morgan's laws:

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B},$$

and $$\overline{(A \cap B)} = \bar{A} \cup \bar{B}$$

PARTITION OF SETS

A partition of a set S is a sub-division of the set into non-empty subsets that are disjoint and exhaustive, i.e. their union is the set S itself.
This implies that:

- i)$A_i \cap A_j = \phi$, where $i \neq j$;
- ii)$A_1 \cap A_2 \cup \ldots \cup A_n = S$.

The subsets in a partition are called cells.

EXAMPLE

Let us consider a set S = {a, b, c, d, e}.
Then {a, b}, and {c, d, e} is a partition of S as each element of S belongs to exactly one cell.

CLASS OF SETS

A set of sets is called a class. For example, in a set of lines, each line is a set of points.

POWER SET

The class of ALL subsets of a set A is called the Power Set of A and is denoted by P(A).
For example, if A = {H, T}, then P(A) = {ϕ, {H}, {T}, {H, T}}.

CARTESIAN PRODUCT OF SETS

The Cartesian product of sets A and B, denoted by $A \times B$, (read as "A cross B"), is a set that contains all ordered pairs (x, y), where x belongs to A and y belongs to B.
Symbolically, we write
$A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\}$
This set is also called the Cartesian set of A and B set of A and B, named after the French mathematician Rene' Descartes (1596-1605).
The product of a set A by itself is denoted by A^2.
This concept of product may be extended to any finite number of sets.

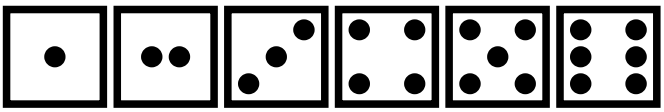
EXAMPLE

Let A = {H, T} and B = {1, 2, 3, 4, 5, 6}. Then the Cartesian product set is the collection of the following twelve (2 × 6) ordered pairs:

A×B = {(H, 1); (H, 2);(H, 3); (H, 4);(H, 6); (H, 6);(T, 1); (T, 2); (T, 3); (T, 4); (T, 5); (T, 6) }

Clearly, these twelve elements together make up the universal set S when a COIN and a DIE are tossed together.

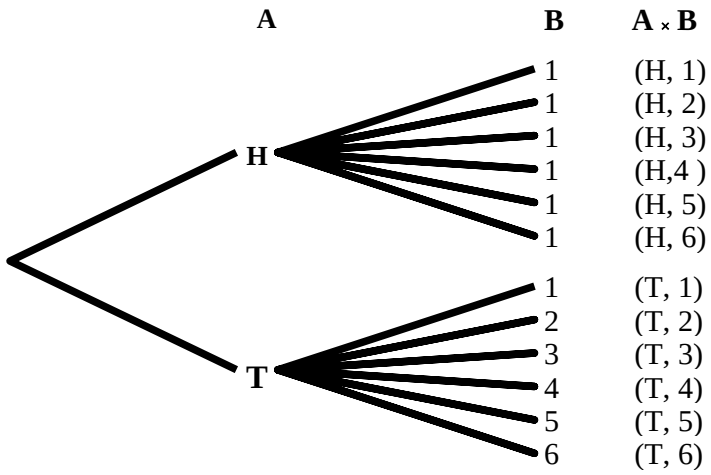
A die is a cube of wood or ivory whose six faces are marked with dots are shown below:



The plural of the word 'die' is 'dice'.

The product A × B may conveniently be found by means of the so-called tree diagram shown below:

Tree Diagram



TREE DIAGRAM

The "tree" is constructed from the left to the right. A "tree diagram" is a useful device for enumerating all the possible outcomes of two or more sequential events.

The possible outcomes are represented by the individual paths or branches of the tree.

It is relevant to note that, in general

$$A \times B \neq B \times A.$$

Having reviewed the basics of set theory, let us now review the COUNTING RULES that facilitate the computation of probabilities in a number of problems.

RULE OF MULTIPLICATION

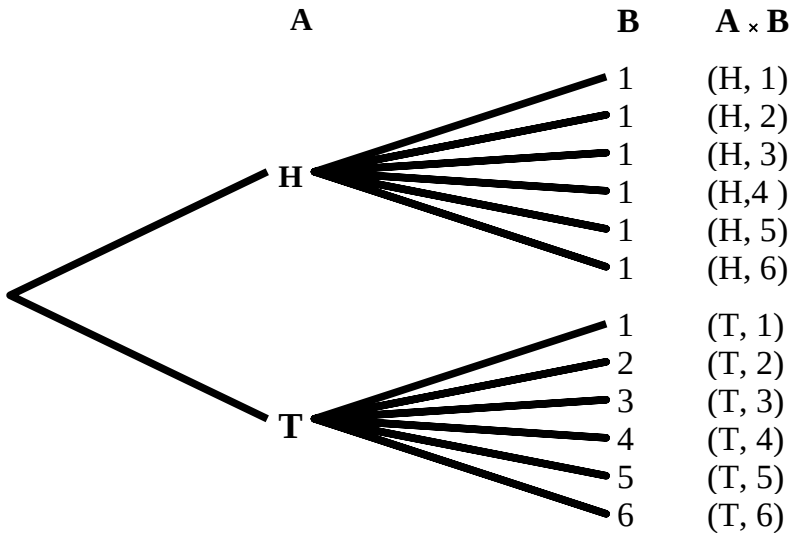
If a compound experiment consists of two experiments which that the first experiment has exactly m distinct outcomes and, if corresponding to each outcome of the first experiment there can be n distinct outcomes of the second experiment, then the compound experiment has exactly mn outcomes.

EXAMPLE

The compound experiment of tossing a coin and throwing a die together consists of two experiments. The coin-tossing experiment consists of two distinct outcomes (H, T), and the die-throwing experiment consists of six distinct outcomes (1, 2, 3, 4, 5, 6).
The total number of possible distinct outcomes of the compound experiment is therefore $2 \times 6 = 12$ as each of the two outcomes of the coin-tossing experiment can occur with each of the six outcomes of die-throwing experiment. As stated earlier, if A = {H, T} and B = {1, 2, 3, 4, 5, 6}, then the Cartesian product set is the collection of the following twelve (2×6) ordered pairs:

A×B = { (H, 1); (H, 2);(H, 3); (H, 4);
(H, 6); (H, 6);(T, 1); (T, 2);
(T, 3); (T, 4); (T, 5); (T, 6) }

Tree Diagram



The rule of multiplication can be readily extended to compound experiments consisting of any number of experiments performed in a given sequence.

This rule can also be called the Multiple Choice Rule, as illustrated by the following example:

EXAMPLE

Suppose that a restaurant offers three types of soups, four types of sandwiches, and two types of desserts. Then, a customer can order any one out of $3 \times 4 \times 2 = 24$ different meals.

EXAMPLE

Suppose that we have a combination lock on which there are eight rings.
In how many ways can the lock be adjusted?
Solution:
The logical way to look at this problem is to see that there are eight rings on the lock, each of which can have any of the 10 figures 0 to 9:

$$\overline{A}\ \overline{B}\ \overline{C}\ \overline{D}\ \overline{E}\ \overline{F}\ \overline{G}\ \overline{H}$$

ring A can have any of the digits 0 to 9 and ring B can have any of the digits 0 to 9 and ring C can have any of the digits 0 to 9 and
.
.
.
ring H can have any of the digits 0 to 9
Hence the total No. of ways in which these 8 rings can be filled is 8
$10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10$
= 10
i.e. 100,000,000 — one hundred million.

LECTURE NO. 17

- Permutations
- Combinations
- Random Experiment
- Sample Space
- Events
  - Mutually Exclusive Events
  - Exhaustive Events
  - Equally Likely Events

COUNTING RULES

As discussed in the last lecture, there are certain rules that facilitate the calculations of probabilities in certain situations. They are known as counting rules and include concepts of;

- Multiple Choice
- Permutations
- Combinations

We have already discussed the rule of multiplication in the last lecture.
Let us now consider the rule of permutations.

RULE OF PERMUTATION

A permutation is any ordered subset from a set of n distinct objects.
For example, if we have the set {a, b}, then one permutation is ab, and the other permutation is ba. The number of permutations of r objects, selected in a definite order from n distinct objects is denoted by the symbol ${}^nP_r$ and is given by

${}^nP_r = n(n-1)(n-2)\ldots(n-r+1)$

$$= \frac{n!}{(n-r)!}.$$

FACTORIALS

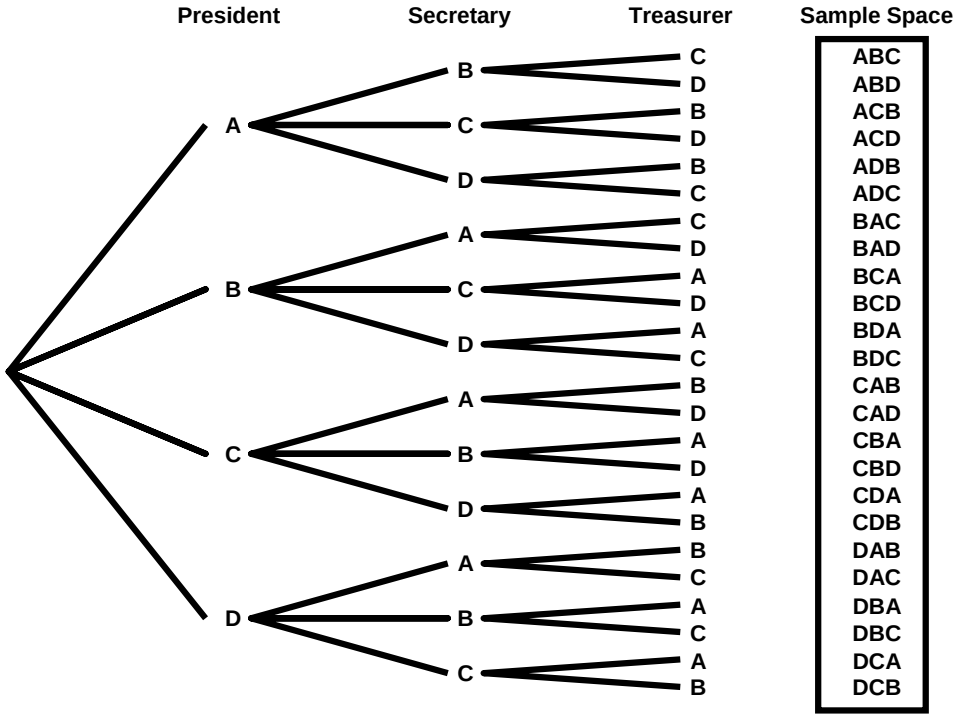
$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$
$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$
.
.
.
$1! = 1$
Also, we define $0! = 1$.

EXAMPLE

A club consists of four members. How many ways are there of selecting three officers: president, secretary and treasurer? It is evident that the order, in which 3 officers are to be chosen, is of significance. Thus there are 4 choices for the first office, 3 choices for the second office, and 2 choices for the third office. Hence the total number of ways in which the three offices can be filled is $4 \times 3 \times 2 = 24$.
The same result is obtained by applying the rule of permutations:

$$
\begin{aligned}
{}^4P_3 &= \frac{4!}{(4-3)!} \\
&= 4 \times 3 \times 2 \\
&= 24
\end{aligned}
$$

Let the four members be, A, B, C and D. Then a tree diagram which provides an organized way of listing the possible arrangements, for this example, is given below:



PERMUTATIONS

In the formula of ${}^{n}P_{r}$, if we put r = n, we obtain:

${}^{n}P_{n}$ = n (n – 1) (n – 2) … 3 × 2 × 1

= n!

I.e. the total number of permutations of n distinct objects, taking all n at a time, is equal to n!

EXAMPLE

Suppose that there are three persons A, B & D, and that they wish to have a photograph taken. The total number of ways in which they can be seated on three chairs (placed side by side) is

3P3 = 3! = 6

These are:
ABD,
ADB,
BAD,
BDA,
DAB,
DBA

The above discussion pertained to the case when all the objects under consideration are distinct objects. If some of the objects are not distinct, the formula of permutations modifies as given below:

The number of permutations of n objects, selected all at a time, when n objects consist of n1 of one kind, n2 of a second kind, …, nk of a kth kind,

$$\text{is } P = \frac{n!}{n_1!\, n_2! \ldots\ldots n_k!}.$$

$$(\text{where } \sum n_i = n)$$

EXAMPLE

How many different (meaningless) words can be formed from the word 'committee'?
In this example:
n = 9 (because the total number of letters in this word is 9)
n1 = 1 (because there is one c)
n2 = 1 (because there is one o)
n3 = 2 (because there are two m's)
n4 = 1 (because there is one i)
n5 = 2 (because there are two t's)
and
n6 = 2 (because there are two e's)
Hence, the total number of (meaningless) words
(permutations) is:

$$P = \frac{n!}{n_1!\, n_2! \ldots\ldots n_k!}.$$

$$= \frac{9!}{1!\ 1!\ 2!\ 1!\ 2!\ 2!}$$

$$= \frac{9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{1\times 1\times 2\times 1\times 1\times 2\times 1\times 2\times 1}$$

$$= 45360$$

Next, let us consider the rule of combinations.

RULE OF COMBINATION

A combination is any subset of r objects, selected without regard to their order, from a set of n distinct objects. The total number of such combinations is denoted by the symbol

and is given by ${}^{n}C_r \text{ or } \binom{n}{r}$,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

where r < n.

It should be noted that

$${}^{n}P_r = r!\binom{n}{r}$$

In other words, every combination of r objects (out of n objects) generates r! Permutations

EXAMPLE

Suppose we have a group of three persons, A, B, & C. If we wish to select a group of two persons out of these three, the three possible groups are {A, B}, {A, C} and {B, C}.In other words, the total number of combinations of size two out of this set of size three is 3.

Now, suppose that our interest lies in forming a committee of two persons, one of whom is to be the president and the other the secretary of a club.
The six possible committees are:

(A, B), (B, A),
(A, C), (C, A),
(B, C) & (C, B)

In other words, the total number of permutations of two persons out of three is 6.
And the point to note is that each of three combinations mentioned earlier generates 2 = 2! Permutations, I.e. the combination {A, B} generates the permutations (A, B) and (B, A) and the combination {A, C} generates the permutations (A, C) and (C, A); and the combination {B, C} generates the permutations (B, C) and (C, B).

The quantity

$$\binom{n}{r}$$

or ${}^{n}C_{r}$ is also called a binomial co-efficient because of its appearance in the binomial expansion of

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r}\ b^r .$$

The binomial co-efficient has two important properties.

i) $\binom{n}{r} = \binom{n}{n-r}$, and

ii) $\binom{n}{n-r} + \binom{n}{r} = \binom{n+1}{r}$

Also, it should be noted that

$$\binom{n}{0} = 1 = \binom{n}{n}$$

and

$$\binom{n}{1} = n = \binom{n}{n-1}$$

EXAMPLE

A three-person committee is to be formed out of a group of ten persons. In how many ways can this be done?
Since the order in which the three persons of the committee are chosen, is unimportant, it is therefore an example of a problem involving combinations. Thus the desired number of combinations is

$$\binom{n}{r} = \binom{10}{3} = \frac{10!}{3!\,(10-3)!} = \frac{10!}{3!\,7!}$$
$$= \frac{10\times 9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{3\times 2\times 1\times 7\times 6\times 5\times 4\times 3\times 2\times 1}$$
$$= 120$$

In other words, there are one hundred and twenty different ways of forming a three-person committee out of a group of only ten persons!

EXAMPLE

In how many ways can a person draw a hand of 5 cards from a well-shuffled ordinary deck of 52 cards?
The total number of ways of doing so is given by

$$\binom{n}{r} = \binom{52}{5} = \frac{52\times 51\times 50\times 49\times 48}{5\times 4\times 3\times 2\times 1} = 2{,}598{,}960$$

Having reviewed the counting rules that facilitate calculations of probabilities in a number of problems, let us now begin the discussion of concepts that lead to the formal definitions of probability. The first concept in this regard is the concept of Random Experiment. The term experiment means a planned activity or process whose results yield a set of data. A single performance of an experiment is called a trial. The result obtained from an experiment or a trial is called an outcome.

RANDOM EXPERIMENT

An experiment which produces different results even though it is repeated a large number of times under essentially similar conditions is called a Random Experiment.

The tossing of a fair coin, the throwing of a balanced die, drawing of a card from a well-shuffled deck of 52 playing cards, selecting a sample, etc. are examples of random experiments.

PROPERTIES OF A RANDOM EXPERIMENT

A random experiment has three properties:

- The experiment can be repeated, practically or theoretically, any number of times.
- The experiment always has two or more possible outcomes. An experiment that has only one possible outcome is not a random experiment.
- The outcome of each repetition is unpredictable, i.e. it has some degree of uncertainty.

Considering a more realistic example, interviewing a person to find out whether or not he or she is a smoker is an example of a random experiment. This is so because this example fulfils all the three properties that have just been discussed:

- This process of interviewing can be repeated a large number of times.
- To each interview, there are at least two possible replies: 'I am a smoker' and 'I am not a smoker'.
- For any interview, the answer is not known in advance i.e. there is an element of uncertainty regarding the person's reply.

A concept that is closely related with the concept of a random experiment is the concept of the Sample Space.

SAMPLE SPACE

A set consisting of all possible outcomes that can result from a random experiment (real or conceptual), can be defined as the sample space for the experiment and is denoted by the letter S. Each possible outcome is a member of the sample space, and is called a sample point in that space. Let us consider a few examples:

EXAMPLE-1

The experiment of tossing a coin results in either of the two possible outcomes: a head (H) or a tail (T). (We assume that it is not possible for the coin to land on its edge or to roll away). The sample space for this experiment may be expressed in set notation as S = {H, T}. 'H' and 'T' are the two sample points.

EXAMPLE-2

The sample space for tossing two coins once (or tossing a coin twice) will contain four possible outcomes denoted by

S = {HH, HT, TH, TT}.

In this example, clearly, S is the Cartesian product $A \times A$, where A = {H, T}.

EXAMPLE-3

The sample space S for the random experiment of throwing two six-sided dice can be described by the Cartesian product $A \times A$, where A = {1, 2, 3, 4, 5,6}. In other words, $S = A \times A = \{(x, y) \mid x \in A \text{ and } y \in A\}$, Where x denotes the number of dots on the upper face of the first die, and y denotes the number of dots on the upper face of the second die. Hence, S contains 36 outcomes or sample points, as shown below:

S = {(1, 1), (1, 2), (1, 3), (1, 5), (1, 6),
(2, 1), (2, 2), (2, 3), (2, 5), (2, 6),
(3, 1), (3, 2), (3, 3), (3, 5), (3, 6),
(4, 1), (4, 2), (4, 3), (4, 5), (4, 6),
(5, 1), (5, 2), (5, 3), (5, 5), (5, 6),
(6, 1), (6, 2), (6, 3), (6, 5), (6, 6)}

The next concept is that of events.

EVENTS

Any subset of a sample space S of a random experiment, is called an event. In other words, an event is an individual outcome or any number of outcomes (sample points) of a random experiment.

SIMPLE & COMPOUND EVENTS

An event that contains exactly one sample point is defined as a simple event. A compound event contains more than one sample point, and is produced by the union of simple events.

EXAMPLE

The occurrence of a 6 when a die is thrown, is a simple event, while the occurrence of a sum of 10 with a pair of dice, is a compound event, as it can be decomposed into three simple events (4, 6), (5, 5) and (6, 4).

OCCURRENCE OF AN EVENT

An event A is said to occur if and only if the outcome of the experiment corresponds to some element of A.

EXAMPLE

Suppose we toss a die, and we are interested in the occurrence of an even number.
If ANY of the three numbers '2', '4' or '6' occurs, we say that the event of our interest has occurred.
In this example, the event A is represented by the set {2, 4, 6}, and if the outcome '2' occurs, then, since this outcome is corresponding to the first element of the set A, therefore, we say that A has occurred.

COMPLEMENTARY EVENT

The event "not-A" is denoted by $\bar{A}$ or Ac and called the negation (or complementary event) of A.

EXAMPLE

If we toss a coin once, then the complement of "heads" is "tails". If we toss a coin four times, then the complement of "at least one head" is "no heads". A sample space consisting of n sample points can produce 2n different subsets (or simple and compound events).

EXAMPLE

Consider a sample space S containing 3 sample points, i.e. S = {a, b, c}.
Then the 23 = 8 possible subsets are

ϕ, {a}, {b}, {c}, {a, b},
{a, c}, {b, c}, {a, b, c}

Each of these subsets is an event. The subset {a, b, c} is the sample space itself and is also an event. It always occurs and is known as the certain or sure event. The empty set ϕ is also an event, sometimes known as impossible event, because it can never occur.

MUTUALLY EXCLUSIVE EVENTS

Two events A and B of a single experiment are said to be mutually exclusive or disjoint if and only if they cannot both occur at the same time i.e. they have no points in common.

EXAMPLE-1

When we toss a coin, we get *either* a head *or* a tail, but *not* both at the same time. The two events head and tail are therefore mutually exclusive.

EXAMPLE-2

When a die is rolled, the events 'even number' and 'odd number' are mutually exclusive as we can get either an even number or an odd number in one throw, not both at the same time. Similarly, a student *either* qualifies or fails, a single birth must be *either* a boy or a girl, it cannot be both, etc., etc. Three or more events originating from the same experiment are mutually exclusive if pair wise they are mutually exclusive. If the two events *can* occur at the same time, they are not mutually exclusive, e.g., if we draw a card from an ordinary deck of 52 playing cars, it *can* be both a king and a diamond.
Therefore, kings and diamonds are not mutually exclusive. Similarly, inflation and recession are not mutually exclusive events. Speaking of playing cards, it is to be remembered that an ordinary deck of playing cards contains 52 cards arranged in 4 suits of 13 each. The four suits are called diamonds, hearts, clubs and spades; the first two are red and the last two are black. The face values called denominations, of the 13 cards in each suit are ace, 2, 3, …, 10, jack, queen and king. The *face cards* are king, queen and jack. These cards are used for various games such as whist, bridge, poker, etc. We have discussed the concepts of mutually exclusive events. Another important concept is that of exhaustive events.

EXHAUSTIVE EVENTS

Events are said to be collectively exhaustive, when the union of mutually exclusive events is equal to the entire sample space S.

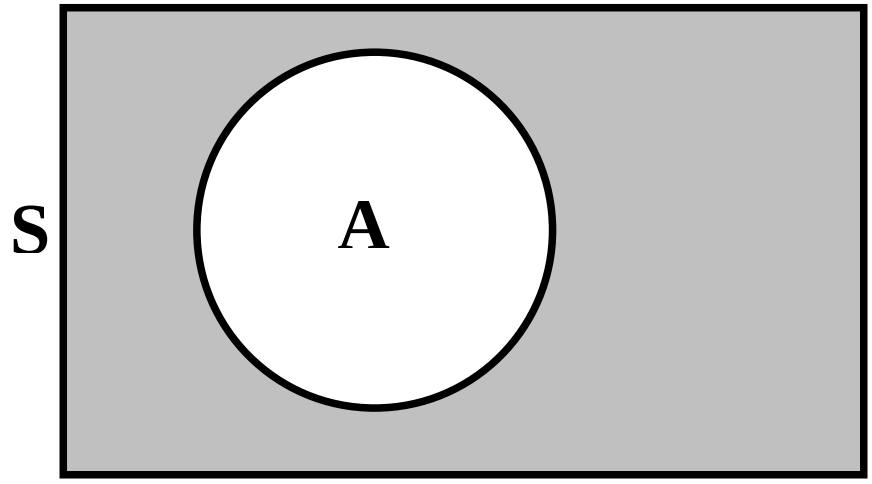
EXAMPLES

- In the coin-tossing experiment, 'head' and 'tail' are collectively exhaustive events.
- In the die-tossing experiment, 'even number' and 'odd number' are collectively exhaustive events.

In conformity with what was discussed in the last lecture:

PARTITION OF THE SAMPLE SPACE

A group of mutually exclusive and exhaustive events belonging to a sample space is called a partition of the sample space. With reference to any sample space S, events A and $\bar{A}$ form a partition as they are mutually exclusive and their union is the entire sample space. The Venn Diagram below clearly indicates this point.



$\bar{A}$ is shaded

Next, we consider the concept of equally likely events:

EQUALLY LIKELY EVENTS

Two events A and B are said to be equally likely, when one event is as likely to occur as the other. In other words, each event should occur in equal number in repeated trials.

EXAMPLE

When a fair coin is tossed, the head is as likely to appear as the tail, and the proportion of times each side is expected to appear is 1/2.

EXAMPLE

If a card is drawn out of a deck of well-shuffled cards, each card is equally likely to be drawn, and the probability that any card will be drawn is 1/52.

LECTURE NO. 18

DEFINITIONS OF PROBABILITY

- Subjective Approach to Probability
- Objective Approach:
- Classical Definition of Probability

RELATIVE FREQUENCY DEFINITION OF PROBABILITY

Before we begin the various definitions of probability, let us revise the concepts of:

- Mutually Exclusive Events
- Exhaustive Events
- Equally Likely Events

MUTUALLY EXCLUSIVE EVENTS

Two events A and B of a single experiment are said to be mutually exclusive or disjoint if and only if they cannot both occur at the same time i.e. they have no points in common.

EXAMPLE-1

When we toss a coin, we get *either* a head *or* a tail, but *not* both at the same time. The two events head and tail are therefore mutually exclusive.

EXAMPLE-2

When a die is rolled, the events 'even number' and 'odd number' are mutually exclusive as we can get either an even number or an odd number in one throw, not both at the same time. Similarly, a student *either* qualifies *or* fails, a person is either a teenager or not a teenager, etc., etc.

Three or more events originating from the same experiment are mutually exclusive if pair wise they are mutually exclusive. If the two events *can* occur at the same time, they are not mutually exclusive, e.g., if we draw a card from an ordinary deck of 52 playing cars, it *can* be both a king and a diamond.

Therefore, kings and diamonds are not mutually exclusive. Speaking of playing cards, it is to be remembered that an ordinary deck of playing cards contains 52 cards arranged in 4 suits of 13 each. The four suits are called diamonds, hearts, clubs and spades; the first two are red and the last two are black. The face values called denominations, of the 13 cards in each suit are ace, 2, 3, …, 10, jack, queen and king. The face values called denominations, of the 13 cards in each suit are ace, 2, 3, …, 10, jack, queen and king. We have discussed the concepts of mutually exclusive events.

Another important concept is that of exhaustive events.

EXHAUSTIVE EVENTS

Events are said to be collectively exhaustive, when the union of mutually exclusive events is equal to the entire sample space S.

EXAMPLES

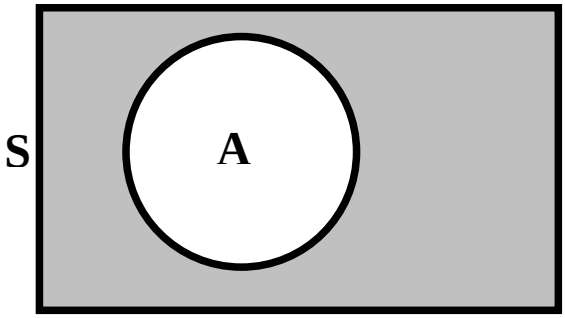
- In the coin-tossing experiment, 'head' and 'tail' are collectively exhaustive events.
- In the die-tossing experiment, 'even number' and 'odd number' are collectively exhaustive events.

In conformity with what was discussed in the last lecture:

PARTITION OF THE SAMPLE SPACE

A group of mutually exclusive and exhaustive events belonging to a sample space is called a partition of the sample space. With reference to any sample space S, events A and $\bar{A}$ form a partition as they are mutually exclusive and their union is the entire sample space. The Venn Diagram below clearly indicates this point.

Venn Diagram



$\bar{A}$ is shaded

EQUALLY LIKELY EVENTS

Two events A and B are said to be equally likely, when one event is as likely to occur as the other. In other words, each event should occur in equal number in repeated trials.

EXAMPLE

When a fair coin is tossed, the head is as likely to appear as the tail, and the proportion of times each side is expected to appear is 1/2.

EXAMPLE

If a card is drawn out of a deck of well-shuffled cards, each card is equally likely to be drawn, and the proportion of times each card can be expected to be drawn in a *very* large number of draws is 1/52.Having discussed basic concepts related to probability theory, we now begin the discussion of THE CONCEPT AND DEFINITIONS OF PROBABILITY. Probability can be discussed from two points of view: the subjective approach, and the objective approach.

SUBJECTIVE OR PERSONALISTIC PROBABILITY

As its name suggests, the subjective or personality probability is a measure of the strength of a person's belief regarding the occurrence of an event A. Probability in this sense is purely subjective, and is based on whatever evidence is available to the individual. It has a disadvantage that two or more persons faced with the same evidence may arrive at different probabilities.
For example, suppose that a panel of three judges is hearing a trial. It is possible that, based on the evidence that is presented, two of them arrive at the conclusion that the accused is guilty while one of them decides that the evidence is NOT strong enough to draw this conclusion. On the other hand, objective probability relates to those situations where everyone will arrive at the same conclusion.
It can be classified into two broad categories, each of which is briefly described as follows:

1. THE CLASSICAL OR 'A PRIORI' DEFINITION OF PROBABILITY

f a random experiment can produce n mutually exclusive and equally likely outcomes, and if m out to these outcomes are considered favorable to the occurrence of a certain event A, then the probability of the event A, denoted by P(A), is defined as the ratio m/n.
Symbolically, we write

$$P(A) = \frac{m}{n} = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

This definition was formulated by the French mathematician P.S. Laplace (1949-1827) and can be very conveniently used in experiments where the total number of possible outcomes and the number of outcomes favorable to an event can be DETERMINED.

Let us now consider a few examples to illustrate the classical definition of probability:

EXAMPLE-1

If a card is drawn from an ordinary deck of 52 playing cards, find the probability that i) the card is a red card, ii) the card is a 10.

SOLUTION:
The total number of possible outcomes is 13+13+13+13 = 52, and we assume that all possible outcomes are equally likely.(It is well-known that an ordinary deck of cards contains 13 cards of diamonds, 13 cards of hearts, 13 cards of clubs, and 13 cards of spades.)
(i) Let A represent the event that the card drawn is a red card.
Then the number of outcomes favorable to the event A is 26 (since the 13 cards of diamonds and the 13 cards of hearts are *red*).

Hence

$$P(A) = \frac{m}{n}$$

$$= \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

$$= \frac{26}{52} = \frac{1}{2}$$

Thus $P(B) = \frac{4}{52} = \frac{1}{13}$.

EXAMPLE-2

A fair coin is tossed three times. What is the probability that at least one head appears?

SOLUTION

The sample space for this experiment is

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

and thus the total number of sample points is 8 i.e. n(S) = 8.Let A denote the event that at least one head appears. Then

A = {HHH, HHT, HTH, THH, HTT, THT, TTH}

Therefore n(A) = 7.

Hence

$$P(A) = \frac{n(A)}{n(S)} = \frac{7}{8}.$$

EXAMPLE-3

Four items are taken at random from a box of 12 items and inspected. The box is *rejected* if more than 1 item is found to be faulty. If there are 3 faulty items in the box, find the probability that the box is *accepted*.

SOLUTION

The sample space S contains $\binom{12}{4} = 495$ Sample points

(Because there are $\binom{12}{4}$ ways of selecting four items out of twelve)

The box contains 3 faulty and 9 good items. The box is accepted if there is (i) no faulty items, or (ii) one faulty item in the sample of 4 items *selected*.
Let A denote the event the number of faulty items chosen is 0 or 1.
Then

$$n(A) = \binom{3}{0}\binom{9}{4} + \binom{3}{1}\binom{9}{3}$$

$$= 126 + 252 = 378 \quad sample \quad po\,\text{int}\,s.$$

$$\therefore \quad P(A) = \frac{m}{n} = \frac{378}{495} = 0.76$$

Hence the probability that the box is accepted is 76% , (in spite of the fact that the box *contains* 3 faulty items).

THE CLASSICAL DEFINITION HAS THE FOLLOWING SHORTCOMINGS

- This definition is said to involve circular reasoning as the term equally likely really means equally probable.
- Thus probability is defined by introducing concepts that presume a *prior* knowledge of the *meaning* of probability.
- This definition becomes vague when the possible outcomes are INFINITE in number, or uncountable.
- This definition is NOT applicable when the assumption of equally likely does *not* hold. And the fact of the matter is that there are NUMEROUS situations where the assumption of equally likely cannot hold.

And these are the situations where we have to look for another definition of probability!

THE **RELATIVE FREQUENCY DEFINITION OF PROBABILITY**

The essence of this definition is that if an experiment is repeated a large number of times under (more or less) identical conditions, and if the event of our interest occurs a certain number of times, then the *proportion* in which this event occurs is regarded as the probability of that event.
For example, we know that a large number of students sit for the matric examination every year. Also, we know that a certain proportion of these students will obtain the first division, a certain proportion will obtain the second division, --- and a certain proportion of the students will fail.
Since the total number of students appearing for the matric exam is very large, hence:

- The proportion of students who obtain the first division --- this proportion can be regarded as the *probability* of obtaining the first division,
- The proportion of students who obtain the second division --- this proportion can be regarded as the *probability* of obtaining the second division, and so on.

LECTURE NO. 19

- Relative Frequency Definition of Probability
- Axiomatic Definition of Probability
- Laws of Probability
    - Rule of Complementation
    - Addition Theorem

THE RELATIVE FREQUENCY DEFINITION OF PROBABILITY ('A POSTERIORI' DEFINITION OF PROBABILITY)

If a random experiment is repeated a large number of times, say n times, under identical conditions and if an event A is observed to occur m times, then the probability of the event A is defined as the LIMIT of the relative frequency m/n as n tends to infinitely.
Symbolically, we write

$$P(A) = \underset{n\to\infty}{Lim}\ \frac{m}{n}$$

The definition assumes that as n increases indefinitely, the ratio m/n tends to become stable at the numerical value P(A).
The relationship between relative frequency and probability can also be represented as follows:

Relative Frequency $\rightarrow$ Probability
as $n \rightarrow \infty$

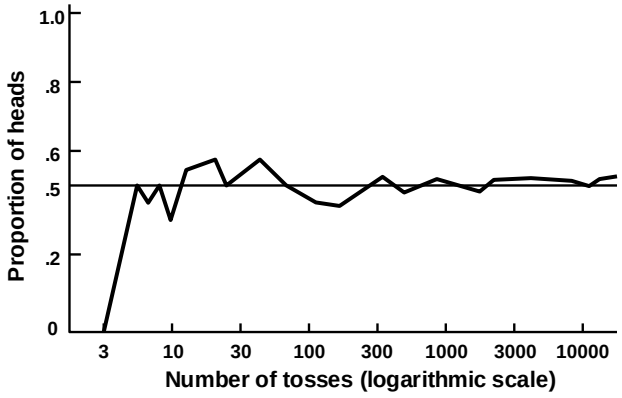
As its name suggests, the relative frequency definition relates to the relative frequency with which are event occurs in the *long run.* In situations where we can say that an experiment has been repeated a very large number of times, the *relative frequency definition* can be applied.
As such, this definition is very useful in those practical situations where we are interested in computing a probability in numerical form but where the classical definition cannot be applied.(Numerous real-life situations are such where various possible outcomes of an experiment are NOT equally likely). This type of probability is also called empirical probability as it is based on *EMPIRICAL* evidence i.e. on *OBSERVATIONAL* data.
It can also be called ***STATISTICAL PROBABILITY*** for it is this very probability that forms the basis of mathematical statistics.
Let us try to understand this concept by means of two examples:

- from a coin-tossing experiment and
- From data on the numbers of boys and girls born.

EXAMPLE-1

Coin-Tossing:
No one can tell which way a coin will fall but we expect the proportion of leads and tails after a large no. of tosses to be nearly equal. An experiment to demonstrate this point was performed by Kerrich in Denmark in 1946. He tossed a coin 10,000 times, and obtained altogether 5067 heads and 4933 tails.
The behavior of the proportion of heads throughout the experiment is shown as in the following figure:

The proportion; of heads in a sequence of tosses of a coin (Kerrich, 1946):



As you can see, the curve fluctuates *widely* at first, but begins to settle down to a more or less *stable* value as the number of spins increases. It seems reasonable to suppose that the fluctuations would continue to diminish if the experiment were continued *indefinitely*, and the proportion of heads would cluster more and more *closely* about a *limiting* value which would be *very near*, if not exactly, one-half. This hypothetical *limiting* value is the (statistical) probability of heads.

Let us now take an example closely related to our *daily* lives --- that relating to the sex ratio:-

In this context, the first point to note is that it has been known since the eighteenth century that in *reliable* birth statistics based on sufficiently *large* numbers (in at least some parts of the world), there is always a slight *excess* of boys, Laplace records that, among the 215,599 births in thirty districts of France in the years 1800 to 1802, there were 110,312 boys and 105,287 girls. The proportions of boys and girls were thus 0.512 and 0.488 respectively (indicating a slight *excess* of boys over girls).In a *smaller* number of births one would, however, expect considerable *deviations* from these proportions. This point can be illustrated with the help of the following example:

EXAMPLE-2

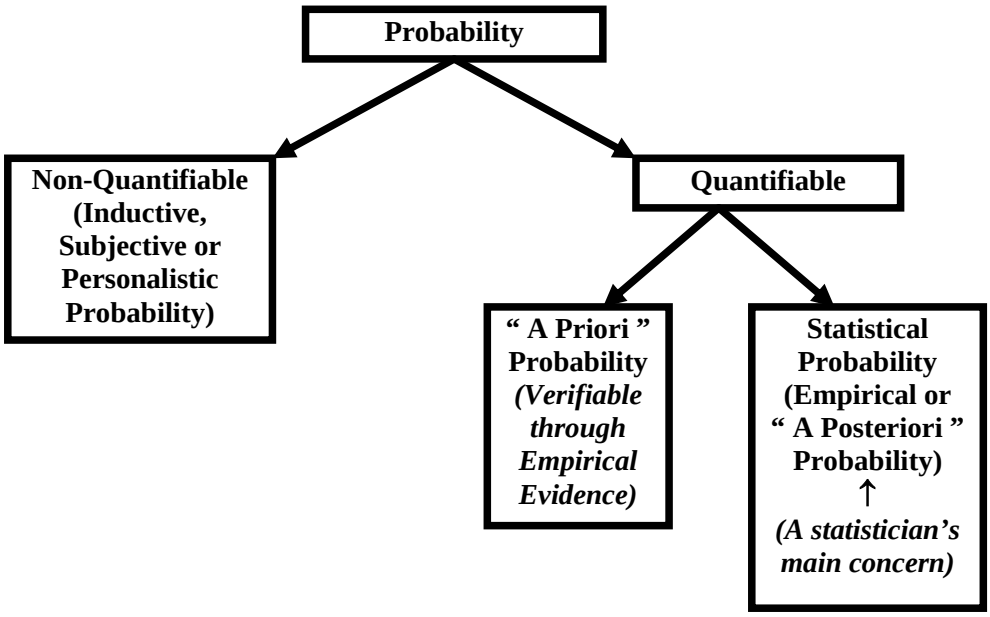
The following table shows the proportions of male births that have been worked out for the major regions of England as well as the rural districts of Dorset (for the year 1956).

Proportions of Male Births in various Regions and Rural Districts of England in 1956

Region of England	Proportion of Male Births	Rural Districts of Dorset	Proportion of Male Births
Northern	.514	Beaminster	.38
E. & W. Riding	.513	Blandford	.47
North Western	.512	Bridport	.53
North Midland	.517	Dorchester	.50
Midland	.514	Shaftesbury	.59
Eastern	.516	Sherborne	.44
London and S. Eastern	.514	Sturminster	.54
Southern	.514	Wareham and Purbeck	.53
South Western	.513	Wimborne & Cranborne	.54
Whole country	.514	All Rural District's of Dorset	.512

(Source: Annual Statistical Review)

As you can see, the figures for the rural districts of Dorset, based on about 200 births each, fluctuate between 0.38 and 0.59. While those for the major regions of England, which are each based on about 100,000 births, do not fluctuate much, rather, they range between 0.512 and 0.517 only. The larger sample size is clearly the reason for the greater constancy of the latter. We can imagine that if the sample were increased indefinitely, the proportion of boys would tend to a *limiting* value which is unlikely to differ much from 0.514, the proportion of male births for the *whole* country. This hypothetical *limiting* value is the (statistical) *probability* of a male birth. The overall discussion regarding the various ways in which probability can be defined is presented in the following diagram:



As far as *quantifiable* probability is concerned, in those situations where the various possible outcomes of our experiment are equally likely, we can compute the probability *prior* to actually conducting the experiment --- otherwise, as is generally the case, we can compute a probability only *after* the experiment has been conducted (and this is why it is also called 'a posteriori' probability*). Non-quantifiable probability is the one that is called Inductive Probability.*
It refers to the degree of belief which it is reasonable to place in a proposition on *given* evidence.
An important point to be noted is that it is difficult to express inductive probabilities numerically — to construct a numerical scale of inductive probabilities, with 0 standing for impossibility and for logical certainty. An important point to be noted is that it is difficult to express inductive probabilities numerically — to construct a numerical scale of inductive probabilities, with 0 standing for impossibility and for logical certainty. Most statisticians have arrived at the conclusion that inductive probability cannot, in general, he measured and, therefore cannot be use in the mathematical theory of statistics.
This conclusion is not, perhaps, very surprising since there seems no reason why rational degree of belief should be measurable any more than, say, degrees of beauty. Some paintings are very beautiful, some are quite beautiful, and some are ugly, but it would be observed to try to construct a numerical scale of beauty, on which Mona Lisa had a beauty value of 0.96.Similarly some propositions are highly probable, some are quite probable and some are improbable, but it does not seem possible to construct a numerical scale of such (inductive) probabilities .Because of the fact that inductive probabilities are not quantifiable and cannot be employed in a mathematical argument, this is the reason why the usual methods of statistical inference such as tests of significance and confidence interval are based entirely on the concept of statistical probability. Although we have discussed three different ways of defining probability, the most formal definition is yet to come.
This is The Axiomatic Definition of Probability.

THE AXIOMATIC DEFINITION OF PROBABILITY

This definition, introduced in 1933 by the Russian mathematician Andrei N. Kolmogrov, is based on a set of AXIOMS. Let S be a sample space with the sample points E1, E2, … Ei, …En. To each sample point, we assign a real number, denoted by the symbol P(Ei), and called the probability of Ei, that must satisfy the following basic axioms:

Axiom 1:

For any event Ei,
$0 < P(Ei) < 1$.

Axiom 2:

$P(S) = 1$
for the sure event S.

Axiom 3:

If A and B are mutually exclusive events (subsets of S), then

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

It is to be emphasized that According to the axiomatic theory of probability:

SOME probability defined as a non-negative real number is to be ATTACHED to each sample point Ei *such that the sum of all such numbers must equal ONE*. The ASSIGNMENT of probabilities may be based on past evidence or on some other underlying conditions. (If this assignment of probabilities is based on past evidence, we are talking about EMPIRICAL probability, and if this assignment is based on underlying conditions that ensure that the various possible outcomes of a random experiment are EQUALLY LIKELY, then we are talking about the CLASSICAL definition of probability. Let us consider another example:

EXAMPLE

Table given below shows the numbers of births in England and Wales in 1956 classified by (a) sex and (b) whether live born or stillborn.

Table-1

Number of births in England and Wales in 1956 by sex and whether live- or still born
(Source Annual Statistical Review)

	Liveborn	Stillborn	Total
Male	359,881 (A)	8,609 (B)	368,490
Female	340,454 (B)	7,796 (D)	348,250
Total	700,335	16,405	716,740

There are four possible events in this double classification:

- Male livebirth (denoted by A),
- Male stillbirth (denoted by B),
- Female livebirth (denoted by C)
- Female stillbirth (denoted by D),

The relative frequencies corresponding to the figures of Table-1 are given in Table-2:

Table-2

Proportion of births in England and Wales in 1956 by sex and whether live- or stillborn
(Source *Annual Statistical Review*)

	Liveborn	Stillborn	Total
Male	.5021	.0120	.5141
Female	.4750	.0109	.4859
Total	.9771	.0229	1.0000

The total number of births is large enough for these relative frequencies to be treated for all practical purposes as *PROBABILITIES*.

Let us denote the compound events 'Male birth' and 'Stillbirth' by the letters M and S. Now a male birth occurs whenever either a male live birth or a male stillbirth occurs, and so the proportion of male birth, regardless of whether they are live-or stillborn, is equal to the sum of the proportions of these two types of birth; that is to say,

$$p(M) = p(A \text{ or } B) = p(A) + p(B)$$
$$= .5021 + .0120 = .5141$$

Similarly, a stillbirth occurs whenever either a male stillbirth or a female stillbirth occurs and so the proportion of stillbirths, regardless of sex, is equal to the sum of the proportions of these two events:

$$p(S) = p(B \text{ or } D) = p(B) + p(D)$$
$$= .0120 + .0109 = .0229$$

Let us now consider some basic LAWS of probability. These laws have important applications in solving probability problems.

LAW OF COMPLEMENTATION

If $\overline{A}$ is the complement of an event A relative to the sample space S, then

$$P(\overline{A}) = 1 - P(A).$$

Hence the probability of the complement of an event is equal to one minus the probability of the event. Complementary probabilities are very useful when we are wanting to solve questions of the type 'What is the probability that, in tossing two fair dice, at least one even number will appear?'

EXAMPLE

A coin is tossed 4 times in succession. What is the probability that at least one head occurs?

- The sample space S for this experiment consists of 24 = 16 sample points (as each toss can result in 2 outcomes),and
- We assume that each outcome is equally likely.

If we let A represent the event that at least one head occurs, then A will consist of MANY sample points, and the process of computing the probability of this event will become somewhat cumbersome! So, instead of denoting this particular event by A, let us denote its complement i.e. "No head" by A.
Thus the event A consists of the SINGLE sample point {TTTT}.
Therefore P(A) = 1/16.
Hence by the law of complementation, we have

$$P(\overline{A}) = 1 - P(A) = 1 - \frac{1}{16} = \frac{15}{16}.$$

The next law that we will consider is the Addition Law or the General Addition Theorem of Probability:

ADDITION LAW

If A and B are any two events defined in a sample space S, then
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
In words, this law may be stated as follows:
"If two events A and B are not mutually exclusive, then the probability that at least one of them occurs, is given by the sum of the separate probabilities of events A and B minus the probability of the joint event $A \cap B$.**"**

LECTURE NO. 20

- Application of Addition Theorem
- Conditional Probability
- Multiplication Theorem

First of all, let us consider in some detail the Addition Law or the General Addition Theorem of Probability:

ADDITION LAW

If A and B are any two events defined in a sample space S, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
In words, this law may be stated as follows:

"If two events A and B are not mutually exclusive, then the probability that at least one of them occurs, is given by the sum of the separate probabilities of events A and B minus the probability of the joint event $A \cap B$."

EXAMPLE

If one card is selected at random from a deck of 52 playing cards, what is the probability that the card is a club or a face card or both?
Let A represent the event that the card selected is a club, B, the event that the card selected is a face card, and $A \cap B$, the event that the card selected is both a club and a face card. Then we need $P(A \cup B)$
Now $P(A) = 13/52$, as there are 13 clubs, $P(B) = 12/52$, as there are 12 faces cards,
$P(A \cap B) = 3/52$, since 3 of clubs are also face cards.
Therefore the desired probability is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 13/52 + 12/52 - 3/52$$
$$= 22/52.$$

COROLLARY-1

If A and B are mutually exclusive events, then
$P(A \cup B) = P(A) + P(B)$ (Since $A \cap B$ is an impossible event, hence $P(A \cap B) = 0$)

EXAMPLE

Suppose that we toss a pair of dice, and we are interested in the event that we get a total of 5 or a total of 11.What is the probability of this event?

SOLUTION

In this context, the first thing to note is that 'getting a total of 5' and 'getting a total of 11' are mutually exclusive events. Hence, we should apply the special case of the addition theorem. If we denote 'getting a total of 5' by A, and 'getting a total of 11' by B, then $P(A) = 4/36$ (since there are four outcomes favorable to the occurrence of a total of 5), and $P(B) = 2/36$ (since there are two outcomes favorable to the occurrence of a total of 11).
Hence the probability that we get a total of 5 or a total of 11 is given by

$$P(A \cup B) = P(A) + P(B)$$
$$= 4/36 + 2/36 = 6/36 = 16.67\%.$$

COROLLARY-2

If $A_1, A_2 \ldots A_k$ are k mutually exclusive events, then the probability that one of them occurs, is the sum of the probabilities of the separate events, i.e.
$P(A_1, \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$.
Let us now consider an interesting example to illustrate the way in which probability problems can be solved:

EXAMPLE

Three horses A, B and C are in a race; A is twice as likely to win as B and B is twice as likely to win as C. What is the probability that A or B wins?
Evidently, the events mentioned in this problem are not equally likely.
Let $P(C) = p$
Then $P(B) = 2p$ as B is twice as likely to win as C.
Similarly
$P(A) = 2P(B) = 2(2p) = 4p$
In this problem, we assume that no two of the horses A, B and C cannot win the race together (i.e. the race cannot end in a draw).
Hence, the events A, B and C are mutually exclusive.

Since A, B and C are mutually exclusive and collectively *exhaustive*, therefore the sum of their probabilities must be equal to 1.

Thus
$p + 2p + 4p = 1$
or $p = 1/7$

$\therefore$ $P(C) = 1/7$,
$P(B) = 2(1/7) = 2/7$,
and $P(A) = 4(1/7) = 4/7$.
Hence
$P(A \cup B) = P(A) + P(B)$
$= 4/7 + 2/7$
$= 6/7$.

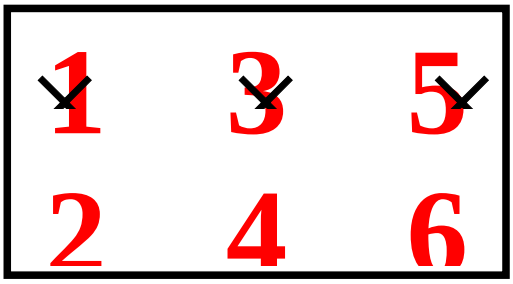
Having discussed the addition theorem in some detail, we would now like to discuss the Multiplication Theorem. But, before we are in a position to take up the multiplication theorem, we need to consider the concept of conditional probability.

CONDITIONAL PROBABILITY

The sample space for an experiment must often be changed when some additional information pertaining to the outcome of the experiment is received.
The effect of such information is to REDUCE the sample space by excluding some outcomes as being impossible which BEFORE receiving the information were believed possible. The probabilities associated with such a reduced sample space are called conditional probabilities.
The following example illustrates the concept of conditional probability

EXAMPLE

Suppose that we toss a fair die. Then the sample space of this experiment is S = {1, 2, 3, 4, 5, 6}. Suppose we wish to know the probability of the outcome that the die shows 6 (say event A).Also, suppose that, before seeing the outcome, we are told that the die shows an EVEN number of dots (say event B).
Then the information that the die shows an even number excludes the outcomes 1, 3 and 5, and thereby reduces the original sample space to a sample space that consists of three outcomes 2, 4 and 6,
i.e. the reduced sample space is
B = {2, 4, 6}.



(The sample space is reduced.)

Then, the desired probability in the reduced sample space B is 1/3.
(since each outcome in the reduced sample space is EQUALLY LIKELY). This probability 1/3 is called the conditional probability of the event A because it is computed under the CONDITION that the die has shown an even number of dots. In other words,
P(die shows 6/die shows even numbers)
= 1/3,
(Where the vertical line is read as given that and the information following the vertical line describes the conditioning event).
Sometimes, it is not very convenient to compute a conditional probability by first determining the number of sample points that belong to the reduced sample space.
In such a situation, we can utilize the following *alternative method of computing a conditional probability*

CONDITIONAL PROBABILITY

If A and B are two events in a sample space S and if P(B) is not equal to zero, then the conditional probability of the event A given that event B has occurred, written as P(A/B), is defined by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Where P (B) > 0
(If P (B) = 0, the conditional probability P(A/B) remains undefined.)
Similarly

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

where P(A) > 0.
It should be noted that P(A/B) SATISFIES all the basic axioms of probability, namely:

- 0 < P(A/B) < 1.
- **ii)** P(S/B) = 1
- P(A1∪A2/B) = P(A1/B) + P(A2/B) (provided that the events A1 and A2 are mutually exclusive).

Let us now apply this concept to a real-world example

EXAMPLE-2

At a certain elementary school in a Western country, the school-record of the past ten years shows that 75% of the students come from a two-parent home and that 20% of the students are low-achievers and belong to two-parent homes. What is the probability that such a randomly selected student will be a low achiever GIVEN THAT he or she comes from a two-parent home?

SOLUTION

Let A denote a low achiever and B a student from a two-parent home. Applying the *relative frequency definition* of probability, we have

P(B) = 0.75 and P(A ∩ B) = 0.20.
Thus, we obtain

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.20}{0.75} = 0.27$$

MULTIPLICATION THEOREM OF PROBABILITY

It is interesting to note that the multiplication theorem is obtained very conveniently from the formula of conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

As discussed earlier, the conditional probability of A given that B has occurred has already been defined as:

$$P(A/B) = \frac{P(A \cap B)}{P(B)},$$ Where P (B) > 0

Multiplying both sides by P (B), we get

P (A ∩ B) = P (B) . P (A/B).

And if we interchange the roles of A and B, we obtain:
P (A ∩ B) = P (A) P (B/A),
Provided P(A) > 0.

MULTIPLICATION LAW

If A and B are any two events defined in a sample space S, then
P (A ∩ B)
= P (A) P (B/A), provided P (A) > 0,
= P (B) P (A/B) provided P (B) > 0.
(The second form is easily obtained by interchanging A and B. This is called the GENERAL rule of multiplication of probabilities. It can be stated as follows:

MULTIPLICATION LAW

"The probability that two events A and B will both occur is equal to the probability that one of the events will occur multiplied by the conditional probability that the other event will occur given that the first event has already occurred."
Let us apply the concept of multiplication theorem to an example

EXAMPLE

A box contains 15 items, 4 of which are defective and 11 is good. Two items are selected. What is the probability that the first is good and the second defective?
Let A represent the event that the first item selected is good, and B, the event that the second items is defective.
Then we need to calculate the probability of the JOINT event $A \cap B$ by the rule
$P(A \cap B) = P(A)P(B/A)$.
We have:

Type of Item	No. of Items
Defective	4
Good	11
Total	15

Since all the items are equally likely to be chosen, hence
$P(A) = 11/15$.
Given the event A has occurred, there remain 14 items of which 4 are defective. Therefore the probability of selecting a defective item after a good item has been selected is 4/14 i.e.

$$P(B/A) = 4/14.$$

Hence

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B/A) \\ &= 11/15 \times 4/14 \\ &= 44/210 \\ &= 0.16. \end{aligned}$$

In this lecture, the concepts of the Addition Theorem and the Multiplication Theorem of probability have been discussed in some detail.
In order to **differentiate** between the situation where the addition theorem is applicable and the situation where the multiplication theorem is applicable, the main point to keep in mind is that whenever we wish to compute the probability that *either* A occurs *or* B occurs, we should think of the Addition Theorem, where as, whenever we wish to compute the probability that *both* A *and* B occur, we should think of the Multiplication Theorem.

LECTURE NO. 21

- Independent and Dependent Events
- Multiplication Theorem of Probability for Independent Events
- Marginal Probability

Before we proceed the concept of independent versus dependent events, let us review the Addition and Multiplication Theorems of Probability that were discussed in the last lecture.
To this end, let us consider an interesting example that illustrates the application of both of these theorems in one problem:

EXAMPLE

A bag contains 10 white and 3 black balls. Another bag contains 3 white and 5 black balls. Two balls are transferred from first bag and placed in the second, and then one ball is taken from the latter.
What is the probability that it is a white ball?
In the beginning of the experiment, we have:

Colour of Ball	No. of Balls in Bag A	No. of Balls in Bag B
White	10	3
Black	3	5
Total	13	8

Let A represent the event that 2 balls are drawn from the first bag and transferred to the second bag. Then A can occur in the following three mutually exclusive ways:
A1 = 2 white balls are transferred to the second bag.
A2 = 1 white ball and 1 black ball are transferred to the second bag.
A3 = 2 black balls are transferred to the second bag.
Then, the total number of ways in which 2 balls can be drawn out of a total of 13 balls is $\binom{13}{2}$.
And, the total number of ways in which 2 white balls can be drawn out of 10 white balls is $\binom{10}{2}$.
Thus, the probability that two white balls are selected from the first bag containing 13 balls (in order to *transfer* to the second bag) is

$$P(A_1) = \binom{10}{2} \div \binom{13}{2} = \frac{45}{78},$$

Similarly, the probability that one white ball and one black ball are selected from the first bag containing 13 balls (in order to *transfer* to the second bag) is

$$P(A_2) = \binom{10}{1}\binom{3}{1} \div \binom{13}{2} = \frac{30}{78},$$

And, the probability that two black balls are selected from the first bag containing 13 balls (in order to *transfer* to the second bag) is

$$P(A_3) = \binom{3}{2} \div \binom{13}{2} = \frac{3}{78}.$$

AFTER having transferred 2 balls from the first bag, the second bag contains
i) 5 white and 5 black balls (if 2 white balls are transferred)

Colour of Ball	No. of Balls in Bag A	No. of Balls in Bag B
White	10 – 2 = 8	3 + 2 = 5
Black	3	5
Total	13 – 2 = 11	8 + 2 = 10

Hence: P(W/A1) = 5/10

ii) 4 white and 6 black balls (if 1 white and 1 black ball are transferred)

Colour of Ball	No. of Balls in Bag A	No. of Balls in Bag B
White	10 – 1 = 7	3 + 1 = 4
Black	3 – 1 = 2	5 + 1 = 4
Total	13 – 2 = 11	8 + 2 = 10

Hence: P(W/A2) = 4/10

iii) 3 white and 7 black balls (if 2 black balls are transferred)

Colour of Ball	No. of Balls in Bag A	No. of Balls in Bag B
White	10	3
Black	3 – 2 = 1	5 + 2 = 7
Total	13 – 2 = 11	8 + 2 = 10

Hence: P(W/A3) = 3/10

Let W represent the event that the WHITE ball is drawn from the second bag after having transferred 2 balls from the first bag.

Then $P(W) = P(A1 \cap W) + P(A2 \cap W) + P(A3 \cap W)$

Now $P(A1 \cap W) = P(A1)\, P(W/A1)$

$= 45/78 \times 5/10$

$= 15/52$

$P(A2 \cap W) = P(A2)\, P(W/A2)$

$= 30/78 \times 4/10$

$= 2/13,$

And

$P(A3 \cap W) = P(A3)\, P(W/A3)$

$= 3/78 \times 3/10$

$= 3/260.$

Hence the required probability is

$P(W)$

$= P(A1 \cap W) + P(A2 \cap W) + P(A3 \cap W)$

$= 15/52 + 2/13 + 3/260$

$= 59/130$

$= 0.45$

INDEPENDENT EVENTS

Two events A and B in the same sample space S, are defined to be independent (or statistically independent) if the probability that one event occurs, is not affected by whether the other event has or has not occurred, that is

$P(A/B) = P(A)$ and $P(B/A) = P(B)$. It then follows that two events A and B are independent if and only if

$P(A \cap B) = P(A)\, P(B)$

and this is known as the *special case* of the Multiplication Theorem of Probability.

RATIONALE

According to the multiplication theorem of probability, we have:

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

Putting P(B/A) = P(B), we obtain

$$P(A \cap B) = P(A)\, P(B)$$

The events A and B are defined to be *DEPENDENT* if $P(A \cap B) \neq P(A) \times P(B)$.

This means that the occurrence of one of the events in some way affects the probability of the occurrence of the other event. Speaking of independent events, it is to be emphasized that two events that are independent, can NEVER be mutually exclusive.

EXAMPLE

Two fair dice, one red and one green, are thrown. Let A denote the event that the red die shows an even number and let B denote the event that the green die shows a 5 or a 6. Show that the events A and B are independent.

The sample space S is represented by the following 36 outcomes:

S = {(1, 1), (1, 2), (1, 3), (1, 5), (1, 6);
(2, 1), (2, 2), (2, 3), (2, 5), (2, 6);
(3, 1), (3, 2), (3, 3), (3, 5), (3, 6);
(4, 1), (4, 2), (4, 3), (4, 5), (4, 6);
(5, 1), (5, 2), (5, 3), (5, 5), (5, 6);
(6, 1), (6, 2), (6, 3), (6, 5), (6, 6) }

Since

A represents the event that red die shows an even number, and B represents the event that green die shows a 5 or a 6,

Therefore $A \cap B$ represents the event that red die shows an even number and green die shows a 5 or a 6.

Since A represents the event that red die shows an even number, hence P(A) = 3/6. Similarly, since B represents the event that green die shows a 5 or a 6, hence P(B) = 2/6.

Now, in order to compute the probability of the joint event $A \cap B$, the first point to note is that, in all, there are 36 possible outcomes when we throw the two dice together, i.e.

S = (1, 1), (1, 2), (1, 3), (1, 5), (1, 6);
(2, 1), (2, 2), (2, 3), (2, 5), (2, 6);
(3, 1), (3, 2), (3, 3), (3, 5), (3, 6);
(4, 1), (4, 2), (4, 3), (4, 5), (4, 6);
(5, 1), (5, 2), (5, 3), (5, 5), (5, 6);
(6, 1), (6, 2), (6, 3), (6, 5), (6, 6)

The joint event $A \cap B$ contains only 6 outcomes out of the 36 possible outcomes.

These are (2, 5), (4, 5), (6, 5), (2, 6), (4, 6), and (6, 6).

and $P(A \cap B) = 6/36$.

Now

P (A) P (B)

$= 3/6 \times 2/6$

$= 6/36$

$= P (A \cap B)$.

Therefore the events A and B are independent. Let us now go back to the example pertaining to live births and stillbirths that we considered in the last lecture, and try to determine whether or not sex of the baby and nature of birth are independent.

EXAMPLE

Table-1 below shows the numbers of births in England and Wales in 1956 classified by (a) sex and (b) whether live born or stillborn.

Table-1

Number of births in England and Wales in 1956 by sex and whether live- or still born
(Source *Annual Statistical Review*)

	Liveborn	Stillborn	Total
Male	359,881 (A)	8,609 (B)	368,490
Female	340,454 (B)	7,796 (D)	348,250
Total	700,335	16,405	716,740

There are four possible events in this double classification:

- Male live birth,
- Male stillbirth,
- Female live birth, and
- Female stillbirth.

The corresponding relative frequencies are given in Table-2.

Table-2

Proportion of births in England and Wales in 1956 by sex and whether live- or stillborn
(Source Annual Statistical Review)

	Liveborn	Stillborn	Total
Male	.5021	.0120	.5141
Female	.4750	.0109	.4859
Total	.9771	.0229	1.0000

As discussed in the last lecture, the total number of births is large enough for these relative frequencies to be treated for all practical purposes as *PROBABILITIES*. The compound events 'Male birth' and 'Stillbirth' may be represented by the letters M and S. If M represents a male birth and S a stillbirth, we find that

$$\frac{n(M \text{ and } S)}{n(M)} = \frac{8609}{368490} = 0.0234$$

This figure is the proportion — and, since the sample size is large, it can be regarded as the *probability* — of males who are still born — in other words, the *CONDITIONAL* probability of a stillbirth *given that* it is a male birth. In other words, the probability of stillbirths *in* males. The corresponding proportion of stillbirths among females is

$$\frac{7796}{348258} = 0.0224.$$

These figures should be contrasted with the OVERALL, or *UNCONDITIONAL*, proportion of stillbirths, which is

$$\frac{16405}{716740} = 0.0229.$$

We observe that the conditional probability of stillbirths among boys is slightly *HIGHER* than the overall proportion. Where as the conditional proportion of stillbirths among girls is slightly *LOWER* than the overall proportion. It can be concluded that sex and stillbirth are statistically *DEPENDENT*, that is to say, the *SEX* of a baby yet to be born *has* an effect, (although a small effect), on its chance of being stillborn. The example, that we just considered point out the concept of *MARGINAL PROBABILITY*.

Let us have another look at the data regarding the live births and stillbirths in England and Wales:

Table-2Proportion of births in England and Wales in 1956 by sex and whether live- or stillborn (Source *Annual Statistical Review*)

	Liveborn	Stillborn	Total
Male	.5021	.0120	.5141
Female	.4750	.0109	.4859
Total	.9771	.0229	1.0000

And, the figures in Table-2 indicate that the probability of male birth is 0.5141, whereas the probability of female birth is 0.4859.Also, the probability of live birth is 0.9771, where as the probability of stillbirth is 0.0229. And since these probabilities appear in the margins of the Table, they are known as *Marginal Probabilities*. According to the above table, the probability that a new born baby is a male and is live born is 0.5021 whereas the probability that a new born baby is a male and is stillborn is 0.0120.Also, as stated earlier; the probability that a new born baby is a male is 0.5141, and, CLEARLY, 0.5141 = 0.5021 + 0.0120. Hence, it is clear that the joint probabilities occurring in any row of the table *ADD UP* to yield the corresponding *marginal* probability. If we reflect upon this situation carefully, we will realize that this equation is totally in accordance with the Addition Theorem of Probability for mutually exclusive events.

P (male birth) = P(male live-born *or* male stillborn)
= P (male live-born) + P (male stillborn)
= 0.5021 + 0.0120
= 0.5141

EXAMPLE

P (stillbirth/male birth)
P (male birth *and* stillbirth)/P(male birth)
=0.0120/0.5141
= 0.0233

LECTURE NO. 22

- Bayes' Theorem
- Discrete Random Variable
  - Discrete Probability Distribution
  - Graphical Representation of a Discrete Probability Distribution
  - Mean, Standard Deviation and Coefficient of Variation of a Discrete Probability Distribution
  - Distribution Function of a Discrete Random Variable.

First of all, let us discuss the BAYES' THEOREM. This theorem deals with conditional probabilities in an interesting way:

BAYES' THEOREM

If events A_1, A_2… A_k form a PARTITION of a sample space S (that is, the events Ai are mutually exclusive and exhaustive (i.e. their union is S)),
and if B is any other event of S such that it can occur ONLY IF ONE OF THE Ai OCCURS, then for any i,

$$P(A_i / B) = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{k} P(A_i)P(B/A_i)},$$

for i = 1, 2, …, k.

Stated differently:
BAYES' THEOREM:
If A_1, A_2... and A_k are mutually exclusive events *of which one must occur*, then

$$P(A_i \mid B) = \frac{P(A_i).P(B \mid A_i)}{P(A_1).P(B \mid A_1) + P(A_2).P(B \mid A_2) + ... + P(A_k).P(B \mid A_k)}$$

If k = 2, we obtain:
Bayes' Theorem for two mutually exclusive events A1 and A2:

$$P(A_i \mid B) = \frac{P(A_i).P(B \mid A_i)}{P(A_1).P(B \mid A_1) + P(A_2).P(B \mid A_2)}$$

Where i = 1, 2.

In other words

$$P(A_1 \mid B) = \frac{P(A_1).P(B \mid A_1)}{P(A_1).P(B \mid A_1) + P(A_2).P(B \mid A_2)}$$

and

$$P(A_2 \mid B) = \frac{P(A_2).P(B \mid A_2)}{P(A_1).P(B \mid A_1) + P(A_2).P(B \mid A_2)}$$

EXAMPLE

In a developed country where cars are tested for the emission of pollutants, 25 percent of all cars emit excessive amounts of pollutants. When tested, 99 percent of all cars that emit excessive amounts of pollutants will fail, but 17 percent of the cars that do not emit excessive amounts of pollutants will also fail. What is the probability that a car that fails the test actually *emits* excessive amounts of pollutants?

SOLUTION

Let A1 denote the event that it emits EXCESSIVE amounts of pollutants, and let A2 denote the event that a car does NOT emit excessive amounts of pollutants. (In other words, A2 is the complement of A1.)
Also, let B denote the event that a car *FAILS* the test.
The first thing to note is that any car will either *emit* or *not* emit excessive amounts of pollutants. In other words, A1 and A2 are mutually exclusive and exhaustive events i.e. A1 and A2 form a PARTITION of the sample space S.
Hence, we are in a position to apply the Bayes' theorem.
We need to calculate P(A1|B), and, according to the Bayes' theorem:

$$P(A_1 \mid B) = \frac{P(A_1).P(B \mid A_1)}{P(A_1).P(B \mid A_1) + P(A_2).P(B \mid A_2)}$$

Now, according to the data given in this problem:

P (A1) = 0.25,

P (A2) = 0.75 (as A2 is simply the complement of A1),

P (B|A1) = 0.99,
and

P (B|A2) = 0.17
Substituting the above values in the Bayes' theorem, we obtain:

$$P(A_1 \mid B) = \frac{P(A_1).P(B \mid A_1)}{P(A_1).P(B \mid A_1) + P(A_2).P(B \mid A_2)}$$
$$= \frac{(0.25)(0.99)}{(0.25)(0.99) + (0.75)(0.17)}$$
$$= \frac{0.2475}{0.2475 + 0.1275}$$
$$= \frac{0.2475}{0.3750}$$
$$= 0.66$$

This is the probability that a car which fails the test ACTUALLY emits excessive amounts of pollutants. The example that we just considered pertained to the simplest case when we have only two mutually exclusive and exhaustive events A1 and A2.
As stated earlier, the Bayes' theorem can be extended to the case of three, four, five or more mutually exclusive and exhaustive events.
Let us consider another example: In the following example, check the percentages of defective bolts from the recorded lecture.

EXAMPLE

In a bolt factory, 25% of the bolts are produced by machine A, 35% are produced by machine B, and the remaining 40% are produced by machine C. Of their outputs, 2%, 4% and 5% respectively are defective bolts. If a bolt is selected at random and found to be defective, what is the probability that it came from machine A?

In this example, we realize that "a bolt is produced by machine A", "a bolt is produced by machine B" and "a bolt is produced by machine C" represent three mutually exclusive and exhaustive events i.e. we can regard them as A1, A2 and A3. The event "defective bolt" represents the event B. Hence, in this example, we need to determine P(A1/B).
The students are encouraged to work on this problem on their own, in order to understand the application and significance of the Bayes' Theorem. This brings us to the *END* of the discussion of various *basic* concepts of probability. We now begin the discussion of a *very important* concept in mathematical statistics, i.e., the concept of *PROBABILITY DISTRIBUTIONS*.
As stated in the very beginning of this course, there are two types of quantitative variables --- the discrete variable, and the continuous variable. Accordingly, we have the discrete probability distribution as well as the continuous probability distribution.
We begin with the discussion of the discrete probability distribution.
In this regard, the first concept that we need to consider is the concept of Random variable.

RANDOM VARIABLE

Such a numerical quantity whose value is determined by the outcome of a random experiment is called a random variable.
For example, if we toss three dice together, and let X denote the number of heads, then the random variable X consists of the values 0, 1, 2, and 3. Obviously, in this example, X is a discrete random variable. Let us now discuss the concept of discrete probability distribution in detail with the help of the following example:
Example:

If a biologist is interested in the number of petals on a particular flower, this number may take the values 3, 4, 5, 6, 7, 8, 9, and each one of these numbers will have its own probability.

Suppose that upon observing a large no. of flowers, say 1000 flowers, of that particular species, the following results are obtained:

No. of Petals X	f
3	50
4	100
5	200
6	300
7	250
8	75
9	25
	1000

Since 1000 is quite a large number, hence the proportions f/∑f can be regarded as probabilities and hence we can write

No. of Petals X	**P(x)**
$x_1 = 3$	0.05
$x_2 = 4$	0.10
$x_3 = 5$	0.20
$x_4 = 6$	0.30
$x_5 = 7$	0.25
$x_6 = 8$	0.075
$x_7 = 9$	0.025
	1

PROPERTIES OF A DISCRETE PROBABILITY DISTRIBUTION

- (1) $0 \le P(X_i) \le 1$ • for each Xi (i = 1, 2, … 7)
- (2) $\sum p(X_i) = 1$

And, since the number of petals on a leaf can only be a *whole* number, hence the variable X is known as a *discrete* random variable, and the probability distribution of this variable is known as a *DISCRETE* probability distribution.
In other words, Any discrete variable that is associated with a random experiment, and attached to whose various values are various probabilities (Such that $\sum_{i=1}^{\infty} P(X_i) = 1$)

is known as a Discrete Random Variable, and its probability distribution is known as a Discrete Probability Distribution. Just as we can depict a frequency distribution graphically, we can draw the GRAPH of a probability distribution.

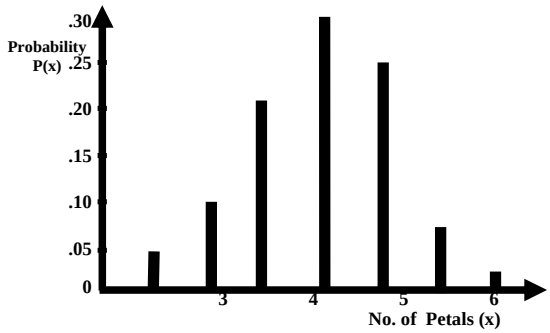
EXAMPLE

Going back to the probability distribution of the number of petals on the flowers of a particulars species, i.e.:

No. of Petals X	P(x)
$x_1 = 3$	0.05
$x_2 = 4$	0.10
$x_3 = 5$	0.20
$x_4 = 6$	0.30
$x_5 = 7$	0.25
$x_6 = 8$	0.075
$x_7 = 9$	0.025
	1

This distribution can be represented in the form of a line chart.

Line Chart Representation of the



Evidently, this particular probability distribution is approximately symmetric. In addition, this graph clearly shows that, just as in the case of a frequency distribution, every discrete probability distribution has a *CENTRAL* point and a *SPREAD*. Hence, similar to a frequency distribution, the discrete probability distribution has a *MEAN* and a *STANDARD DEVIATION*. How do we *calculate* the mean and the standard deviation of a probability distribution?
Let us first consider the computation of the MEAN:
We know that in the case of a frequency distribution such as

X	f
1	1
2	2
3	4
4	2
5	1

the mean is given by

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{\sum Xf}{\sum f}$$

In case of a discrete probability distribution, such as the one that we have been considering i.e

No. of Petals X	P(x)
$x_1 = 3$	0.05
$x_2 = 4$	0.10
$x_3 = 5$	0.20
$x_4 = 6$	0.30
$x_5 = 7$	0.25
$x_6 = 8$	0.075
$x_7 = 9$	0.025
	1

the mean is given by:

$$\mu = E(X) = \frac{\sum XP(X)}{\sum p(X)} = \frac{\sum XP(X)}{1} = \sum XP(X)$$

Hence we construct the column of XP(X), as shown below:

No. of Petals x	P(x)	xP(x)
$x_1 = 3$	0.05	0.15
$x_2 = 4$	0.10	0.40
$x_3 = 5$	0.20	1.00
$x_4 = 6$	0.30	1.80
$x_5 = 7$	0.25	1.75
$x_6 = 8$	0.075	0.60
$x_7 = 9$	0.025	0.225
Total	1	5.925

Hence $\mu = E(X) = \sum XP(X) = 5.925$ i.e. the mean of the given probability distribution is 5.925. In other words, considering a very large number of flowers of that particular species, we would expect that, on the average, a flower contains 5.925 petals --- or, *rounding* this number, 6 petals. This interpretation points to the reason why the mean of the probability distribution of a random variable X is technically called the EXPECTED VALUE of the random variable X. ("Given that the probability that the flower has 3 petals is 5%, the probability that the flower has 4 petals is 10%, and so *ON*, we *EXPECT* that on the *average* a flower contains 5.925 petals.)

COMPUTATION OF THE STANDARD DEVIATION

Just as in case of a frequency distribution, we have

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

$$= \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2} = \sqrt{\frac{\sum X^2 f}{\sum f} - \left(\frac{\sum Xf}{\sum f}\right)^2}$$

Similarly, in case of a probability distribution, we have

$$\sigma = \text{S.D.}(X) = \sqrt{\frac{\sum X^2 P(X)}{\sum P(X)} - \left[\frac{\sum XP(X)}{\sum P(X)}\right]^2}$$

$$= \sqrt{\sum X^2 P(X) - [\sum XP(X)]^2}$$

$$\left(\because \quad \sum P(X) = 1\right)$$

In the above example

No. of Petals x	P(x)	xP(x)	$x^2P(x)$
$x_1 = 3$	0.05	0.15	0.45
$x_2 = 4$	0.10	0.40	1.60
$x_3 = 5$	0.20	1.00	5.00
$x_4 = 6$	0.30	1.80	10.80
$x_5 = 7$	0.25	1.75	12.25
$x_6 = 8$	0.075	0.60	4.80
$x_7 = 9$	0.025	0.225	2.025
Total	1	5.925	36.925

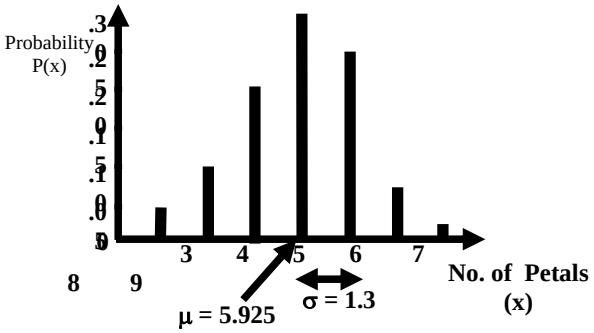
Hence

$$S.D.(X) = \sqrt{36.925 - (5.925)^2}$$

$$= \sqrt{36.925 - 35.106}$$

$$= \sqrt{1.819} = 1.3$$

Graphical Representation:



Now that we have both the mean and the standard deviation, we are in a position to compute the *coefficient of variation* of this distribution:

Coefficient of Variation

$$C.V. = \frac{\sigma}{\mu} \times 100$$

$$= \frac{1.3}{5.925} \times 100$$

$$= 21.9\ \%$$

Let us consider another example to understand the concept of discrete probability distribution.

EXAMPLE

a) Find the probability distribution of the sum of the dots when two fair dice are thrown
b) Use the probability distribution to find the probabilities of obtaining (i) a sum that is greater than 8, and (ii) a sum that is greater than 5 but less than or equal to 10.

SOLUTION

a) The sample space S is represented by the following 36 outcomes:

S = {(1, 1), (1, 2), (1, 3), (1, 5), (1, 6);
(2, 1), (2, 2), (2, 3), (2, 5), (2, 6);
(3, 1), (3, 2), (3, 3), (3, 5), (3, 6);
(4, 1), (4, 2), (4, 3), (4, 5), (4, 6);
(5, 1), (5, 2), (5, 3), (5, 5), (5, 6);
(6, 1), (6, 2), (6, 3), (6, 5), (6, 6) }

Since each of the 36 outcomes is equally likely to occur, therefore each outcome has probability 1/36.
Let X be the random variable representing the sum of dots which appear on the dice. Then the values of the r.v. are 2, 3, 4… 12.
The probabilities of these values are computed as below:

$f(2) = P(X = 2) = P[\{1,1\}] = \frac{1}{36}$, as there is only one outcome resulting in a sum of 2,

$f(3) = P(X = 3) = P[\{(1,2), (2,1)\}] = \frac{2}{36}$,

$f(4) = P(X = 4) = P[\{(1,3), (2,2), (3,1)\}] = \frac{3}{36}$,

Similarly

$f(5) = \frac{4}{36}, f(6) = \frac{5}{36}, f(7) = \frac{6}{36}, f(8) = \frac{5}{36}, f(9) = \frac{4}{36},$

$f(10) = \frac{3}{36}, f(11) = \frac{2}{36}$ and $f(12) = \frac{1}{36}$.

Therefore the desired probability distribution of the r.v X is

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The probabilities in the above table clearly indicate that if we draw the line chart of this distribution, we will obtain a triangular-shaped graph. The students are encouraged to draw the graph of this probability distribution, in order to be able to develop a visual picture in their minds.

b) Using the probability distribution, we get the required probabilities as follows:

i) P(a sum that is greater than 8)

$= P(X > 8)$

$= P(X=9) + P(X=10) + P(X=11) + P(X=12)$

$= f(9) + f(10) + f(11) + f(12)$

$= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36}$

ii) P(a sum that is greater than 5 but less than or equal to 10)

$= P(5 < X \leq 10)$

$= P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$

$= f(6) + f(7) + f(8) + f(9) + f(10)$

$= \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} = \frac{23}{36}.$

Next, we consider the concept of the DISTRIBUTION FUNCTION of a discrete random variable:

DISTRIBUTION FUNCTION

The distribution function of a random variable X, denoted by F(x), is defined by F(x) = P(X < x).
The function F(x) gives the probability of the event that X takes a value LESS THAN OR EQUAL TO a specified value x. The distribution function is abbreviated to *d.f.* and is also called the *cumulative distribution function (cdf)* as it is the cumulative probability function of the random variable X from the smallest value upto a *specific* value x.
Let us illustrate this concept with the help of the same example that we have been considering --- that of the probability distribution of the sum of the dots when two fair dice are thrown. As explained earlier, the probability distribution of this example is:

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The term 'distribution function' implies the cumulating of the probabilities similar to the cumulation of frequencies in the case of the frequency distribution of a discrete variable.

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F(x_i)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

If we are interested in finding the probability that we obtain a sum of five or less, the column of cumulative probabilities immediately indicates that this probability is 10/36.

LECTURE NO. 23

- Graphical Representation of the Distribution Function of a Discrete Random Variable
- Mathematical Expectation
- Mean, Variance and Moments of a Discrete Probability Distribution
- Properties of Expected Values

First, let us consider the concept of the DISTRIBUTION FUNCTION of a discrete random variable.

DISTRIBUTION FUNCTION

The distribution function of a random variable X, denoted by F(x), is defined by $F(x) = P(X < x)$. The function F(x) gives the probability of the event that X takes a value LESS THAN OR EQUAL TO a specified value x. The distribution function is abbreviated to d.f. and is also called the cumulative distribution function (cdf) as it is the cumulative probability function of the random variable X from the smallest value up to a specific value x.

EXAMPLE

Find the probability distribution and distribution function for the number of heads when 3 balanced coins are tossed. Depict both the probability distribution and the distribution function graphically. Since the coins are balanced, therefore the equally probable sample space for this experiment is

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

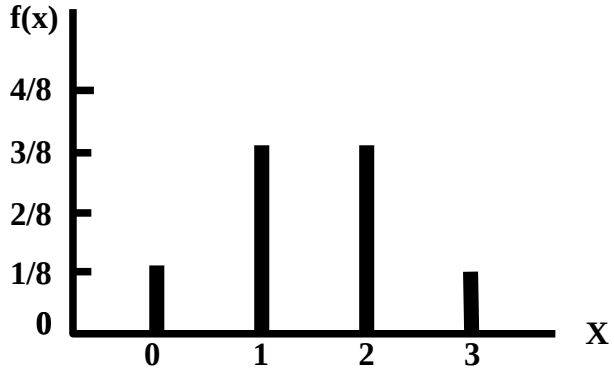
Let X be the random variable that denotes the number of heads.
Then the values of X are 0, 1, 2 and 3.
And their probabilities are:

f (0) = P(X = 0)
= P [{TTT}] = 1/8
f(1) = P (X = 1)
= P [{HTT, THT, TTH}] = 3/8
f(2) = P (X = 2)
= P [{HHT, HTH, THH}] = 3/8
f(2) = P (X = 3)
= P [{HHH}] = 1/8

Expressing the above information in the tabular form, we obtain the desired probability distribution of X as follows:

Number of Heads (x_i)	Probability $f(x_i)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$
Total	1

The line chart of the above probability distribution is as follows:



In order to obtain the distribution function of this random variable, we compute the cumulative probabilities as follows:

Number of Heads (x_i)	Probability $f(x_i)$	Cumulative Probability $F(x_i)$
0	$\frac{1}{8}$	$\frac{1}{8}$
1	$\frac{3}{8}$	$\frac{1}{8}+\frac{3}{8}=\frac{4}{8}$
2	$\frac{3}{8}$	$\frac{4}{8}+\frac{3}{8}=\frac{7}{8}$
3	$\frac{1}{8}$	$\frac{7}{8}+\frac{1}{8}=1$

Hence the desired distribution function is

$$F(x)=\begin{cases}0, & for\ \ x<0\\ \frac{1}{8}, & for\ \ 0\le x<1\\ \frac{4}{8}, & for\ \ 1\le x<2\\ \frac{7}{8}, & for\ \ 2\le x<3\\ 1, & for\ \ x\ge 3\end{cases}$$

Why has the distribution function been expressed in this manner? The answer to this question is:

INTERPRETATION

If $x < 0$, we have $P(X < x) = 0$, the reason being that it is not possible for our random variable X to assume value less than zero.(The minimum number of heads that we can have in tossing three coins is zero.)
If $0 < x < 1$, we note that it is not possible for our random variable X to assume any value between zero and one. (We will have no head or one head but we will NOT have 1/3 heads or 2/5 heads!)
Hence, the probabilities of all such values will be zero, and hence we will obtain a situation which can be explained through the following table:

Number of Heads (x_i)	Probability $f(x_i)$	Cumulative Probability $F(x_i)$
0	$\frac{1}{8}$	$\frac{1}{8}$
0.2	**0**	$\frac{1}{8}+0=\frac{1}{8}$
0.4	**0**	$\frac{1}{8}+0=\frac{1}{8}$
0.6	**0**	$\frac{1}{8}+0=\frac{1}{8}$
0.8	**0**	$\frac{1}{8}+0=\frac{1}{8}$
1	$\frac{3}{8}$	$\frac{1}{8}+\frac{3}{8}=\frac{4}{8}$

The above table clearly shows that the probability that X is LESS THAN any value lying between zero and 0.9999… will be equal to the probability of X = 0 i.e. For $0 < x < 1$,

Similarly, $$P(X < x) = P(X = 0) = \frac{1}{8};$$

- For $1 < x < 2$, we have

$$P(X < x) = P(X = 0) + P(X = 1)$$
$$= \frac{1}{8} + \frac{3}{8} = \frac{4}{8};$$

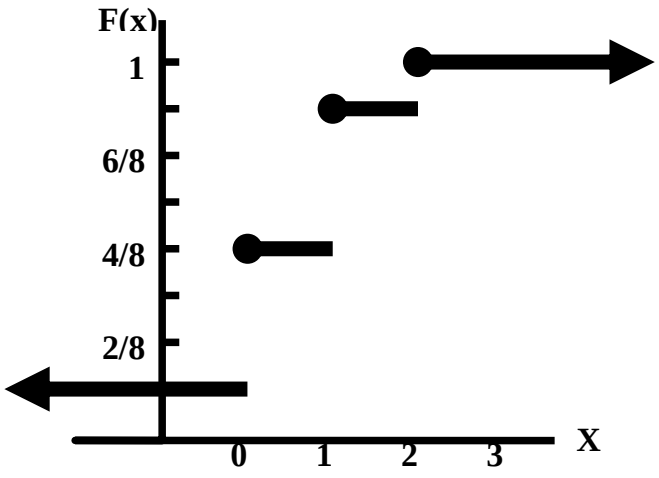
- For $2 < x < 3$, we have

$$P(X < x) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8};$$

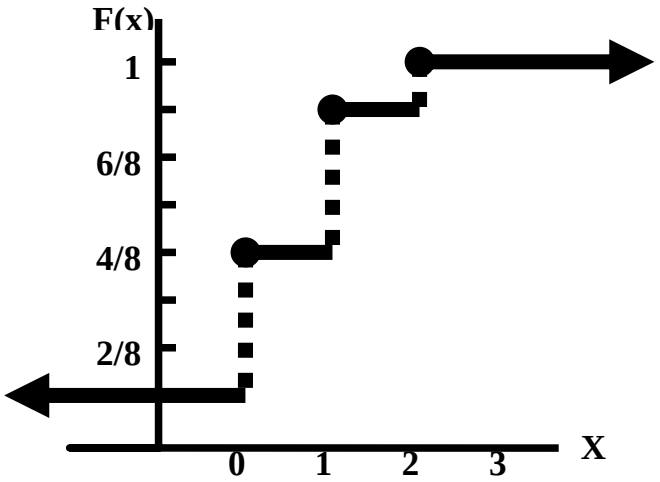
And, finally, for $x > 3$, we have

$$P(X < x) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{8}{8} = 1.$$

Hence, the graph of the DISTRIBUTION FUNCTION is as follows:



As this graph resembles the steps of a staircase, it is known as a step function. It is also known as a jump function (as it takes jumps at integral values of X).In some books, the graph of the distribution function is given as shown in the following figure:



In what way do we interpret the above distribution function from a REAL-LIFE point of view? If we toss three balanced coins, the probability that we obtain at the most one head is 4/8, the probability that we obtain at the most two heads is 7/8, and so on. Let us consider another interesting example to illustrate the concepts of a discrete probability distribution and its distribution function:

EXAMPLE

A large store places its last 15 clock radios in a clearance sale. Unknown to any one, 5 of the radios are defective. If a customer tests 3 different clock radios selected at random, what is the probability distribution of X, where X represent the number of defective radios in the sample?

SOLUTION

We have:

Type of Clock Radio	Number of Clock Radios
Good	10
Defective	5
Total	15

The total number of ways of selecting 3 radios out of 15 is $\binom{15}{3}$.

Also, the total number of ways of selecting 3 good radios (and no defective radio) is $\binom{10}{3}\binom{5}{0}$.
Hence, the probability of X = 0 is

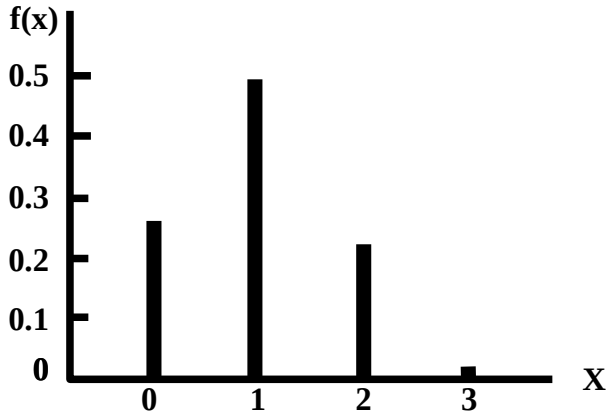
$$\frac{\binom{10}{3}\binom{5}{0}}{\binom{15}{3}} = 0.26.$$

The probabilities of X = 1, 2, and 3 are computed in a similar way. Hence, we obtain the following probability distribution:

Number of defective clock radios in the sample X	Probability f(x)
0	0.26
1	0.49
2	0.22
3	0.02
Total	0.99 ≈1

The line chart of this distribution is:

LINE CHART



As indicated by the above diagram, it is not necessary for a probability distribution to be symmetric; it can be positively or negatively skewed. The distribution function of the above probability distribution is obtained as follows:

Number of defective clock radios in the sample X	f(x)	F(x)
0	0.26	0.26
1	0.49	0.75
2	0.22	0.97
3	0.02	0.99 ≈ 1
Total	0.99 ≈ 1	

INTERPRETATION

The probability that the sample of 3 clock radios contains at the most one defective radio is 0.75, the probability that the sample contains at the most two defective radios is 0.97, and so on.
Let a discrete random variable X have possible values x1, x2, …, xn with corresponding probabilities f(x1), f(x2), …, f(xn) such that Σf(xi) =1. Then the mathematical expectation or the expectation or the expected value of X, denoted by E(x), is defined as
E(X) = x1f(x1) + x2f(x2) + … + xnf(xn)

$$= \sum_{i=1}^{n} x_i f(x_i),$$

E(X) is also called the mean of X and is usually denoted by the letter μ.

The expression

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

may be regarded as a weighted mean of the variable's possible values x1, x2, …,xn, each being weighted by the respective probability.
In case the values are equally likely,

$$E(X) = \frac{1}{n} \sum x_i,$$

Which represents the ordinary arithmetic mean of the n possible values
It should be noted that E(X) is the average value of the random variable X over a very large number of trials.

EXAMPLE
If it rains, an umbrella salesman can earn $ 30 per day. If it is fair, he can lose $ 6 per day. What is his expectation if the probability of rain is 0.3?

SOLUTION

Let X represents the number of dollars the salesman earns. Then X is a random variable with possible values 30 and –6, (where -6 corresponds to the fact that the salesman loses), and the corresponding probabilities are 0.3 and 0.7 respectively. Hence, we have:

EVENT	AMOUNT EARNED ($) x	PROBABILITY P(x)
Rain	30	0.3
No Rain	–6	0.7
	Total	1

In order to compute the expected value of X, we carry out the following computation

EVENT	AMOUNT EARNED ($) x	PROBABILITY P(x)	xP(x)
Rain	30	0.3	9.0
No Rain	–6	0.7	-4.2
	Total	1	4.8

Hence
E(X) = $ 4.80 per day
i.e. on the average, the salesman can expect to earn 4.8 dollars per day.
Until now, we have considered the mathematical expectation of the random variable X.

But, in many situations, we may be interested in the mathematical expectation of some FUNCTION of X:

EXPECTATION OF A FUNCTION OF A RANDOM VARIABLE

Let H(X) be a function of the random variable X. Then H(X) is also a random variable and also has an expected value, (as any function of a random variable is also a random variable). If X is a discrete random variable with probability distribution f(x), then, since H(X) takes the value H(xi) when X = xi, the expected value of the function H(X) is
E[H(X)] = H(x1) f(x1) + H(x2)f(x2) + … + H(xn) f(xn)

$$= \sum_i H(x_i) f(x_i),$$

Provided the series converges absolutely. Again, if H(X) = (X - μ)2,where μ is the population mean, then
E(X – μ)2 = Σ(xi - μ)2 f(x).
We call this expected value the variance and denote it by Var (X) or σ2.
And, since
E(X – μ)2 = E(X2) – [E(X)]2,
hence the short cut formula for the variance is
σ2 = E(X2) – [E(X)]2
The positive square root of the variance, a before, is called the standard deviation. More generally, if
H(X) = Xk, k = 1, 2, 3… then
E(Xk) = Σxik f(x)
which we call the kth moment about the origin of the random variable X and we denote it by ′μk. Similarly, if H(X) = (X – μ)k, k = 1, 2, 3, …, then we get an expected value, called the kth moment about the mean of the random variable X, which we denote by μk. That is:
μk = E(X – μ)k = Σ(xi – μ)k f(x)
The skewness of a probability distribution is often measured by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

and kurtosis by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}.$$

These moment-ratios assist us in determining the Skewness and kurtosis of our probability distribution in exactly the same way as was discussed in the case of frequency distributions.

PROPERTIES OF MATHEMATICAL EXPECTATION

The important properties of the expected values of a random variable are as follows:

- If c is a constant, then E(c) = c. Thus the expected value of a constant is constant itself. This point can be understood easily by considering the following interesting example: Suppose that a very difficult test was given to students by a professor, and that every student obtained 2 marks out of 20! It is obvious that the mean mark is also 2. Since the variable 'marks' was a constant, therefore its expected value was equal to itself.
- If X is a discrete random variable and if a and b are constants, then E(aX + b) = a E(X) + b.

EXAMPLE
Let X represents the number of heads that appear when three fair coins are tossed. The probability distribution of X is:

X	P(x)
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

The expected value of X is obtained as follows:

x	P(x)	xP(x)
0	1/8	0
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8
Total	1	12/8=1.5

Hence, E(X) = 1.5
Suppose that we are interested in finding the expected value of the random variable 2X+3.Then we carry out the following computations:

x	2x+3	P(x)	(2x+3)P(x)
0	3	1/8	3/8
1	5	3/8	15/8
2	7	3/8	21/8
3	9	1/8	9/8
	Total	1	48/8=6

Hence E(2X+3) = 6It should be noted that
E(2X+3) = 6= 2(1.5) + 3= 2E(X) + 3
i.e. E (aX + b) = a E(X) + b.

LECTURE NO. 24

- Chebychev's Inequality
- Concept of Continuous Probability Distribution
- Mathematical Expectation, Variance & Moments of a Continuous Probability Distribution

We begin with the discussion of the concept of the Chebychev's Inequality in the case of a discrete *probability* distribution

Chebychev's Inequality

If X is a random variable having mean μ and variance σ2 > 0, and k is any positive constant, then the probability that a value of X falls within k standard deviations of the mean is at least
That is:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2},$$

Alternatively, we may state Chebychev's theorem as follow: Given the probability distribution of the random variable X with mean μ and standard deviation σ, the probability of the observing a value of X that differs the μ by k or more standard deviations cannot exceed 1/k2. As indicated earlier, this inequality is due to the Russian mathematician P.L. Chebychev (1821-1894), and it provides a means of understanding *how the standard deviation measures variability about the mean* of a random variable. It holds for all probability distributions having finite mean and variance.
Let us apply this concept to the example of the number of petals on the flowers of a particular species that we considered earlier:

EXAMPLE

If a biologist is interested in the number of petals on a particular flower, this number may take the values 3, 4, 5, 6, 7, 8, 9, and each one of these numbers will have its own probability

The probability distribution of the random variable X is:

No. of Petals X	P(x)
$x_1 = 3$	0.05
$x_2 = 4$	0.10
$x_3 = 5$	0.20
$x_4 = 6$	0.30
$x_5 = 7$	0.25
$x_6 = 8$	0.075
$x_7 = 9$	0.025
	1

The mean of this distribution is:
$\mu = E(X) = \sum XP(X) = 5.925 \cong 5.9$
And the standard deviation of this distribution is:

$\sigma =$

$$S.D.(X) = \sqrt{36.925 - (5.925)^2}$$

$$= \sqrt{36.925 - 35.106}$$

$$= \sqrt{1.819} = 1.3$$

According to the Chebychev's inequality, the probability is at least 1 - 1/22 = 1 - 1/4 = 3/4 = 0.75 that X will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$ i.e. between 5.9 - 2(1.3) and 5.9 + 2(1.3) i.e. between 3.3 and 8.5
Let us have another look at the probability distribution:

No. of Petals X	P(x)
$x_1 = 3$	0.05
$x_2 = 4$	0.10
$x_3 = 5$	0.20
$x_4 = 6$	0.30
$x_5 = 7$	0.25
$x_6 = 8$	0.075
$x_7 = 9$	0.025
	1

According to this distribution, the probability that X lies between 3.3 and 8.5 is
0.10 + 0.20 + 0.30 + 0.25 + 0.075
= 0.925
which is *greater* than 0.75(*AS* indicated by the Chebychev's inequality).
Finally, and most importantly, we will use the concepts in Chebychev's Rule and the Empirical Rule to build the foundation for statistical inference-making. The method is illustrated in next example.

EXAMPLE

Suppose you invest a fixed sum of money in each of five business ventures. Assume you know that 70% of such ventures are successful, the outcomes of the ventures are independent of one another, and the probability distribution for the number, x, of successful ventures out of five is:

x	0	1	2	3	4	5
P(x)	.002	.029	.132	.309	.360	.168

a) Find $\mu = E(X)$.
Interpret the result.
b)Find

$$\sigma = \sqrt{E\left[(X-\mu)^2\right]}.$$

Interpret the result.
c) Graph P(x).
d) Locate μ and the interval $\mu + 2\sigma$ on the graph. Use either Chebychev's Rule or the Empirical Rule to approximate the probability that x falls in this interval. Compare this result with the actual probability.
e) Would you expect to observe fewer than two successful ventures out of five?

SOLUTION

a) Applying the formula,
$\mu = E(X) = \sum xP(x)$
= 0(.002) +1(.029) + 2(.132) + 3(.309) + 4 (.360) + 5(.168)
= 3.50

INTERPRETATION

On average, the number of successful ventures out of five will equal 3.5. (It should be remembered that this expected value has meaning only when the experiment – investing in five business ventures – is repeated a large number of times.)
b) Now we calculate the variance of X:
We know that
$\sigma 2 = E[(X - \mu)2] = \sum(x - \mu)2\ P(x)$
Hence, we will need to construct a column of $x - \mu$:

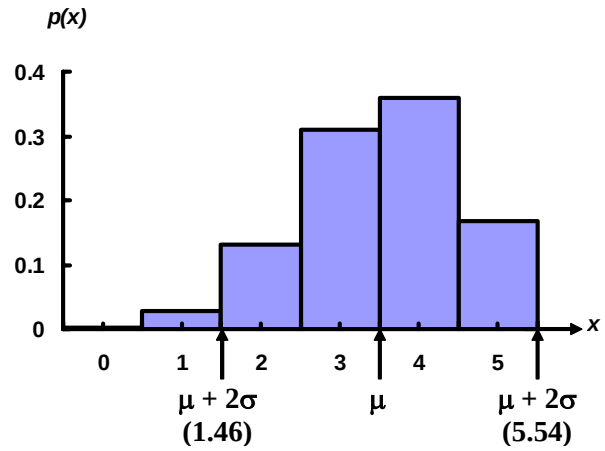
x	P(x)	x-μ	$(x-\mu)^2$	$(x-\mu)^2P(x)$
0	.002	–3.5	12.25	0.02
1	.029	–2.5	6.25	0.18
2	.132	–1.5	2.25	0.30
3	.309	–0.5	0.25	0.08
4	.360	+0.5	0.25	0.09
5	.168	+1.5	2.25	0.38
			Total	1.05

Thus, the variance is σ2 = 1.05 and the standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} = 1.02$$

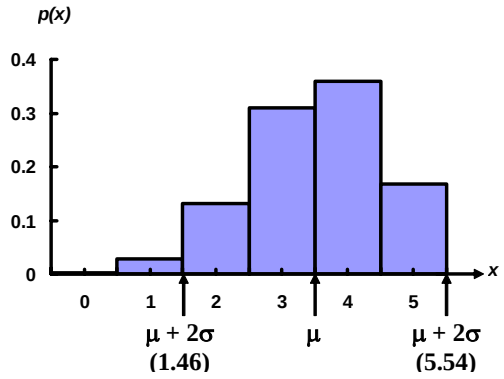
This value measures the spread of the probability distribution of X, the number of successful ventures out of five.

c) The graph of P(x) is shown in the following figure with the mean μ and the interval
μ + 2σ =3.50+2(1.02)
=3.50+2.04
= (1.46, 5.54) shown on the graph.



Note particularly that μ = 3.5 locate the *centre* of the probability distribution. Since this distribution is a theoretical relative frequency distribution that is moderately mound-shaped, we expect (from Chebychev's Rule) at least 75% and, more likely (from the Empirical Rule), approximately 95% of observed x values to fall in the interval μ + 2σ ------ that is, between 1.46 and 5.54.

It can be seen from the above figure that the actual probability that X falls in the interval μ + 2σ includes the sum of P(x) for the values
X = 2, X = 3, X = 4, and X = 5.

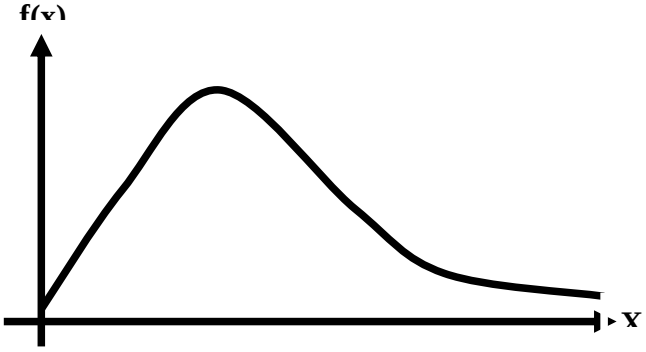


This probability is P(2) + P(3) + P(4) + P(5)
= .132 +.309 + .360 + .168
= .969.
Therefore, 96.9% of the probability distribution lies within 2 standard deviations of the mean. This percentage is *CONSISTENT* with both the Chebychev's rule and the Empirical Rule.

d) Fewer than two successful ventures out of five implies that x = 0 or x = 1. Since both these values of x lie outside the interval $\mu + 2\sigma$, we know from the Empirical Rule that such a result is unlikely (with approximate probability of only .05).The exact probability, P(x < 1), is P(0) + P(1) = .002 + .029 = .031.
Consequently, in a *single* experiment where we invest in five business ventures, we would *not* expect to observe fewer than two successful ones. The *key* question: What is the ***significance*** of the Chebychev's Inequality and the Empirical Rule?
The answer to this question is that both these rules assist us in having a certain *IDEA* regarding amount of data lying between the mean minus a certain number of standard deviations and mean plus that same number of standard deviations. Given any data-set, the moment we compute the mean and standard deviation, we *HAVE* an idea regarding the two points (i.e. mean minus two standard deviations, and mean plus two standard deviations) between which the *BULK* of our data lies. If our data-set is hump-shaped, we obtain this idea through the *Empirical Rule*, and if we don't have any reason to believe that our data-set is hump-shaped, then we obtain this idea through the Chebychev's Rule
We now begin the discussion of CONTINUOUS RANDOM VARIABLES – quantities that are measurable. As stated in the very first lecture, continuous variables result from measurement, and can therefore take any value within a certain range. For example, the height of a normal Pakistani adult male may take any value between 5 feet 4 inches and 6 feet. The temperature at a place, the amount of rainfall, time to failure for an electronic system, etc. are all *examples* of continuous random variable. Formally speaking, a continuous random variable can be defined as follows:

CONTINUOUS RANDOM VARIABLE

A random variable X is defined to be continuous if it can assume every possible value in an interval [a, b], a < b, where a and b may be $-\infty$ and $+\infty$ respectively. The function f(x) is called the *probability density function*, abbreviated to *p.d.f.*, or simply density function of the random variable X. A continuous probability distribution looks something like this:



A p.d.f. has the following properties:

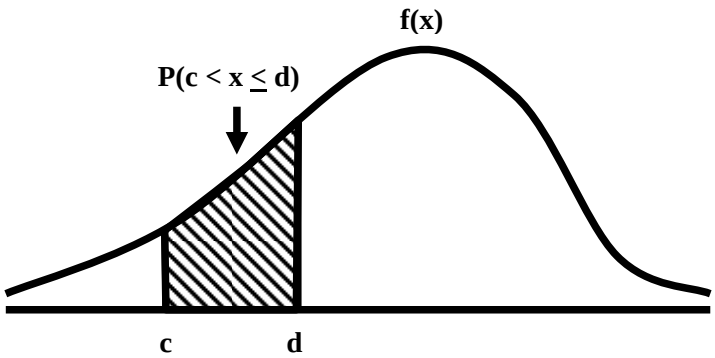
i) $\mathbf{f(x) \geq 0, \text{ for all } x}$

ii) $\int_{-\infty}^{\infty} f(x)\ dx = 1$

iii) The probability that X takes on a value in the interval [c, d], c < d is given by:

$$P(c < x < d) = \int_{c}^{d} f(x)\, dx$$

which is the area under the curve y = f(x) between X = c and X = d, as shown in the following figure:



The *TOTAL* area under the curve is 1. In other words:

- f(x) a non-negative function,
- The integration takes place over all possible values of the random variable X *between the specified limits*, and
- The probabilities are given by appropriate areas under the curve.

Since

$$P(X = k) = \int_{k}^{k} f(x)\, dx = 0,$$

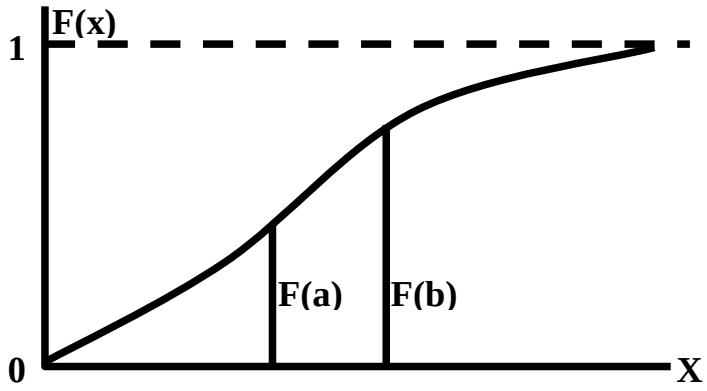
It should therefore be noted that the probability of a continuous random variable X taking any *particular* value k is always *zero*. That is why probability for a continuous random variable is measurable only over a given interval.
Further, since for a continuous random variable X, P(X = x) = 0 for every x, the following four probabilities are regarded as the same:

P(c < X < d), P(c < X < d),
P(c < X < d) and P(c < X d).

They may be different for a discrete random variable. The values (expressed as intervals) of a continuous random variable and their associated probabilities can be expressed by means of a formula.
We now discuss the *distribution function* of a continuous random variable.

CONTINUOUS RANDOM VARIABLE

A random variable X may also be defined as continuous if its *distribution function* F(x) is continuous and is differentiable everywhere except at isolated points in the given range. In contrast with the graph of the distribution function of a discrete variable, the graph of F(x) in the case of a continuous variable has no jumps or steps but is a *continuous* function for all x-values, as shown in the following figure:



Since F(x) is a non-decreasing function of x, we have

i) f(x) > 0,

ii) $F(x) = \int_{-\infty}^{x} f(x)\,dx,$ for all x.

The relationship between f(x) and F(x) is as follows: f(x) is obtained by finding the derivative of F(x), i.e.

$$\frac{d\,F(x)}{dx} = f(x)$$

EXAMPLE

a) Find the value of k so that the function f(x) defined as follows, may be a density function

f(x) = kx, 0 < x < 2
= 0, elsewhere

b) Compute P(X = 1).

c) Compute P(X > 1).

d) Compute the distribution function F(x).

e) $P(X < 1/2 \mid 1/3 < X < 2/3)$

SOLUTION

a) The function f(x) will be a density function, if

i) f(x) > 0 for every x, and

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

The first condition is satisfied when k > 0. The second condition will be satisfied, if

$$\int_{-\infty}^{\infty} f(x)\,dx = 1,$$

$$\text{i.e. if} \quad 1 = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{2} f(x)\,dx + \int_{2}^{\infty} f(x)\,dx$$

$$\text{i.e. if} \quad 1 = \int_{-\infty}^{0} 0\,dx + \int_{0}^{2} kx\,dx + \int_{2}^{\infty} 0\,dx$$

$$\text{i.e. if} \quad 1 == 0 + \left[k\frac{x^2}{2} \right]_{0}^{2} + 0 = 2k$$

We had

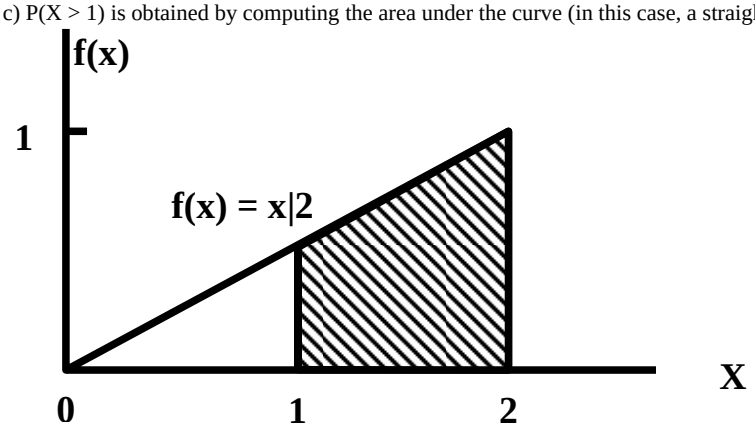
This gives k = 1/2

f(x) = kx, 0 < x < 2
= 0, elsewhere
and since we have obtained
k = 1/2, hence:

$$f(x)=\begin{cases}\frac{x}{2}, & for\ 0 \le x \le 2\\ 0, & elsewhere\end{cases}$$

b) Since f(x) is continuous probability function, therefore (X = 1) = 0.

c) P(X > 1) is obtained by computing the area under the curve (in this case, a straight line) between X=1 and X=2:



This area is obtained as follows:

$$P(X > 1) = \text{area of shaded region}$$
$$= \int_1^2 f(x)\,dx$$
$$= \int_1^2 \frac{x}{2}dx = \left[\frac{x^2}{4}\right]_1^2 = \frac{3}{4}$$

d) To compute the distribution function, we need to find:

$$\mathbf{F(x) = P(X < x)} = \int_{-\infty}^{x} f(x)\,dx$$

We do so *step by step*, as shown below:

For any x such that $-\infty < x \le 0$,

$$\mathbf{F(x)} = \int_{-\infty}^{x} 0\,dx = 0,$$

If $0 < x \le 2$, we have

$$F(x) = \int_{-\infty}^{0} 0\,dx + \int_0^x \left(\frac{x}{2}\right)dx = \left[\frac{x^2}{4}\right]_0^x = \frac{x^2}{4},$$

and, finally, for x > 2 we have

$$F(x) = \int_{-\infty}^{0} 0\,dx + \int_0^2 \frac{x}{2}dx + \int_0^2 0\,dx = 1$$

Hence

$F(x) = 0$, for $x < 0$

$= \frac{x^2}{4}$, for $0 \leq x \leq 2$

=1, for x > 2.

We will discuss the computation of the conditional probability

$$P\left(X < 1/2 \;/\; 1/3 < X < 2/3\right)$$

LECTURE NO. 25

- Mathematical Expectation, Variance & Moments of a Continuous Probability Distribution
- BIVARIATE Probability Distribution

In the last lecture, we were dealing with an example of a continuous probability distribution in which we were interested in computing a conditional probability. We now discuss this particular concept

EXAMPLE

a) Find the value of k so that the function f(x) defined as follows, may be a density function

f(x) = kx, 0 < x < 2
= 0, elsewhere

b) Compute P(X = 1).
c) Compute P(X > 1).
d) Compute the distribution function F(x).
e) $P\left(X < 1/2 \;/\; 1/3 < X < 2/3\right)$

SOLUTION

We had
f(x) = kx, 0 < x < 2
= 0, elsewhere
and we obtained k = 1/2.

Hence:

$$f(x) = \begin{cases} \frac{x}{2}, & for\ 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$$

e) Applying the definition of conditional probability, we get

$$P\left(X \le \tfrac{1}{2} \mid \tfrac{1}{3} \le X \le \tfrac{2}{3}\right) = \frac{P\left(\frac{1}{3} \le X \le \frac{1}{2}\right)}{P\left(\frac{1}{3} \le X \le \frac{2}{3}\right)} = \frac{\int_{\frac{1}{3}}^{\frac{1}{2}} \frac{x}{2}\,dx}{\int_{\frac{1}{3}}^{\frac{2}{3}} \frac{x}{2}\,dx}$$

$$= \left[\frac{x^2}{4}\right]_{\frac{1}{3}}^{\frac{1}{2}} \div \left[\frac{x^2}{4}\right]_{\frac{1}{3}}^{\frac{1}{2}}$$

The above example was of the simplest case when the graph of our continuous probability distribution is in the form of a straight line.

Let us now consider a slightly more *complicated* situation.

EXAMPLE

A continuous random variable X has the d.f. F(x) as follows:

$$F(x) = 0, \qquad \text{for } x < 0,$$

$$= \frac{2x^2}{5}, \qquad \text{for } 0 < x \le 1,$$

$$= -\frac{3}{5} + \frac{2}{5}\left(3x - \frac{x^2}{2}\right), \text{for } 1 < x \le 2,$$

$$= 1 \qquad \text{for } x > 2.$$

Find the p.d.f. and P(|X| < 1.5).

SOLUTION

By definition, we have $f(x) = \frac{d}{dx} F(x)$.

Therefore $f(x) = \frac{4x}{5}$ for $0 < x \leq 1$

$= \frac{2}{5}(3 - x)$ for $1 < x \leq 2$

$= 0$ elsewhere.

Let us now discuss the mathematical expectation of *continuous* random variables through the following example:

EXAMPLE

Find the expected value of the random variable X having the p.d.f

f(x) = 2 (1 – x), $0 < x < 1$

= 0, elsewhere

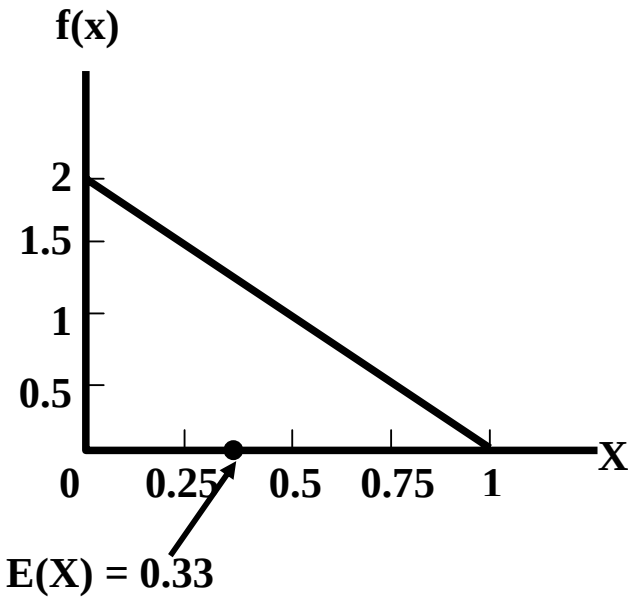
SOLUTION

Now $E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$

$= 2\int_{0}^{1} x(1-x)\, dx$

$= 2\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1 = 2\left[\frac{1}{2} - \frac{1}{3}\right] = \frac{1}{3}$

As indicated earlier, the term 'expected value' implies the *mean* value. The graph of the above probability density function and its *mean* value are presented in the following figure:



Suppose that we are interested in verifying the properties of mathematical expectation that are valid in the case of univariate probability distributions. In the last lecture, we noted that if X is a discrete random variable and if a and b are constants, then

E (aX + b) = a E(X) + b.

This property is equally valid in the case of continuous probability distributions. In this example, suppose that a = 3 and b = 5. Then, we wish to verify that

E(3X + 5) = 3 E(X) + 5.

The right-hand-side of the above equation is:

3 E(X) + 5 = 3( ) + 5 = 1 + 5 = 6

In order to compute the left-hand-side, we proceed as follows:

$$E(3X+5) = 2\int_0^1 (3x+5)(1-x)dx$$

$$= 2\int_0^1 \left(5-2x-3x^2\right)dx$$

$$= 2\left[5x-x^2-x^3\right]_0^1$$

$$= 2[5-1-1] = 2(3) = 6.$$

Since the left-hand-side is equal to the right-hand-side, therefore the property is verified.

SPECIAL CASE

We have

E(aX + b) = a E(X) + b.

If b = 0, the above property takes the following simple form:

E(aX) = a E(X).

Next, let us consider the computation of the *moments* and *moment-ratios* in the case of a continuous probability distribution:

EXAMPLE

A continuous random variable X has the p.d.f.

$$f(x) = \frac{3}{4}x(2-x),\ 0 \le x \le 2.$$

$$= 0, \qquad \text{otherwise}$$

Find the first four moments about the mean and the moment-ratios.

We first calculate the moments about origin as

$$\mu'_1 = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$= \frac{3}{4}\int_0^2 x\left(2x-x^2\right)dx = \frac{3}{4}\left[\frac{2x^3}{3}-\frac{x^4}{4}\right]_0^2$$

$$= \frac{3}{4}\left[\frac{16}{3}-\frac{16}{4}\right] = \frac{3}{4}\left[\frac{16}{12}\right] = 1;$$

$$\mu'_2 = E\left(X^2\right) = \int_{-\infty}^{\infty} x^2\, f(x)\, dx$$

$$= \frac{3}{4}\int_0^2 x^2\left(2x-x^2\right)dx = \frac{3}{4}\left[\frac{2x^4}{4}-\frac{x^5}{5}\right]_0^2$$

$$= \frac{3}{4}\left[8-\frac{32}{5}\right] = \frac{3}{4}\left[\frac{8}{5}\right] = \frac{6}{5};$$

$$\mu'_3 = E\left(X^3\right) = \int_{-\infty}^{\infty} x^3 f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 x^3\left(2x - x^2\right)dx = \frac{3}{4}\left[\frac{2x^5}{5} - \frac{x^6}{6}\right]_0^2$$

$$= \frac{3}{4}\left[\frac{64}{5} - \frac{64}{6}\right] = \frac{3}{4}\left[\frac{64}{30}\right] = \frac{8}{5};$$

$$\mu'_4 = E\left(X^4\right) = \int_{-\infty}^{\infty} x^4 f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 x^4\left(2x - x^2\right)dx = \frac{3}{4}\left[\frac{2x^6}{6} - \frac{x^7}{7}\right]_0^2$$

$$= \frac{3}{4}\left[\frac{64}{3} - \frac{128}{7}\right] = \frac{3}{4}\left[\frac{64}{21}\right] = \frac{16}{7}.$$

Next, we find the moments about the mean as follows:

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - \left(\mu'_1\right)^2 = \frac{6}{5} - (1)^2 = \frac{1}{5}$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\left(\mu'_1\right)^3$$

$$= \frac{8}{5} - 3(1)\left(\frac{6}{5}\right) + 2(1)^3 = \frac{8}{5} - \frac{18}{5} + 2 = 0;$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\left(\mu'_1\right)^2\mu'_2 - 3\left(\mu'_1\right)^4$$

$$= \frac{16}{7} - 4(1)\left(\frac{8}{5}\right) + 6(1)^2\left(\frac{6}{5}\right) - 3(1)^4$$

$$= \frac{16}{7} - \frac{32}{5} + \frac{36}{5} - 3 = \frac{3}{35}.$$

The first moment-ratio is

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0^2}{\left(\frac{1}{5}\right)^3} = 0.$$

This implies that this particular continuous probability distribution is *absolutely* symmetric
The second moment-ratio is

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\frac{3}{35}}{\left(\frac{1}{5}\right)^2} = 2.14.$$

This implies that this particular continuous probability distribution may be regarded as playkurtic, i.e. flatter than the normal distribution.
The students are encouraged to draw the graph of this distribution in order to develop a visual picture in their minds.

We begin the concept of Bivariate probability distribution by introducing the term 'Joint Distributions':

JOINT DISTRIBUTIONS

The distribution of two or more random variables which are observed simultaneously when an experiment is performed is called their *JOINT* distribution. It is customary to call the distribution of a single random variable as univariate. Likewise, a distribution involving two, three or many r.v.'s simultaneously is referred to as bivariate, trivariate or multivariate. A bivariate distribution may be discrete when the possible values of (X, U) are finite or count ably infinite. It is continuous if (X, Y) can assume all values in some non-countable set of the plane. A bivariate distribution is said mixed when one r.v. is discrete and the other is continuous.

BIVARIATE PROBABILITY FUNCTION

Let X and Y be two discrete r.v.'s defined on the same sample space S, X taking the values x1, x2, …, xm and Y taking the values y1, y2, …, yn. Then the probability that X takes on the value xi and, at the same time, Y takes on the value, denoted by f(xi, yj) or pij, is defined to be the joint probability function or simply the joint distribution of X and Y. Thus the joint probability function, also called the bivariate probability function f(x, y) is a function whose value at the point (xi, yj) is given byf(xi, yj) = P(X = xi and Y = yj),

i = 1, 2, …, m.

j = 1, 2, …, n.

The joint or bivariate probability distribution consisting of all pairs of values (xi, yj) and their associated probabilities f(xi, yj) i.e. the set of triples [xi, yj, f(xi, yj)] can either be shown in the following two-way table:

Joint Probability Distribution of X and Y

X\Y	y_1	y_2	…	y_j	…	y_n	$P(X = x_i)$
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	…	$f(x_1,y_j)$	…	$f(x_1,y_n)$	$g(x_1)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$	…	$f(x_2,y_j)$	…	$f(x_2,y_n)$	$g(x_2)$
⋮	⋮					⋮	⋮
x_i	$f(x_i,y_1)$	$f(x_i,y_2)$	…	$f(x_i,y_j)$	…	$f(x_i,y_n)$	$g(x_i)$
⋮	⋮					⋮	⋮
x_m	$f(x_m,y_1)$	$f(x_m,y_2)$	…	$f(x_m,y_j)$	…	$f(x_m,y_n)$	$g(x_m)$
$P(Y=y_j)$	$h(y_1)$	$h(y_2)$	…	$h(y_j)$	…	$h(y_n)$	1

or be expressed by mean of a formula for f(x, y). The probabilities f(x, y) can be obtained by substituting appropriate values of x and y in the table or formula. A joint probability function has the following properties:

PROPERTIES

i) f(xi, yj)>0,for all(xi, yj),i.e. for i=1,2,…,m; j = 1, 2, …, n.

ii) $$\sum_i \sum_j f(x_i, y_j) = 1$$

MARGINAL PROBABILITY FUNCTIONS

The point to be understood here is that, from the joint probability function for (X, Y), we can obtain the INDIVIDUAL probability function of X and Y. Such individual probability functions are called *MARGINAL* probability functions.

Let f(x, y) be the joint probability function of two discrete r.v.'s X and Y. Then the marginal probability function of X is defined as

$$g(x_i) = \sum_{j=1}^{n} f(x_i, y_j)$$

f(xi, y1) + f(xi, y2) + … + f(xi, yn)
as xi must occur either with y1 or y2 or … or yn
= P(X = xi);

that is, the individual probability function of X is found by adding over the rows of the two-way table. Similarly, the marginal probability function for Y is obtained by adding over the column as

$$h(y_j) = \sum_{i=1}^{m} f(x_i, y_j) = P(Y = y_j)$$

The values of the marginal probabilities are often written in the margins of the joint table as they are the row and column totals in the table. The probabilities in each marginal probability function add to 1.

CONDITIONAL PROBABILITY FUNCTION

Let X and Y be two discrete r.v.'s with joint probability function f(x, y). Then the conditional probability function for X given Y = y, denoted as f(x|y), is defined by

$$\begin{aligned} f(x_i \mid y_j) &= P(X = x_i \mid Y = y_j) \\ &= \frac{P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)} \\ &= \frac{f(x_i, y_j)}{h(y_j)}, \\ &\quad \text{for } i = 1, 2, \ldots, j = 1, 2, \ldots \end{aligned}$$

Where h(y) is the marginal probability, and h(y) > 0

It gives the probability that X takes on the value xi given that Y has taken on the value yj. The conditional probability f(xi | yj) is non-negative and (for a given fixed yj) adds to 1 on i and hence is a *probability function*. Similarly, the conditional probability function for Y given X = x is

$$\begin{aligned} f(y_j \mid x_i) &= P(Y = y_j \mid X = x_i) \\ &= \frac{P(Y = y_j \text{ and } X = x_i)}{P(X = x_i)} \\ &= \frac{f(x_i, y_j)}{g(x_i)}, \quad \text{where } g(x) > 0. \end{aligned}$$

INDEPENDENCE

Two discrete r.v.'s X and Y are said to be statistically independent, if and only if, for all possible pairs of values (xi, yj) the joint probability function f(x, y) can be expressed as the *product* of the two marginal probability functions.
That is, X and Y are independent, if

$$\begin{aligned} f(x, y) &= P(X = x_i \text{ and } Y = y_j) \\ &= P(X = x_i).\ P(Y = y_j) \\ &\qquad \text{for all i and j.} \\ &= g(x)\, h(y). \end{aligned}$$

It should be noted that the joint probability function of X and Y when they are *independent*, can be obtained by *MULTIPLYING* together their marginal probability functions.

EXAMPLE

An urn contains 3 black, 2 red and 3 green balls and 2 balls are selected at random from it. If X is the number of black balls and Y is the number of red balls selected, then find

i) the joint probability function f(x, y);
ii) $P(X + Y \leq 1)$;
iii) the marginal p.d. g(x)
and h(y);
iv) the conditional p.d. f(x | 1),
v) P(X = 0 | Y = 1); and
vi) Are x and Y independent?

i) The sample space S for this experiment contains sample points. The possible values of X are 0, 1, and 2, and those for Y are 0, 1, and 2. The values that (X, Y) can take on are (0, 0), (0, 1), (1, 0), (1, 1), (0, 2) and (2, 0). We desire to find f(x, y) for each value (x, y).
The total number of ways in which 2 balls can be drawn out of a total of 8 balls is

$$\binom{8}{2} = \frac{8x7}{2} = 28.$$

Now f(0, 0) = P(X = 0 and Y = 0), where the event (X = 0 and Y = 0) represents that neither black nor red ball is selected, implying that the 2 selected are green balls. This event therefore contains $\binom{3}{0}\binom{2}{0}\binom{3}{2} = 3$ sample points,

and

$$f(0, 0) = P(X = 0 \text{ and } Y = 0) = 3/28$$

Again f(0, 1) = P(X = 0 and Y = 1)

= P (none is black, 1 is red and 1 is green)

$$= \frac{\binom{3}{0}\binom{2}{1}\binom{3}{1}}{28} = \frac{6}{28}$$

Similarly, f(1, 1)

= P(X = 1 and Y = 1)

= P(1 is black 1 is red and none is green)

$$= \frac{\binom{3}{1}\binom{2}{1}\binom{3}{0}}{28} = \frac{6}{28}$$

Similar calculations give the probabilities of other values and the joint probability function of X and Y is given as:

Joint Probability Distribution

X \ Y	0	1	2	$P(X = x_i)$ g(x)
0	3/28	6/28	1/28	10/28
1	9/28	6/28	0	15/28
2	3/28	0	0	3/28
$P(Y = y_j)$ h(y)	15/28	12/28	1/28	1

LECTURE NO. 26

- BIVARIATE Probability Distributions (Discrete and Continuous)
- Properties of Expected Values in the case of Bivariate Probability Distributions

In the last lecture we began the discussion of the example in which we were drawing 2 balls out of an urn containing 3 black, 2 red and 3 green balls, and you will remember that, in this example, we were interested in computing quite a few quantities.

EXAMPLE

An urn contains 3 black, 2 red and 3 green balls and 2 balls are selected at random from it. If X is the number of black balls and Y is the number of red balls selected, then find

i) the joint probability function f(x, y)
ii) $P(X + Y \leq 1)$
iii) the marginal p.d. g(x) and h(y)
iv) the conditional p.d. f(x | 1)
v) P(X = 0 | Y = 1)
vi) Are x and Y independent?

As indicated in the last lecture, using the rule of combinations in conjunction with the classical definition of probability, the probability of the first cell came out to be 3/28. By similar calculations, we obtain all the remaining probabilities, and, as such, we obtain the following bivariate table:

Joint Probability Distribution

X \ Y	0	1	2	$P(X = x_i)$ g(x)
0	3/28	6/28	1/28	10/28
1	9/28	6/28	0	15/28
2	3/28	0	0	3/28
$P(Y = y_j)$ h(y)	15/28	12/28	1/28	1

This joint p.d. of the two r.v.'s (X, Y) can be represented by the formula

$$f(x,y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{28} = \begin{matrix} x=0,1,2 \\ y=0,1,2 \\ 0 \leq x+y \leq 2. \end{matrix}$$

ii) To compute $P(X + Y < 1)$, we see that $x + y < 1$ for the cells (0, 0), (0, 1) and (1, 0).

Therefore

$P(X + Y < 1)$
$= f(0, 0) + f(0, 1) + f(1, 0)$
$= 3/28 + 6/28 + 9/28$
$= 18/28 = 9/14$

iii) The marginal p.d.'s are:

x	0	1	2
g(x)	10/28	15/28	3/28

y	0	1	2
h(x)	15/28	12/28	1/28

iv) By definition, the conditional p.d. f(x | 1) is

$$f(x \mid 1) = P(X = x \mid Y = 1)$$

$$= \frac{P(X = x \text{ and } Y = 1)}{P(Y = 1)} = \frac{f(x,1)}{h(1)}$$

Now

$$h(1) = \sum_{x=0}^{2} f(x,1)$$

$$= \frac{6}{28} + \frac{6}{28} + 0$$

$$= \frac{12}{28} = \frac{3}{7}$$

Therefore

$$f(x \mid 1) = \frac{f(x,1)}{h(1)}$$

That is, $$= \frac{3}{7} f(x,1), \qquad x = 0,1,2$$

$$f(0 \mid 1) = \frac{7}{3} f(0,1) = \left(\frac{7}{3}\right)\left(\frac{6}{28}\right) = \frac{1}{2}$$

$$f(1 \mid 1) = \frac{7}{3} f(1,1) = \left(\frac{7}{3}\right)\left(\frac{6}{28}\right) = \frac{1}{2}$$

$$f(2 \mid 1) = \frac{7}{3} f(2,1) = \left(\frac{7}{3}\right)(0) = 0$$

Hence the conditional p.d. of X given that Y = 1, is

x	0	1	2
f(x|1)	1/2	1/2	0

vi) We find that f(0, 1) = 6/28,

$$g(0) = \sum_{y=0}^{2} f(0, y)$$

$$= \frac{3}{28} + \frac{6}{28} + \frac{1}{28} = \frac{10}{28}$$

$$h(1) = \sum_{x=0}^{2} f(x,1)$$

$$= \frac{6}{28} + \frac{6}{28} + 0 = \frac{12}{28}$$

v) Finally,

P(X = 0 | Y = 1)
= f(0 | 1) = 1/2

Now $\frac{6}{28} \neq \frac{10}{28} \times \frac{12}{28}$,

i.e. $f(0,1) \neq g(0)h(1)$,

therefore X and Y are ***NOT***
Satistically independent.

CONTINUOUS BIVARIATE DISTRIBUTIONS

The bivariate probability density function of continuous r.v.'s X and Y is an integral function f(x,y) satisfying the following properties:

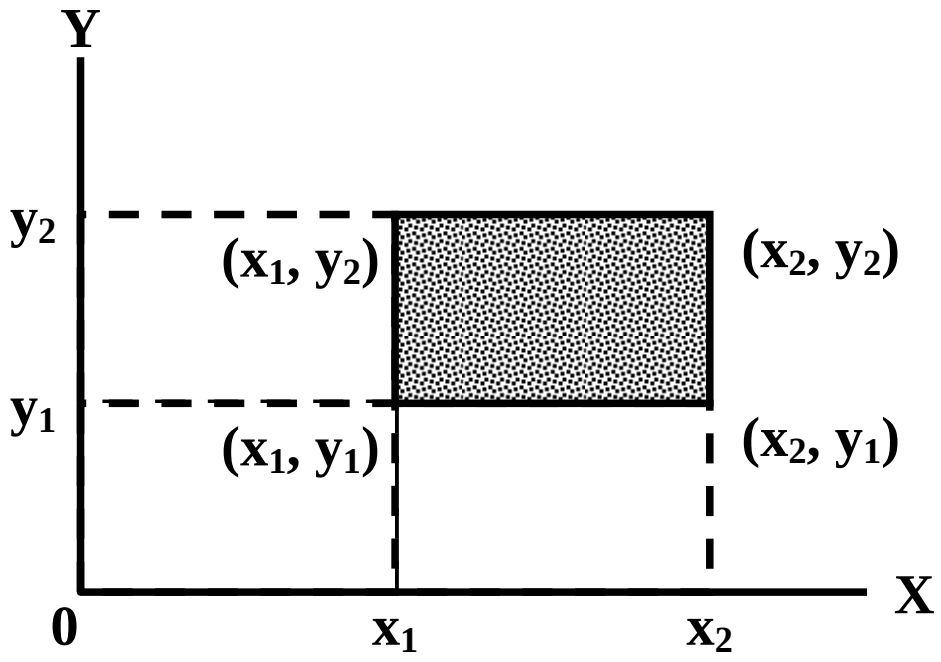
i) $f(x,y) \geq 0$ for all (x, y)

ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dx\ dy = 1$, and

$P(a \leq X \leq b, c \leq Y \leq d)$

iii) $= \int_{a}^{b} \int_{c}^{d} f(x,y)\ dy\ dx.$

Let us try to understand the graphic picture of a bivariate continuous probability distribution: The region of the XY-plane depicted by the interval (x1 < X < x2; y1 < Y < y2) is shown graphically:



Just as in the case of a continuous univariate situation, the probability function f(x) gives us a curve under which we compute areas in order to find various probabilities, in the case of a continuous bivariate situation, the probability function f(x,y) gives a SURFACE and, when we compute the probability that our random variable X lies between x1 and x2 *AND*, simultaneously, the random variable Y lies between y1 and y2, we will be computing the *VOLUME* under the surface given by our probability function f(x, y) encompassed by this region. The MARGINAL p.d.f. of the continuous r.v. X is

and that of the r.v. Y is $g(x) = \int_{-\infty}^{\infty} f(x, y)\ dy$

$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

That is, the marginal p.d.f. of any of the variables is obtained by integrating out the *other* variable from the joint p.d.f. between the limits –∞ and +∞.The CONDITIONAL p.d.f. of the continuous r.v. X given that Y takes the value y, is defined to be

$$f(x \mid y) = \frac{f(x, y)}{h(y)},$$

where f(x,y) and h(y) are respectively the joint p.d.f. of X and Y, and the marginal p.d.f. of Y, and h(y) > 0. Similarly, the conditional p.d.f. of the continuous r.v. Y given that X = x, is

$$f(y \mid x) = \frac{f(x, y)}{g(x)},$$

provided that g(x) > 0

It is worth noting that the conditional p.d.f's satisfy all the requirements for the UNIVARIATE density function.

FINALLY

Two continuous r.v.'s X and Y are said to be Statistically Independent, if and only if their joint density f(x,y) can be factorized in the form f(x,y) = g(x)h(y) for all possible values of X and Y.

EXAMPLE

Given the following joint p.d.f

$$f(x, y) = \frac{1}{8}(6 - x - y),\ 0 \le x \le 2;\ 2 \le y \le 4,$$
$$= 0, \quad \text{elsewhere}$$

a) Verify that f(x,y) is a joint density function.

b) Calculate $P\left(X \le \frac{3}{2}, Y \le \frac{5}{2}\right)$,

c) Find the marginal p.d.f. g(x) and h(y).

d) Find the conditional p.d.f. f(x | y) and f(y | x).

SOLUTION

a) The joint density f(x,y) will be a p.d.f *if*

(i) f(x,y) > 0 and

(ii)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dx\, dy = 1$$

Now f(x,y) is clearly greater than zero for all x and y in the given region, and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\ dx\ dy = \frac{1}{8}\int_{0}^{2}\int_{2}^{4} (6 - x - y)\ dy\ dx$$

$$= \frac{1}{8}\int_{0}^{2}\left[6y - xy - \frac{y^2}{2}\right]_2^4 dx$$

$$= \frac{1}{8}\int_{0}^{2}(6 - 2x)dx = \frac{1}{8}\left[6x - x^2\right]_0^2$$

$$= \frac{1}{8}[12 - 4] = 1$$

Thus f(x,y) *has* the properties of a joint p.d.f.

b) To determine the probability of a value of the r.v. (X, Y) falling in the region X < 3/2, Y < 5/2,

We find $$P\left(X \le \frac{3}{2}, Y \le \frac{5}{2}\right)$$

$$= \int_{x=0}^{\frac{3}{2}} \int_{y=2}^{\frac{5}{2}} \frac{1}{8}(6 - x - y)\, dy\, dx$$

$$= \frac{1}{8} \int_{0}^{\frac{3}{2}} \left[6y - xy - \frac{y^2}{2} \right]_2^{\frac{5}{2}} dx$$

$$= \frac{1}{8} \int_{0}^{\frac{3}{2}} \left(\frac{15}{8} - \frac{x}{2} \right) dx = \frac{1}{64} \left[15x - 2x^2 \right]_0^{\frac{3}{2}} = \frac{9}{32}$$

c) The marginal p.d.f. of X is

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\, dy, \qquad -\infty < x < \infty$$

$$= \frac{1}{8} \int_{2}^{4} (6 - x - y)\, dy, \qquad 0 \le x \le 2$$

$$= \frac{1}{8} \left[6y - xy - \frac{y^2}{2} \right]_2^4 \qquad 0 \le x \le 2$$

$$= \frac{1}{4}(3 - x), \qquad 0 \le x \le 2$$

$$= 0, \qquad x < 0 \ \underline{\text{OR}} \ x \ge 2$$

Similarly, the marginal p.d.f. of Y is

$$h(y) = \frac{1}{8} \int_{0}^{2} (6 - x - y) dx, \ 2 \le y \le 4$$

$$= \frac{1}{4}(5 - y), \qquad 2 \le y \le 4$$

$= 0,$ elsewhere.

d) The conditional p.d.f. of X given Y = y, is

$$f(x \mid y) = \frac{f(x,y)}{h(y)}, \text{ where } h(y) > 0$$

Hence

$$f(x|y) = = \frac{\left(\frac{1}{8}\right)(6 - x - y)}{\left(\frac{1}{4}\right)(5 - y)} = \frac{6 - x - y}{2(5 - y)}, \ 0 \le x \le 2$$

and the conditional p.d.f. of Y given
X = x, is

$$f(y \mid x) = \frac{f(x,y)}{g(x)}, \text{ where } g(x) > 0$$

Hence

$$f(y \mid x) = \frac{\left(\frac{1}{8}\right)(6-x-y)}{\left(\frac{1}{4}\right)(3-x)} = \frac{6-x-y}{2(3-x)}, \quad 2 \le y \le 4$$

Next, we consider *two* important properties of mathematical expectation which are valid in the case of *BIVARIATE* probability distributions:

PROPERTY NO. 1

The expected value of the sum of any two random variables is equal to the *sum* of their expected values, i.e.
E(X + Y) = E(X) + E(Y).
The result also holds for the *difference* of r.v.'s i.e.

E(X – Y) = E(X) – E(Y).

PROPERTY NO. 2

The expected value of the product of two *independent* r.v.'s is equal to the *product* of their expected values, i.e.

E(XY) = E(X) E(Y).

It should be noted that these properties are valid for *continuous* random variable's in which case the summations are replaced by *integrals.*

EXAMPLE

Let X and Y be two discrete r.v.'s with the following joint p.d

y \ x	2	4
1	0.10	0.15
3	0.20	0.30
5	0.10	0.15

Find E(X), E(Y), E(X + Y), and E(XY).

SOLUTION

To determine the expected values of X and Y, we first find the marginal p.d. g(x) and h(y) by adding over the columns and rows of the two-way table as below:

y \ x	2	4	h(y)
1	0.10	0.15	0.25
3	0.20	0.30	0.50
5	0.10	0.15	0.25
g(x)	0.40	0.60	1.00

Now $E(X) = \sum xj\, g(xj)$

$= 2 \times 0.40 + 4 \times 0.60$

$= 0.80 + 2.40 = 3.2$

$E(Y) = \sum yi\, h(yi)$

$= 1 \times 0.25 + 3 \times 0.50 + 5 \times 0.25$

$= 0.25 + 1.50 + 1.25$

$= 3.0$

Hence

$E(X) + E(Y) = 3.2 + 3.0 = 6.2$

In order to compute E(XY) directly, we apply the formula:

$$E(X+Y) = \sum_i \sum_j (x_i + y_j) f(x_i, y_j)$$

$$E(XY) = \sum_i \sum_j (x_i y_j) f(x_i, y_j)$$

LECTURE NO. 27

- Properties of Expected Values in the case of Bivariate Probability Distributions (*Detailed* discussion)
- Covariance & Correlation
- Some Well-known Discrete Probability Distributions:
  - Discrete Uniform Distribution
  - An Introduction to the Binomial Distribution

EXAMPLE

Let X and Y be two discrete r.v.'s with the following joint p.d.

x \ y	1	3	5
2	0.10	0.20	0.10
4	0.15	0.30	0.15

Find E(X), E(Y), E(X + Y), and E(XY).

SOLUTION

To determine the expected values of X and Y, we first find the marginal p.d. g(x) and h(y) by adding over the columns and rows of the two-way table as below:

x \ y	1	3	5	g(x)
2	0.10	0.20	0.10	0.40
4	0.15	0.30	0.15	0.60
h(y)	0.25	0.50	0.25	1.00

Now $E(X) = \sum x_i\, g(x_i)$
$= 2 \times 0.40 + 4 \times 0.60$
$= 0.80 + 2.40 = 3.2$

$E(Y) = \sum y_j\, h(y_j)$
$= 1 \times 0.25 + 3 \times 0.50 + 5 \times 0.25$
$= 0.25 + 1.50 + 1.25$
$= 3.0$

Hence
$E(X) + E(Y) = 3.2 + 3.0 = 6.2$

$$E(X+Y) = \sum_i \sum_j (x_i + y_j) f(x_i, y_j)$$

$= (2 + 1)(0.10) + (2 + 3)(0.20) +$
$(2 + 5)(0.10) + (4 + 1)(0.15) +$
$(4 + 3)(0.30) + (4 + 5)(0.15)$
$= 0.30 + 1.00 + 0.70 + 0.75 +$
$2.10 + 1.35 = 6.20$
$= E(X) + E(Y)$

In order to compute E(XY) directly, we apply the formula:

$$E(XY) = \sum_i \sum_j (x_i y_j) f(x_i, y_j)$$

In this example,

$$E(XY) = \sum_i \sum_j (x_i y_j) f(x_i, y_j)$$

= (2 × 1) (0.10) + (2 × 3) (0.20) + (2 × 5) (0.10) + (4 × 1) (0.15) + (4 × 3) (0.30) + (4 × 5) (0.15)
= 9.6

Now

E(X) E(Y)
= 3.2 × 3.0
= 9.6

Hence E(XY) = E(X) E(Y) implying that X and Y are independent.
This was the discrete situation; let us now consider an example of the *continuous* situation:

EXAMPLE

Let X and Y be independent r.v.'s with joint p.d.f.

$$f(x,y) = \frac{x(1+3y^2)}{4},$$
$$0 < x < 2,\ 0 < y < 1$$
$$= 0, \quad \text{elsewhere.}$$

Find E(X), E(Y), E(X + Y) and E(XY). To determine E(X) and E(Y), we first find the marginal p.d.f. g(x) and h(y) as below:

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 \frac{x(1+3y^2)}{4}\, dy$$

$$= \frac{1}{4}\left[xy + xy^3\right]_0^1 = \frac{x}{2}, \quad \text{for } 0 < x < 2.$$

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

$$= \int_0^2 \frac{x(1+3y^2)}{4}dx = \frac{1}{4}\left[\frac{x^2}{2} + 3xy^2\right]_0^2$$

$$= \frac{1}{2}(1+3y^2), \quad \text{for } 0 < y < 1.$$

Hence

$$E(X) = \int_{-\infty}^{\infty} x\, g(x)dx$$

$$= \int_0^2 x\left(\frac{x}{2}\right)dx = \frac{1}{2}\left[\frac{x^3}{3}\right]_0^2 = \frac{4}{3}, \text{and}$$

$$E(Y) = \int_{-\infty}^{\infty} y\ h(y)dy = \frac{1}{2}\int_0^1 y(1+3y^2)\, dy$$

$$= \frac{1}{2}\left[\frac{y^2}{2} + \frac{3y^4}{4}\right]_0^1 = \frac{1}{2}\left[\frac{1}{2} + \frac{3}{4}\right] = \frac{5}{8},$$

And

$$E(X+Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)\,f(x,y)dx\;dy$$

$$\int_0^2\int_0^1 (x+y)\,\frac{x(1+3y^2)}{4}dy\;dx$$

$$=\int_0^2\int_0^1 \frac{x^2+3x^2y^2}{4}dx\;dy+\int_0^2\int_0^1 \frac{xy+3xy^3}{4}dy\;dx$$

$$=\int_0^2\frac{1}{4}\left[x^2y+x^2y^3\right]_0^1 dx+\int_0^2\frac{1}{4}\left[\frac{xy^2}{2}+\frac{3xy^4}{4}\right]_0^1 dx$$

$$=\int_0^2\frac{1}{4}\left(2x^2\right)dx+\int_0^2\frac{1}{4}\left(\frac{x}{2}+\frac{3x}{4}\right)dx$$

$$=\frac{1}{2}\quad\left[\frac{x^3}{3}\right]_0^2=\frac{1}{4}\left[\frac{x^2}{4}+\frac{3x^2}{8}\right]_0^2$$

$$=\frac{4}{3}+\frac{5}{8}=\frac{47}{24},\text{ and}$$

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\;f(x,y)dx\;dy$$

$$=\int_0^2\int_0^1 (xy)\,\frac{x(1+3y^2)}{4}dy\;dx=\int_0^2\int_0^1 \frac{x^2y+3x^2y^3}{4}dy\;dx$$

$$=\int_0^2\frac{1}{4}\left[\frac{x^2y^2}{2}+\frac{3x^2y^4}{4}\right]_0^1 dx=\int_0^2\frac{1}{4}\left(\frac{5x^2}{4}\right)dx=\frac{1}{4}\quad\left[\frac{5x^3}{12}\right]_0^2=\frac{5}{6}$$

It should be noted that

i) E(X) + E(Y) = 4/3 + 5/8
= 47/24 = E(X + Y), and

ii) E(X) E(Y) = (4/3) (5/8)
= 5/6 = E(XY).

Hence, the two properties of mathematical expectation valid in the case of bivariate probability distributions are verified.

COVARIANCE OF TWO RANDOM VARIABLES

The covariance of two r.v.'s X and Y is a numerical measure of the extent to which their values tend to increase or decrease *together*. It is denoted by σXY or Cov (X, Y), and is defined as the expected value of the product

[X – E(X)] [Y – E(Y)]. That is
Cov (X, Y) = E {[X – E(X)] [Y – E(Y)]}
And the short cut formula is:

Cov (X, Y) = E(XY) – E(X) E(Y).

If X and Y are independent, then
E(XY) = E(X) E(Y), and
Cov (X, Y) = E(XY) – E(X) E(Y)
= 0
It is very important to note that covariance is zero when the r.v.'s X and Y are independent but its converse is not generally true. The covariance of a r.v. with itself is obviously its variance.

CORRELATION CO-EFFICIENT OF TWO RANDOM VARIABLES

Let X and Y be two r.v.'s with non-zero variances σ2X and σ2Y. Then the correlation coefficient which is a measure of linear relationship between X and Y, denoted by ρXY (the Greek letter rho) or Corr(X, Y), is defined as

$$\rho_{XY} = \frac{E[X - E(X)][Y - E(X)]}{\sigma_X \sigma_Y}$$
$$= \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}$$

If X and Y are independent r.v.'s, then ρXY will be zero but zero correlation does not necessarily imply independence.

EXAMPLE

From the following joint p.d. of X and Y, find Var(X), Var(Y), Cov(X,Y) and ρ.

x \ y	0	1	2	3	g(x)
0	0.05	0.05	0.10	0	0.20
1	0.05	0.10	0.25	0.10	0.50
2	0	0.15	0.10	0.05	0.30
h(y)	0.10	0.30	0.45	0.15	1.00

Now
E(X) = Σxig(xi)
= 0 × 0.20 + 1 × 0.50 +2 × 0.30
= 0 + 0.50 + 0.60 = 1.10
E(Y) = Σyjh(yj)
= 0 × 0.10 + 1 × 0.30 + 2 × 0.45 + 3 × 0.15
= 0 + 0.30 + 0.90 + 0.45 = 1.65
E(X2) = Σxi2g(xi)
= 0 × 0.20 + 1 × 0.50 + 4 × 0.30
= 1.70
E(Y2) = Σyj2h(yj)
= 0 × 0.10 + 1 × 0.30 + 4 × 0.45 + 9 × 0.15
= 3.45

Thus
Var(X) = E(X2) – [E(X)]2
= 1.70 – (1.10)2 = 0.49,
and

Var(Y) = E(Y2) – [E(Y)]2
= 3.45 – (1.65)2 = 0.7275

Again:
E(XY) $= \sum_i \sum_j (x_i y_j) f(x_i, y_j)$
= 1 × 0.10 + 2 × 0.15 + 2 × 0.25 + 4 × 0.10 + 3 × 0.10 + 6 × 0.05

= 0.10 + 0.30 + 0.50 + 0.40 + 0.30 + 0.30
= 1.90

$\therefore$ $\mathrm{Cov}(X,Y) = E(XY) - E(X)\,E(Y)$
$= 1.90 - 1.10 \times 1.65 = 0.085$, and

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

$$= \frac{0.085}{\sqrt{(0.49)(0.7275)}} = \frac{0.085}{0.595}$$

$$= 0.14$$

Hence, we can say that there is a weak positive linear correlation between the random variables X and Y.

EXAMPLE

If f(x, y)
= x2 + xy/3 , 0 < x < 1, 0 < y < 2
= 0, elsewhere,

Find
Var(X), Var(Y) and Corr(X,Y)

SOLUTION

The marginal p.d.f.'s are

$$g(x) = \int_0^2 \left(x^2 + \frac{xy}{3}\right) dy = 2x^2 + \frac{3}{2}x, \qquad 0 \le x \le 1$$

and

$$h(y) = \int_0^1 \left(x^2 + \frac{xy}{3}\right) dx = \frac{1}{3} + \frac{y}{6}, \qquad 0 \le y \le 2$$

Now

$$E(X) = \int_{-\infty}^{\infty} x\,g(x)\,dx$$

$$= \int_0^1 x\left(2x^2 + \frac{2x}{3}\right) dx = \frac{13}{18},$$

$$E(Y) = \int_{-\infty}^{\infty} y\,h(y)\,dy$$

Thus

$$= \int_0^2 y\left(\frac{1}{3} + \frac{y}{6}\right) dy = \frac{10}{9}.$$

$$\mathrm{Var}(X) = E[X - E(X)]^2$$

$$= \int_{-\infty}^{\infty} (x + \mu_x)^2\, g(x)\,dx$$

$$= \int_0^1 \left(x - \frac{13}{18}\right)^2 \left(2x^2 + \frac{2x}{3}\right) dx = \frac{73}{1620}$$

$$\text{Var}(Y) = E[Y - E(Y)]^2$$

$$= \int_{-\infty}^{\infty} (y - \mu_y)^2 h(y)\, dy$$

$$= \int_0^2 \left(y - \frac{10}{9}\right)^2 \left(\frac{1}{3} + \frac{y}{6}\right) dy = \frac{26}{81}, \text{ and}$$

Cov(X, Y)

$$= E\{[X - E(X)]\,[Y - E(Y)]\}$$

$$= \int_0^1 \int_0^2 \left(x - \frac{13}{18}\right)\left(y - \frac{10}{9}\right)\left(x^2 + \frac{xy}{3}\right) dy\, dx$$

$$= \int_0^1 \left(-\frac{2}{9}x^3 + \frac{25}{81}x^2 - \frac{26}{243}x\right) dx = \frac{-1}{162}.$$

Hence

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\ \text{Var}(Y)}}$$

$$= \frac{-1/162}{\sqrt{(73/1620)(26/81)}}$$

$$= -0.05$$

Hence we can say that there is a *VERY* weak negative linear correlation between X and Y. In other words, X and Y are almost uncorrelated. This brings us to the end of the discussion of the BASIC concepts of discrete and continuous *Univariate* and *bivariate* probable. We now begin the discussion of some probability distributions that are WELL-KNOWN, and are encountered in *real-life* situations.

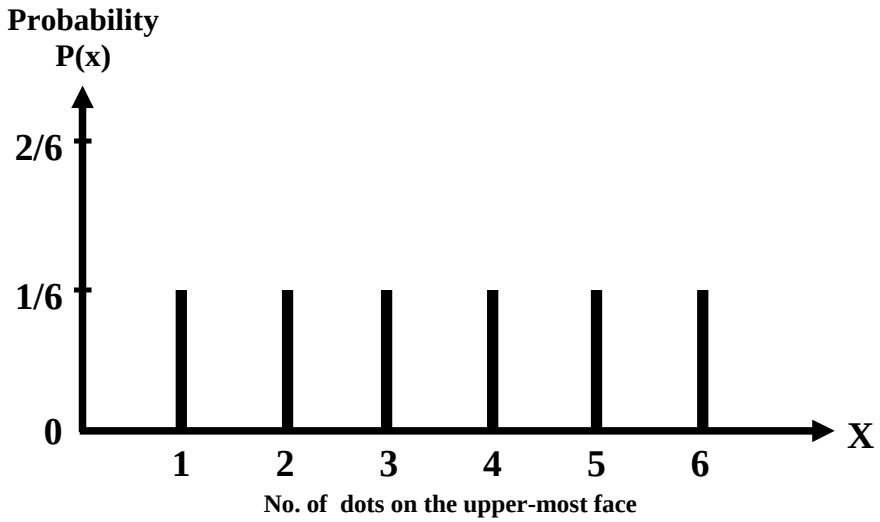
DISCRETE UNIFORM DISTRIBUTION

EXAMPLE

Suppose that we toss a fair die and let X denote the number of dots on the upper-most face. Since the die is *fair*, hence each of the X-values from 1 to 6 is equally likely to occur, and hence the probability distribution of the random variable X is as follows:

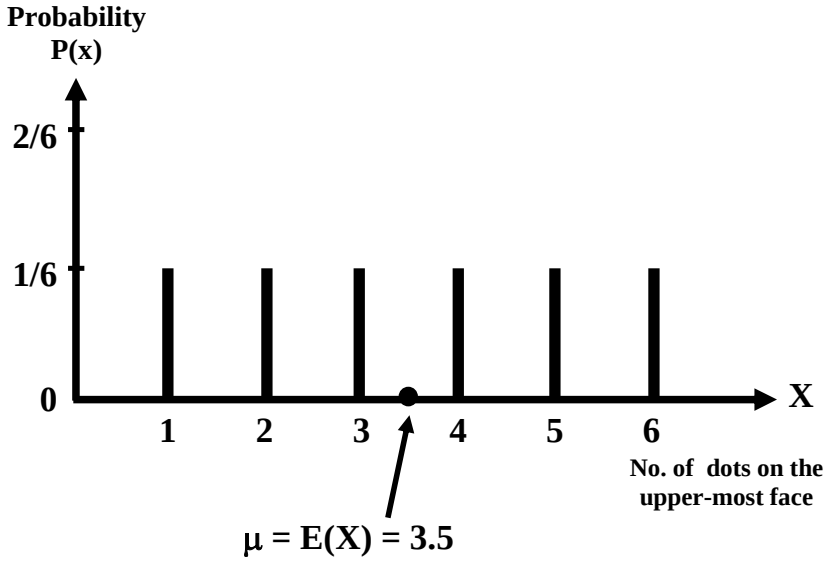
X	P(x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

If we draw the line chart of this distribution, we obtain Line Chart Representation of the Discrete Uniform Probability Distribution



As all the vertical line segments are of equal height, hence this distribution is called a uniform distribution.
As this distribution is absolutely symmetrical, therefore the mean lies at the *exact centre* of the distribution i.e. the mean is equal to 3.5.

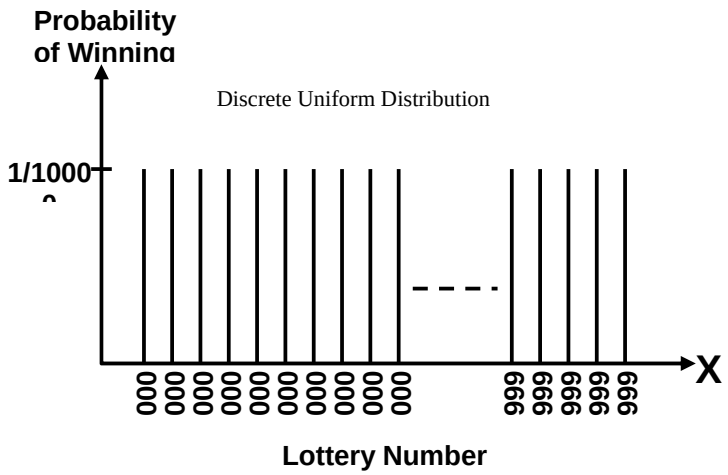
LINE CHART REPRESENTATION OF THE DISCRETE UNIFORM PROBABILITY DISTRIBUTION



What about the spread of this distribution? You are encouraged to compute the standard deviation as well as the coefficient of variation of this distribution on their own. Let us consider another interesting example.

EXAMPLE

The lottery conducted in various countries for purposes of money-making provides a good example of the discrete uniform distribution. Suppose that, in a particular lottery, as many as ten thousand lottery tickets are issued, and the numbering is 0000 to 9999. Since each of these numbers is *equally likely* to occur, hence we have the following situation:



INTERPRETATION

It reflects the fact that winning lottery numbers are selected by a random procedure which makes all numbers equally likely to be selected. The point to be kept in mind is that, whenever we have a situation where the various outcomes are equally likely, and of a form such that we have a random variable X with values 0, 1, 2, … or , as in the above example, 0000, 0001 …, 9999, we will be dealing with the discrete uniform distribution.

BINOMIAL DISTRIBUTION

The binomial distribution is a very important discrete probability distribution. It was discovered by James Bernoulli about the year 1700.We illustrate this distribution with the help of the following example:

EXAMPLE

Suppose that we toss a fair coin 5 times, and we are interested in determining the probability distribution of X, where X represents the number of heads that we obtain.
We note that in tossing a fair coin 5 times:

- every toss results in either a head or a tail,
- the probability of heads (denoted by p) is equal to ½ every time (in other words, the probability of heads remains *constant*),
- every throw is *independent* of every other throw, and
- the total number of tosses i.e. 5 is *fixed* in *advance*.

The above four points represents the *four basic* and vitally important *PROPERTIES* of a binomial experiment

PROPERTIES OF A BINOMIAL EXPERIMENT

- Every trial results in a success or a failure.
- The successive trials are independent.
- The probability of success, p, remains constant from trial to trial.
- The number of trials, n, is fixed in advanced.

LECTURE NO. 28

- Binomial Distribution
- Fitting a Binomial Distribution to Real Data
- An Introduction to the Hyper geometric Distribution

The binomial distribution is a very important discrete probability distribution. We illustrate this distribution with the help of the following example:

EXAMPLE

Suppose that we toss a fair coin 5 times, and we are interested in determining the probability distribution of X, where X represents the number of heads that we obtain. We note that in tossing a fair coin 5 times:

- Every toss results in either a head or a tail,
- The probability of heads (denoted by p) is equal to ½ every time (in other words, the probability of heads remains *constant*),
- Every throw is *independent* of every other throw, and
- The total number of tosses i.e. 5 is *fixed* in *advance*.

The above four points represents the *four basic* and vitally important *PROPERTIES* of binomial experiment. Now, in 5 tosses of the coin, there can be 0, 1, 2, 3, 4 or 5 heads, and the no. of heads is thus a random variable which can take one of these six values. In order to compute the probabilities of these X-values, the formula is:

Binomial Distribution

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Where
n = the total no. of trials
p = probability of success in each trial
q = probability of failure in
each trial (i.e. q = 1 - p)
x = no. of successes in n trials.
x = 0, 1, 2, … n

The binomial distribution has two parameters, n and p. In this example, n = 5 since the coin was thrown 5 times, p = ½ since it is a fair coin, q = 1 – p = 1 – ½ = ½ Hence

Putting x = 0

$$P(X = x) = \binom{5}{x}\left(\tfrac{1}{2}\right)^x \left(\tfrac{1}{2}\right)^{5-x}$$

$$P(X = 0) = \binom{5}{0}\left(\tfrac{1}{2}\right)^0 \left(\tfrac{1}{2}\right)^{5-0}$$

$$= \frac{5!}{0!\,5!}(1)\left(\tfrac{1}{2}\right)^5$$

$$= 1(1)\left(\tfrac{1}{2}\right)^5 = \frac{1}{32}$$

Putting x = 1

$$P(X = 1) = \binom{5}{1}\left(\tfrac{1}{2}\right)^1 \left(\tfrac{1}{2}\right)^{5-1}$$

$$= \frac{5!}{1!\,4!}\left(\tfrac{1}{2}\right)^1 \left(\tfrac{1}{2}\right)^4$$

$$= \frac{(5)}{1}\left(\tfrac{1}{2}\right)$$

$$= 5\left(\tfrac{1}{2}\right)^5 = 5\left(\frac{1}{32}\right) = \frac{5}{32}$$

Similarly, we have:

$$P(X=2)=\binom{5}{2}\left(\tfrac{1}{2}\right)^2\left(\tfrac{1}{2}\right)^{5-2}=\frac{10}{32}$$

$$P(X=3)=\binom{5}{3}\left(\tfrac{1}{2}\right)^3\left(\tfrac{1}{2}\right)^{5-3}=\frac{10}{32}$$

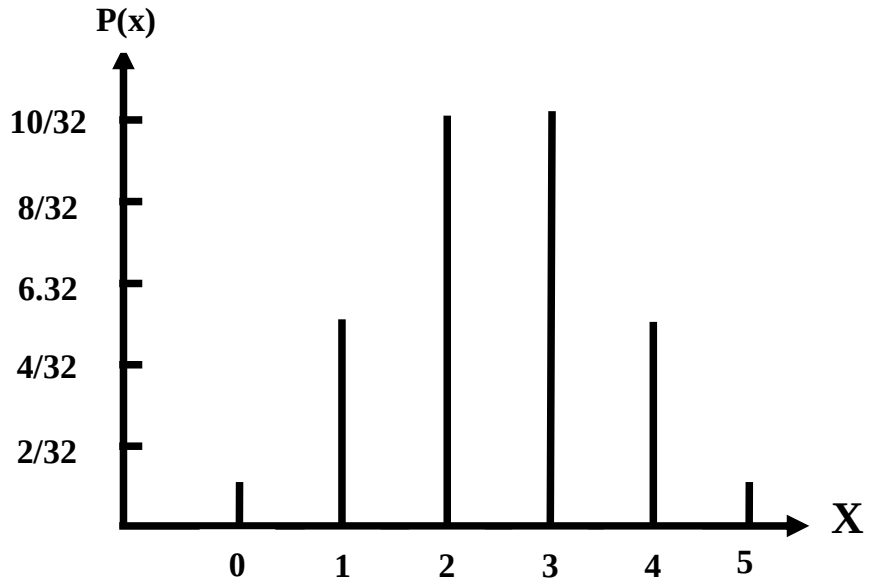
$$P(X=4)=\binom{5}{4}\left(\tfrac{1}{2}\right)^4\left(\tfrac{1}{2}\right)^{5-4}=\frac{5}{32}$$

$$P(X=5)=\binom{5}{5}\left(\tfrac{1}{2}\right)^5\left(\tfrac{1}{2}\right)^{5-5}=\frac{1}{32}$$

Hence, the binomial distribution for this particular example is as follows. Binomial Distribution in the case of tossing a fair coin five times:

Number of Heads	Probability
X	P(x)
0	1/32
1	5/32
2	10/32
3	10/32
4	5/32
5	1/32
Total	32/32 = 1

Graphical Representation of the above binomial distribution:



The next question is: What about the mean and the standard deviation of this distribution? We can calculate them just as before, using the formulas

Mean of X = E(X) = ∑XP(X)
Var(X) = ∑X2 P(X) – [∑XP(X)]2

but it has been mathematically proved that for a binomial distribution given by

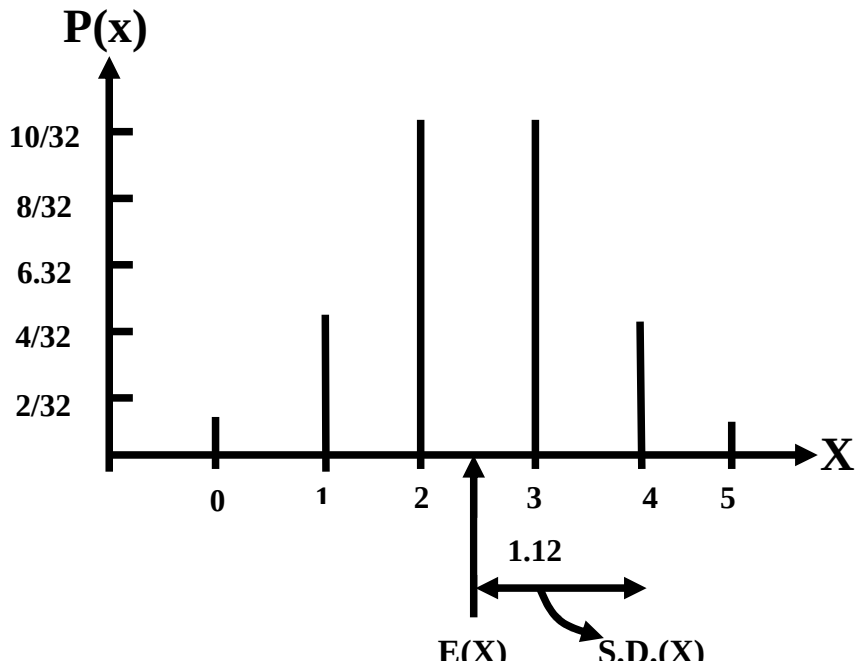
$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

For a binomial distribution
E(X) = np
and Var(X) = npq
so that $S.D.(X) = \sqrt{npq}$

For the above example, n = 5, p = ½ and q = ½
Hence
Mean = E(X) = np = 5(½) = 2.5

and S.D.(X) $= \sqrt{\text{npq}} = \sqrt{5\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = \sqrt{\left(\frac{5}{4}\right)} = \underline{1.12}$

We would have got exactly the same answers if we had applied the LENGTHIER procedure.
E(X) = ∑XP(X) and Var X = ∑X2 P(X)-[∑XP(X)]2
Graphical Representation of the Mean and Standard Deviation of the Binomial Distribution (n=5, p=1/2)



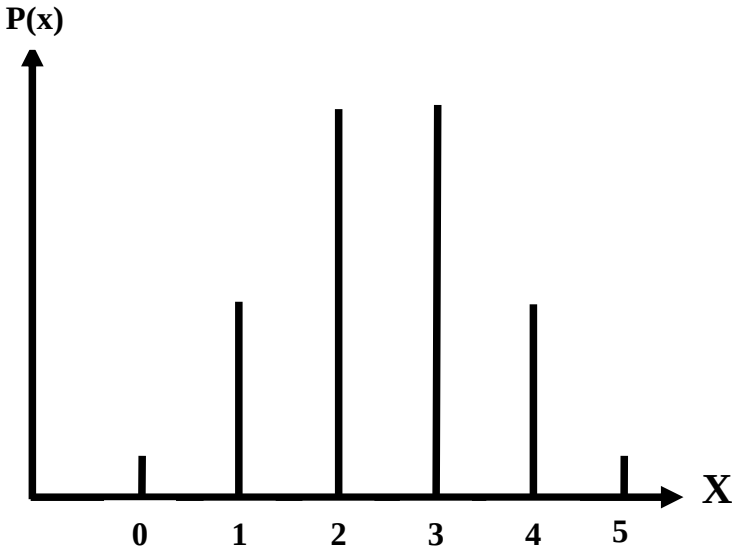
WHAT DOES THIS MEAN?

What this mean is that if 5 *fair* coins are tossed an *INFINITE* no. of times, sometimes we will get no head out of 5, sometimes/head… sometimes all 5 heads. But on the *AVERAGE* we should expect to get 2.5 heads in 5 tosses of the coin, or, a total of 25 heads in 50 tosses of the coin And 1.12 gives a measure of the possible *variability* in the various numbers of heads that can be obtained in 5 tosses. (As you know, in this problem, the number of heads can range from 0 to 5 had the coin been tossed 10 times, the no. of heads possible would vary from 0 to 10 and the standard deviation would probably have been different).

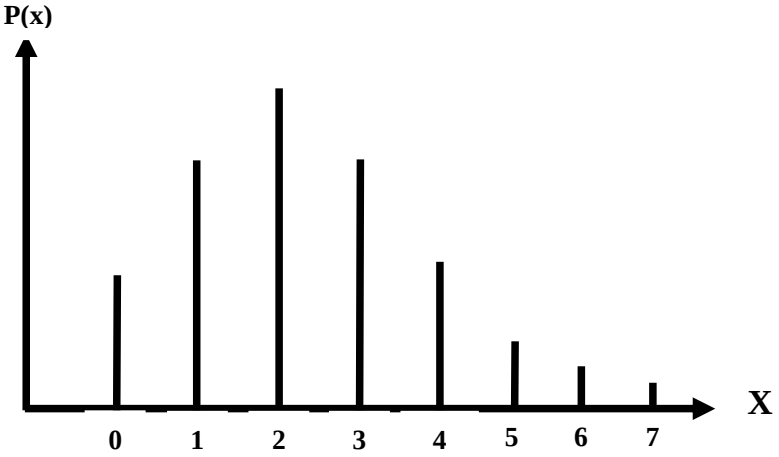
Coefficient of Variation:

$$\text{C.V.} = \frac{\sigma}{\mu} \times 100 = \frac{1.12}{2.5} \times 100 = 44.8\%$$

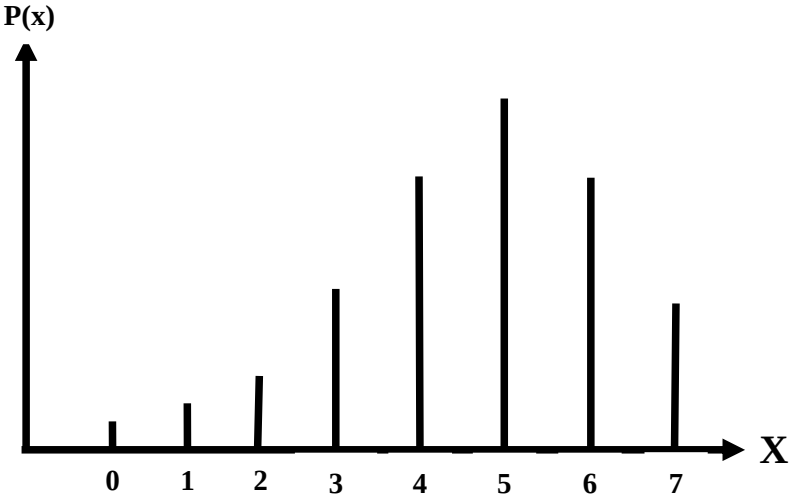
Note that the binomial distribution is not always symmetrical as in the above example. It will be symmetrical *only* when p = q = ½ (as in the above example).



It is skewed to the right if p < q:



It is skewed to the left if p > q:



But the degree of Skewness (or asymmetry) *decreases* as n increases. Next, we consider the *Fitting* of a Binomial Distribution to *Real* Data. We illustrate this concept with the help of the following example:

EXAMPLE

The following data has been obtained by tossing a *LOADED* die 5 times, and noting the number of times that we obtained a *six*. Fit a binomial distribution to this data.

No. of Sixes	0	1	2	3	4	5	Total
Frequency	12	56	74	39	18	1	200

SOLUTION

To fit a binomial distribution, we need to find n and p.

Here n = 5, the largest x-value.

To find p, we use the relationship $\bar{x} = np$.

The rationale of this step is that, as indicated in the last lecture, the mean of a binomial *probability* distribution is equal to np, i.e.

$\mu = np$

But, here, we are not dealing with a *probability* distribution i.e. the entire *population* of all possible sets of throws of a loaded die --- we only have a *sample* of throws at our disposal.

As such, μ is not available to us, and all we can do is to replace it by its estimate $\bar{X}$.

Hence, our equation becomes $\bar{X} = np$.

Now, we have:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{0+56+148+117+72+5}{200}$$

$$= \frac{398}{200} = 1.99$$

Using the relationship $\bar{x} = np$, we get $5p = 1.99$ or $p = 0.398$.This value of p seems to indicate *clearly* that the die is not fair at all! (Had it been a fair die, the probability of getting a six would have been 1/6 i.e. 0.167; a value of p = 0.398 is *very* different from 0.167.) Letting the random variable X represent the number of sixes, the above calculations yield the fitted binomial distribution as

$$b(x;5,\ 0.398) = \binom{5}{x} (0.398)^x (0.602)^{5-x}$$

Hence the *probabilities* and *expected frequencies* are calculated as below:

No. of Sixes (x)	Probability f(x)		Expected frequency
0	$\binom{5}{0} q^5 = (0.602)^5$	= 0.07907	15.8
1	$\binom{5}{1} q^5 p = 5.(0.602)^4 (0.398)$	= 0.26136	52.5
2	$\binom{5}{2} q^3 p^2 = 10.(0.602)^3 (0.398)^2$	= 0.34559	69.1
3	$\binom{5}{3} q^2 p^3 = 10.(0.602)(0.398)^3$	= 0.22847	45.7
4	$\binom{5}{4} qp^4 = (0.602)(0.398)^4$	= 0.07553	15.1
5	$\binom{5}{5} p^5 = (0.398)^5$	= 0.00998	2.0
Total		= 1.00000	200.0

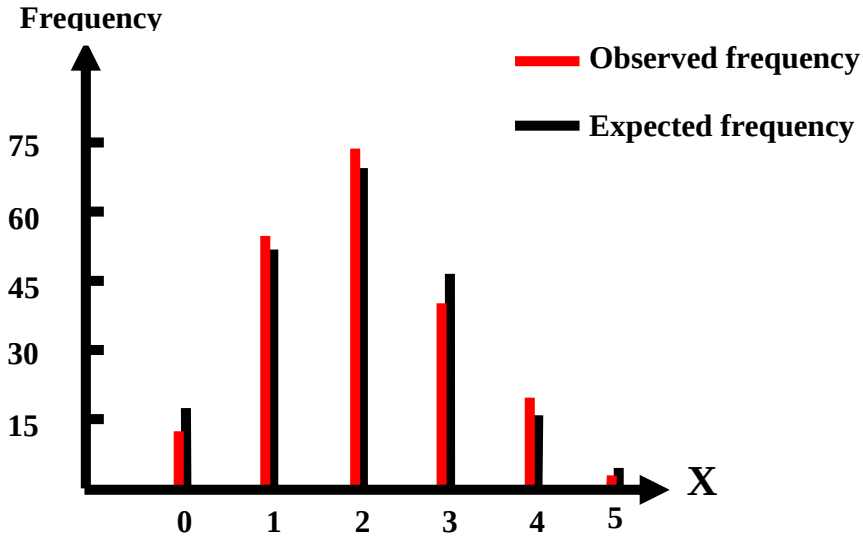
In the above table, the expected frequencies are obtained by multiplying each of the probabilities by 200.

In the entire above procedure, we are assuming that the given frequency distribution has the characteristics of the fitted theoretical binomial distribution, comparing the observed frequencies with the expected frequencies, we obtain:

No. of Sixes x	Observed Frequency f_0	Expected Frequency f_e
0	12	15.8
1	56	52.5
2	74	69.1
3	39	45.7
4	18	15.1
5	1	2.0
Total	200	200.0

The graphical representation of the observed frequencies as well as the expected frequencies is as follows:

Graphical Representation of the Observed and Expected Frequencies:



The above graph quite clearly indicates that there is not much discrepancy between the observed and the expected frequencies. Hence, we can say that it is a reasonably good fit.

There is a procedure known as the Chi-Square Test of Goodness of Fit which enables us to determine in a formal, mathematical manner whether or not the theoretical distribution fits the observed distribution reasonably well. This test comes under the realm of Inferential Statistics --- that area which we will deal with during the last 15 lectures of this course. Let us consider a *real-life* application of the binomial distribution:

AN EXAMPLE FROM INDUSTRY

Suppose that the past record indicates that the proportion of defective articles produced by this factory is 7%.And suppose that a law *NEWLY* instituted in this particular country states that there should not be more than 5% defective.

Suppose that the factory-owner makes the statement that his machinery has been *overhauled* so that the number of defectives has *DECREASED*.

In order to examine this claim, the relevant government department decides to send an inspector to examine a sample of 20 items.

What is the probability that the inspector will find 2 or more defective items in his sample (so that a fine will be imposed on the factory)?

SOLUTION

The first step is to identify the NATURE of the situation, If we study this problem closely, we realize that we are dealing with a binomial experiment because of the fact that all four properties of a binomial experiment are being fulfilled:

PROPERTIES OF A BINOMIAL EXPERIMENT

- Every item selected will either be defective (i.e. *success*) or not defective (i.e. *failure*)
- Every item drawn is independent of every other item
- The probability of obtaining a defective item i.e. 7% is the same (constant) for all items. (This probability figure is according to relative frequency definition of probability.
- The number of items drawn is fixed in advance i.e. 20 hence; we are in a position to apply the binomial formula

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

$$P(X = x) = \binom{20}{x} 0.07^x 0.93^{20-x}$$

Substituting n = 20 and p = 0.07, we obtain:

Now

P(X > 2) = 1 - P(X < 2)
= 1- [P(X = 0) + P(X =1)]

$$= 1 - \binom{20}{0} 0.07^0 0.93^{20-0} - \binom{20}{1} 0.07^1 0.93^{20-1}]$$

$$= 1 - 1 \times 1 \times 0.93^{20} - 20 \times 0.07 \times 0.93^{19}$$

$$= 1 - 0.234 - 0.353$$

$$= 0.413$$

$$= 41.3\%$$

Hence the probability is SUBSTANTIAL i.e. more than 40% that the inspector will find two or more defective articles among the 20 that he will inspect. In other words, there is CONSIDERABLE chance that the factory will be fined.
The point to be realized is that, generally speaking, whenever we are dealing with a 'success / failure' situation, we are dealing with what can be a binomial experiment. (For EXAMPLE, if we are interested in determining any of the following proportions, we are dealing with a BINOMIAL situation:

- Proportion of smokers in a city smoker → success, non-smokers → failure.
- Proportion of literates in a community → literacy rate, literate → success, illiterate → failure.
- Proportion of males in a city → *sex ratio*).

HYPERGEOMETRIC PROBABILITY DISTRIBUTION

There are many experiments in which the condition of independence is violated and the probability of success does not remain constant for all trials. Such experiments are called hyper geometric experiments.

In other words, a hyper geometric experiment has the following properties:

PROPERTIES OF HYPERGEOMETRIC EXPERIMENT

- The outcomes of each trial may be classified into one of two categories, success and failure.
- The probability of success changes on each trial.
- The successive trials are not independent.
- The experiment is repeated a fixed number of times.

The number of success, X in a hyper geometric experiment is called a hyper geometric random variable and its probability distribution is called the hyper geometric distribution. When the hyper geometric random variable X assumes a value x, the hyper geometric probability distribution is given by the formula

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

Where
N = number of units in the population,
n = number of units in the sample, and
k = number of successes in the population.
The hyper geometric probability distribution has three parameters N, n and k.
The hyper geometric probability distribution is appropriate when

- a random sample of size n is drawn *WITHOUT REPLACEMENT* from a *finite* population of N units;
- k of the units are of one kind (classified as success) and the remaining N – k of another kind (classified as failure).

LECTURE NO. 29

- Hyper geometric Distribution
  (in some detail)
- Poisson Distribution
- Limiting Approximation to the Binomial
- Poisson Process
- Continuous Uniform Distribution

In the last lecture, we began the discussion of the HYPERGEOMETRIC PROBABILITY DISTRIBUTION. We now consider this distribution in some detail. As indicated in the last lecture, there are many experiments in which the condition of independence is violated and the probability of success does not remain constant for all trials. Such experiments are called hyper geometric experiments. In other words, a hyper geometric experiment has the following properties:

PROPERTIES OF HYPERGEOMETRIC EXPERIMENT

- The outcomes of each trial may be classified into one of two categories, success and failure.
- The probability of success changes on each trial.
- The successive trials are not independent.
- The experiment is repeated a fixed number of times.

The number of success, X in a hyper geometric experiment is called a hyper geometric random variable and its probability distribution is called the hyper geometric distribution. When the hyper geometric random variable X assumes a value x, the hyper geometric probability distribution is given by the formula

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

where
N = number of units in the population,
n = number of units in the sample,
and
k = number of successes in the population.
The hyper geometric probability distribution
has three parameters N, n and k.

- The hyper geometric probability distribution is appropriate when
- a random sample of size n is drawn *WITHOUT REPLACEMENT* from a *finite* population of N units;
- k of the units are of one kind (classified as success) and the remaining N – k of another kind (classified as failure).

EXAMPLE

The names of 5 men and 5 women are written on slips of paper and placed in a hat. Four names are drawn. What is the probability that 2 are men and 2 are women? Let us regard 'men' as success. Then X will denote the number of men. We have N = 5 + 5 = 10 names to be drawn from; Also, n = 4, (since we are drawing a sample of size 4 out of a 'population' of size 10) In addition, k = 5 (since there are 5 men in the population of 10). In this problem, the possible values of X are 0, 1, 2, 3, 4, i.e. n): The hyper geometric distribution is given by

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

Since N = 10, k = 5 and n = 4, hence, in this problem, the hyper geometric distribution is given by

$$P(X = x) = \frac{\binom{5}{x}\binom{5}{4-x}}{\binom{10}{4}}$$

and the required probability,
i.e P(X = 2) is

$$P(X=2)=\frac{\binom{5}{2}\binom{5}{4-2}}{\binom{10}{4}}$$
$$=\frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}}$$
$$=\frac{10\times 10}{210}$$
$$=\frac{10}{21}$$

In other words, the probability is a little less than 50% that two of the four names drawn will be those of MEN. In the above example, just as we have computed the probability of X = 2, we could also have computed the probabilities of X = 0, X = 1, X = 3 and X = 4 (i.e. the probabilities of having zero, one, three *OR* four men among the four names drawn).The students are encouraged to compute these probabilities on their own, to check that the sum of these probabilities is 1, and to draw the line chart of this distribution.
Additionally, the students are encouraged to think about the *centre*, *spread* and *shape* of the distribution. Next, we consider some important *PROPERTIES* of the Hyper
geometric Distribution:

PROPERTIES OF THE HYPERGEOMETRIC DISTRIBUTION

- The mean and the hyper geometric probability distribution are
$$\mu = n\frac{k}{N} \text{ and } \sigma^2 = n\frac{k}{N}\frac{N-k}{N}\frac{N-n}{N-1},$$

- If N becomes *indefinitely* large, the hyper geometric probability distribution tends to the *BINOMIAL* probability distribution.

The above property will be best understood with reference to the following important points:

- There are two ways of drawing a sample from a population, sampling with replacement, and sampling without replacement.
- Also, a sample can be drawn from either a finite population or an infinite population.

This leads to the following bivariate table: With reference to sampling, the various possible situations are:

Population / Sampling	Finite	Infinite
With replacement		
Without replacement		

The point to be understood is that, whenever we are sampling with replacement, the population remains undisturbed (because any element that is drawn at any one draw, is re-placed into the population before the next draw).Hence, we can say that the various trials (i.e. draws) are independent, and hence we can use the binomial formula. On the other hand, when we are sampling without replacement from a finite population, the constitution of the population changes at every draw (because any element that is drawn at any one draw is not re-placed into the population before the next draw). Hence, we cannot say that the various trials are independent, and hence the formula that is appropriate in this particular situation is the hyper geometric formula. *But*, if the population size is much larger than the sample size (so that we can regard it as an 'infinite' population), then we note that, although we are not re-placing any element that has been drawn back into the population, the population remains almost undisturbed. As such, we can assume that the various trials (i.e. draws) are independent, and, once again, we can apply the binomial formula.
In this regard, the generally accepted rule is that the *binomial* formula *can* be applied when we are drawing a sample from a finite population *without replacement* and the sample size n is not more than 5 percent of the population size N, or, to put it in another way, when $n < 0.05\ N$.
When n is greater than 5 percent of N, the *hyper geometric* formula should be used.

Next, we discuss the Poisson Distribution.

POISSON DISTRIBUTION

The Poisson distribution is named after the French mathematician Sime'on Denis Poisson (1781-1840) who published its derivation in the year 1837.THE POISSON DISTRIBUTION ARISES IN THE FOLLOWING TWO SITUATIONS:

- It is a limiting approximation to the binomial distribution, when p, the probability of success is very small but n, the number of trials is so large that the product np = μ is of a moderate size;
- a distribution in its *own* right by considering a *POISSON PROCESS* where events occur *randomly* over a specified interval of *time* or *space* or *length*.

Such random events might be the number of typing errors per page in a book, the number of traffic accidents in a particular city in a 24-hour period, etc.
With regard to the *first* situation, if we assume that n goes to infinity and p approaches zero in such a way that μ = np remains constant, then the limiting form of the binomial probability distribution is

$$\lim_{\substack{n\to\infty \\ p\to 0}} b(x;n,p) = \frac{e^{-\mu}\mu^{x}}{x!}, \quad x = 0,1,2,...,\infty$$

where **e = 2.71828.**
The Poisson distribution has only one parameter μ > 0.

The parameter μ may be interpreted as the mean of the distribution.

Although the theoretical requirement is that n should tend to infinity, and p should tend to zero, but in *PRACTICE*, generally, most statisticians use the Poisson approximation to the binomial when

p is 0.05 or less,
& n is 20 or more,

but *in fact*, the *LARGER* n is and the *SMALLER* p is, the *better* will be the approximation. We illustrate *this* particular application of the Poisson distribution with the help of the following example:

EXAMPLE

Two hundred passengers have made reservations for an airplane flight. If the probability that a passenger who has a reservation will not show up is 0.01, what is the probability that exactly three will not show up?

SOLUTION

Let us regard a "no show" as success. Then this is essentially a *binomial* experiment with n = 200 and p = 0.01. Since p is very small and n is considerably large, we shall apply the Poisson distribution, using
μ= np = (200) (0.01) = 2.
Therefore, if X represents the number of successes (not showing up), we have

$$P(X=3) = \frac{e^{-2}(2)^3}{3!}$$

$$= \frac{(0.1353)(8)}{3\times 2\times 1} = 0.1804$$

$$\left(\because e^{-2} = \frac{1}{(2.71828)^2} = 0.1353\right)$$

POISSON PROCESS

may be defined as a *physical* process governed at least in *part* by some *random* mechanism.
Stated differently a poisson process represents a situation where events occur *randomly* over a specified interval of *time* or *space* or *length*. Such random events might be the number of taxicab arrivals at an intersection per day; the number of traffic deaths per month in a city; the number of radioactive particles emitted in a given period; the number of flaws per unit length of some material; the number of typing errors per page in a book; etc.
The formula valid in the case of a Poisson process is:

$$P(X=x) = \frac{e^{-\lambda t}(\lambda t)^{x}}{x!},$$

where
λ = average number of occurrences of the outcome of interest per unit of time,
t = number of time-units under consideration, and
x = number of occurrences of the outcome of interest in t units of time.
We illustrate this concept with the help of the following example:

EXAMPLE

Telephone calls are being placed through a certain exchange at random times on the average of four per minute. Assuming a Poisson Process, determine the probability that in a 15-second interval, there are 3 or more calls.

SOLUTION

Step-1: Identify the *unit* of time:
In this problem we take a minute as the unit of time.

Step-2: Identifyλ, the *average* number of occurrences of the outcome of interest per unit of time,
In this problem we have the information that, on the average, 4 calls are received per minute, hence:

$\lambda = 4$

Step-3: Identify t, the number of time-units under consideration. In this problem, we are interested in a 15-second interval, and since 15 seconds are equal to 15/60 = ¼ minutes i.e. 1/4 units of time, therefore t = 1/4
Step-4: Compute λt: In this problem,

$\lambda = 4$, &
$t = 1/4$,

Hence:

$\lambda t = 4 \times \frac{1}{4} = 1$

Step-5: Apply the Poisson formula

$$P(X = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!},$$

In this problem, since λt = 1, therefore and since we are interested in 3 or more calls in a 15-second interval, therefore

$P(X > 3) = 1 - P(X < 3)$
$= 1 - [P(X=0)+P(X=1)+P(X=2)]$

$$= 1 - \sum_{x=0}^{2} \frac{e^{-}(1)^x}{x!}$$

$$= 1 - \sum_{x=0}^{2} \frac{(0.3679)(1)^x}{x!} \quad (\because e^{-1} = 0.3679)$$

$$= 1 - (0.91975) = 0.08025$$

Hence the probability is only 8% (i.e. a very low probability) that in a 15-second interval, the telephone exchange receives 3 or more calls.

PROPERTIES OF THE POISSON DISTRIBUTION

Some of the main properties of the Poisson distribution are given below:

- If the random variable X has a Poisson distribution with parameter μ, then its mean and variance are given by $E(X) = \mu$ and $Var(X) = \mu$.
- (In other words, we can say that the mean of the Poisson distribution is *equal* to its variance.)
- The shape of the Poisson distribution is *positively skewed*. The distribution tends to be symmetrical as μ becomes *larger and larger*.

Comparing the Poisson distribution with the binomial, we note that, whereas the binomial distribution can be symmetric, positively skewed, *or* negatively skewed (depending on whether $p = 1/2$, $p < 1/2$, *or* $p > 1/2$), the Poisson distribution can never be negatively skewed.

FITTING OF A POISSON DISTRIBUTION TO REAL DATA

Just as we discussed the fitting of the binomial distribution to real data in the last lecture, the Poisson distribution can *also* be fitted to real-life data. The procedure is very similar to the one described in the case of the fitting of the binomial distribution: The population mean μ is replaced by the sample mean $\bar{X}$, and the probabilities of the various values of X are computed using the Poisson formula. The *chi-square test of goodness of fit* enables us to determine whether or not it is a good fit i.e. whether or not the discrepancy between the expected frequencies and the observed frequencies is small. Next, we discuss some important mathematical points regarding Poisson distribution.

- **1)** The Poisson approximation to the binomial formula works well when
  $n > 20$ and $p < 0.05$.
- **2)** Suppose that the Poisson is used to approximate the binomial which, in *turn*, is being used to approximate the hyper geometric. Then the *Poisson* is being used to approximate the hyper geometric Putting the two approximation conditions *together*, the rule of *thumb* is that the Poisson distribution can be used to approximate the hyper geometric distribution when $n < 0.05N$, $n > 20$, and $p < 0.05$

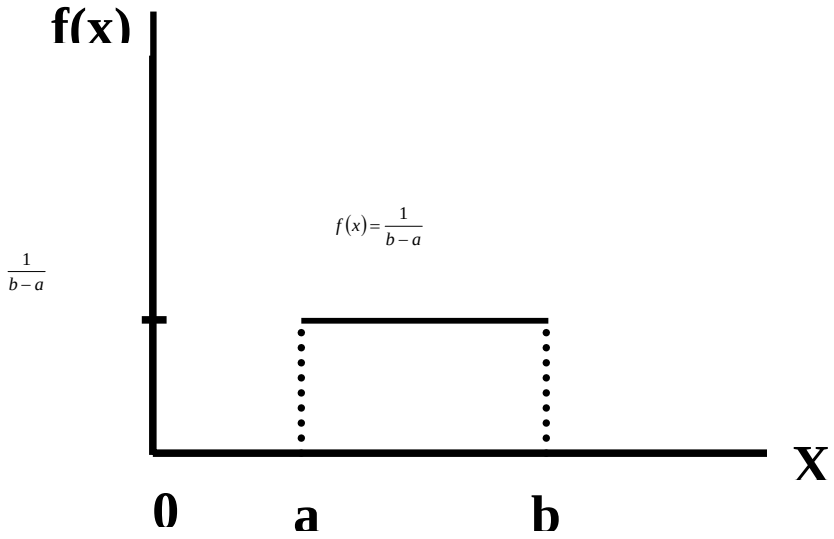
This brings to the *end* of the discussion of some of the most important and well-known Univariate discrete probability distributions. We now begin the discussion some of the well-known Univariate continuous probability distribution. There are different types of continuous distributions e.g. the *uniform* distribution, the *normal* distribution, the *geometric* distribution, and the *exponential* distribution. Each one has its *own* shape and its *own* mathematical properties. In this course, we will discuss the uniform distribution and the normal distribution.
We begin with the continuous UNIFORM DISTRIBUTION (also known as the RECTANGULAR DISTRIBUTION).

UNIFORM DISTRIBUTION

A random variable X is said to be uniformly distributed if its density function is defined as

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b$$

The graph of this distribution is as follows



The above function is a *proper* probability density function because of the fact that:
i) Since a < b, therefore f(x) > 0

ii) $$\int_{-\infty}^{\infty} f(x)dx = \int_{a}^{b} \frac{1}{b-a}dx = \frac{1}{b-a}\left[x\right]_{a}^{b} = \frac{b-a}{b-a} = 1$$

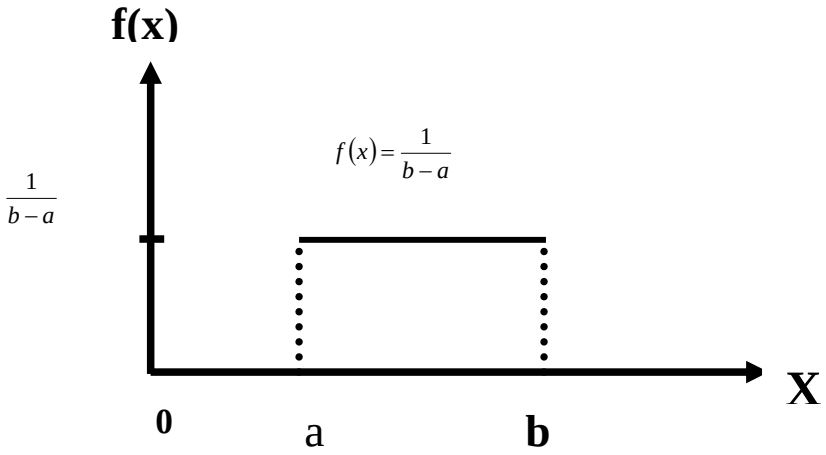
Since the shape of the distribution is like that of a rectangle, therefore the total area of this distribution can *also* be obtained from the simple formula:
Area of rectangle
= (Base) × (Height)

$$= (b-a) \times \left(\frac{1}{b-a}\right) = 1$$

Area under the Uniform Distribution

= Area of the rectangle
= (Base) × (Height)

$$= (b-a)\times\left(\frac{1}{b-a}\right)=1$$



The distribution derives its *name* from the fact that its density is constant or *uniform* over the interval [a, b] and is 0 elsewhere. It is also called the rectangular distribution because its total probability is confined to a rectangular region with base equal to (b – a) and height equal to 1/(b – a). The *parameters* of this distribution are a and b with

$$\mu=\frac{a+b}{2} \text{ and } \underline{\text{variance}} \text{ is } \sigma^2=\frac{(b-a)^2}{12}$$

PROPERTIES OF THE UNIFORM DISTRIBUTION

Let X has the uniform distribution over [a, b]. Then its mean is

The uniform probability distribution provides a *model* for continuous random variables that are *evenly distributed over a certain interval*. That is, a uniform random variable is one that is just *as* likely to assume a value in *one* interval as it is to assume a value in any *other* interval *of equal size*. There is *no clustering* of values around *any* value. Instead, there is an *even spread* over the *entire* region of possible values. As far as the *real-life application* of the uniform distribution is concerned, the point to be noted is that, for *continuous* random variables there is an *infinite* number of values in the sample space, but in *some* cases, *the values may appear to be equally likely*.

EXAMPLE-1

If a short exists in a 5 meter stretch of electrical wire, it may have an equal probability of being in any particular 1 centimeter segment along the line.

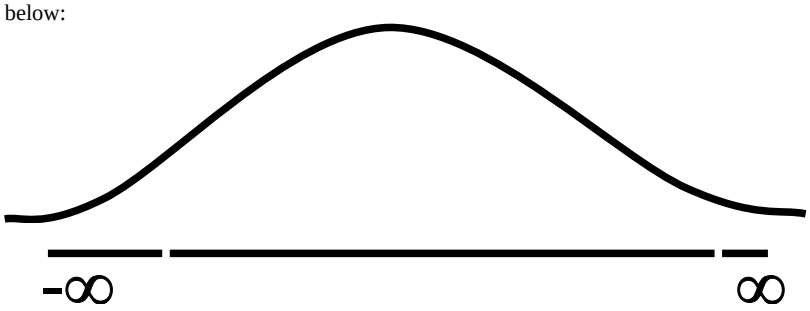
EXAMPLE-2

If a safety inspector plans to choose a time at random during the 4 afternoon work-hours to pay a surprise visit to a certain area of a plant, then each 1 minute time-interval in this 4 work-hour period will have an *equally likely* chance to being selected for the visit. Also, the uniform distribution arises in *the study of rounding off errors*, etc.

LECTURE NO. 30

- Normal Distribution.
  - Mathematical Definition
  - Important Properties
- The Standard Normal Distribution
  - Direct Use of the Area Table
  - Inverse Use of the Area Table
- Normal Approximation to the Binomial Distribution

The normal distribution was discovered in 1733. The normal distribution has a bell-shaped curve of the type shown below:



Let us begin its detailed discussion by considering its formal MATHEMATICAL DEFINITION, and its main PROPERTIES.

NORMAL DISTRIBUTION

A continuous random variable is said to be normally distributed with mean μ and standard deviation σ if its probability density function is given by

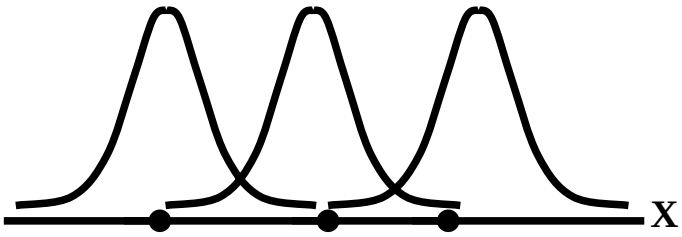
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}, \quad -\infty<x<\infty \qquad \left(\begin{array}{l}\text{where}\\ \pi=3.1416\simeq 22/7,\\ e\simeq 2.71828\end{array}\right)$$

For any particular value of μ and any particular value of σ, giving different values to x and we obtain a set of ordered pairs (x, f(x)) that yield the bell-shaped curve given above. The formula of the normal distribution defines a *FAMILY* of distributions depending on the values of the two *parameters* μ and σ (as these are the two values that determine the shape of the distribution).

PROPERTIES OF THE NORMAL DISTRIBUTION

Property No. 1

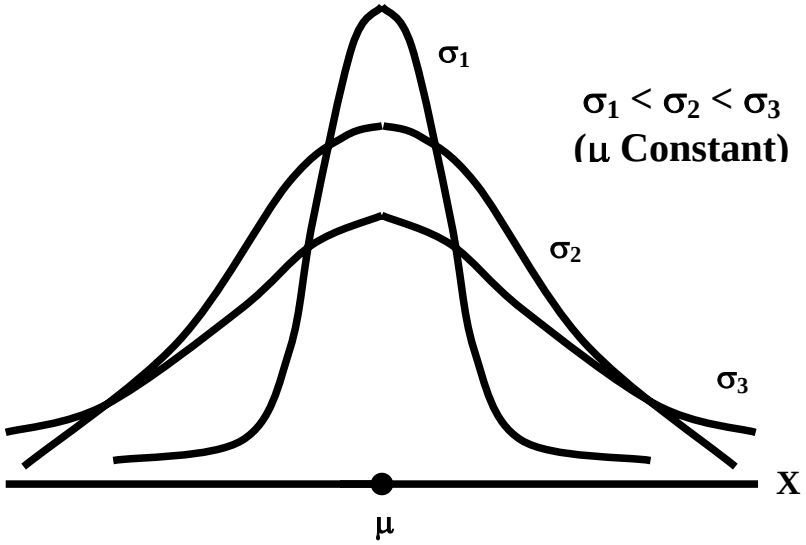
It can be mathematically proved that, for the normal distribution N(μ,σ2), μ represents the *mean*, and σ represents the *standard deviation* of the normal distribution. A change in the mean μ *shifts* the distribution to the left or to the right along the x-axis:



$\mu_1 < \mu_2 < \mu_3$

(σ Constant)

The different values of the standard deviationσ, (which is a measure of *dispersion*), determine the *flatness* or *peakedness* of the normal curve. In other words, achange in the standard deviation on σ *flattens* it or *compresses* it while leaving its centre in the same position:



Property No. 2

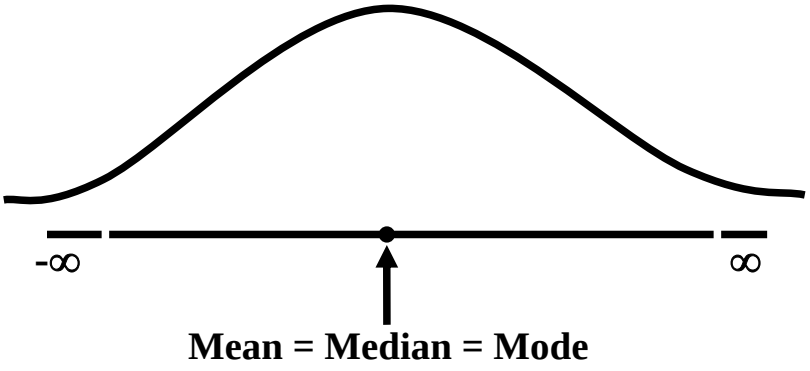
The normal curve is asymptotic to the x-axis as $x \to \pm \infty$.

Property No. 3

Because of the *symmetry* of the normal curve, 50% of the area is to the right of a vertical line erected at the mean, and 50% is to the left.(Since the total area under the normal curve from $-\infty$ to $+\infty$ is unity, therefore the area to the left of μ is 0.5 and the area to the right of μ is also 0.5.)

Property No. 4

The density function attains its maximum value at $x = \mu$ and falls off symmetrically on each side of μ. This is why the mean, median and mode of the normal distribution are all equal to μ.



Property No. 5

Since the normal distribution is absolutely symmetrical, hence $\mu 3$, the third moment about the mean is zero.

Property No. 6

For the normal distribution, it can be mathematically proved that $\mu 4 = 3\ \sigma 4$

Property No. 7

The moment ratios of the normal distribution come out to be 0 and 3 respectively:

Moment Ratios:

$$\beta_1 = \frac{\mu^2{}_3}{\mu_2{}^3} = \frac{0^2}{\left(\sigma^2\right)^3} = 0,$$

$$\beta_2 = \frac{\mu^4}{\mu_2^{\,2}} = \frac{3\sigma^4}{\left(\sigma^2\right)^2} = 3$$

NOTE

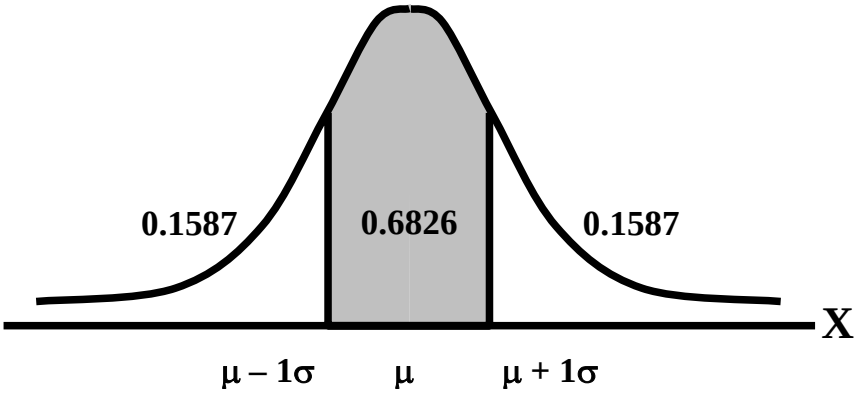
Because of the fact that, for the normal distribution, β2 comes out to be 3, this is why this value has been taken as a criterion for measuring the kurtosis of any distribution: The amount of peakedness of the *normal* curve has been taken as a *standard*, and we say that this particular distribution is masochistic. Any distribution for which β2 is greater than 3 is more peaked than the normal curve, and is called leptokurtic; Any distribution for which β2 is less than 3 is less peaked than the normal curve, and is called platykurtic.

Property No. 8

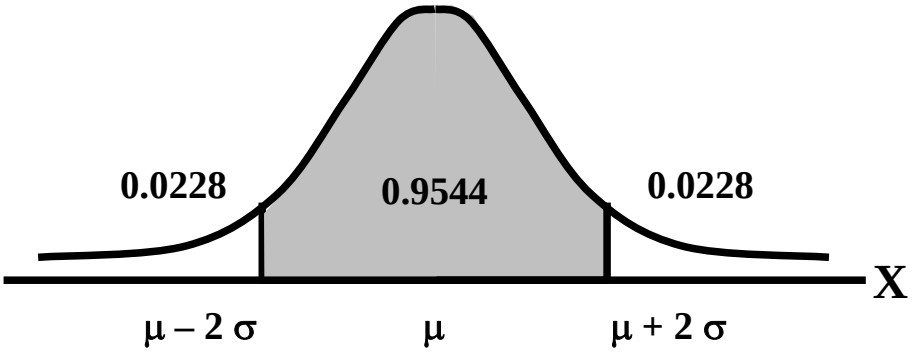
No matter what the values of μ and σ are, areas under the normal curve remain in certain *fixed* proportions within a *specified* number of standard deviations on either side of μ.

For the normal distribution:

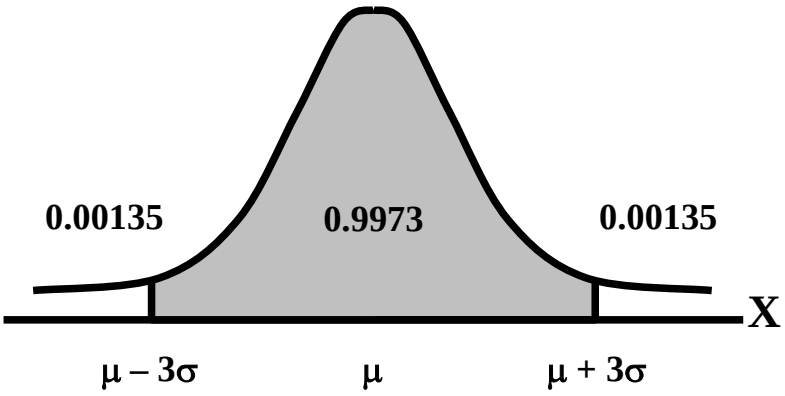
- The interval μ ± σ will always contain 68.26% of the total area.



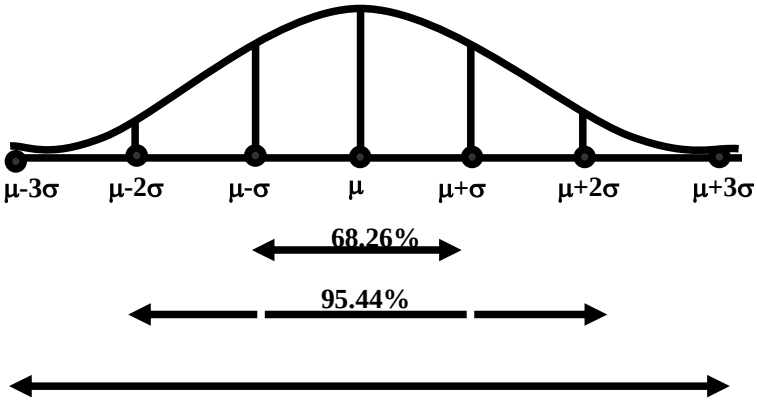
- The interval μ + 2σ will always contain 95.44% of the total area.



- The interval μ ± 3σ will always contain 99.73% of the total area.



Combining the above three results, we have:



At this point, the student are reminded of the Empirical Rule that was discussed during the first part of this course --- that on descriptive statistics. You will recall that, in the case of any approximately symmetric hump-shaped frequency distribution, approximately 68% of the data-values lie between $\bar{X}$ + S, approximately 95% between the $\bar{X}$ + 2S, and approximately 100% between $\bar{X}$ + 3S.You can now recognize the similarity between the empirical rule and the property given above. (In case a distribution is absolutely normal, the areas in the above-mentioned ranges are 68.26%, 95.44% and 99.73%; in case a distribution approximately normal, the areas in these ranges will be *approximately* equal to these percentages.)

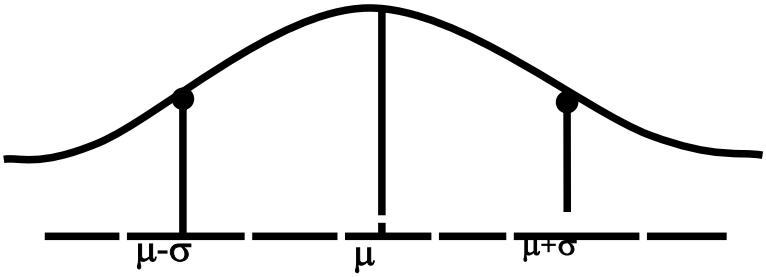
Property No. 9

The normal curve contains points of inflection (where the direction of concavity changes) which are equidistant from the mean. Their coordinates on the XY-plane are

$$\left(\mu-\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right) and \left(\mu+\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$$

respectively.

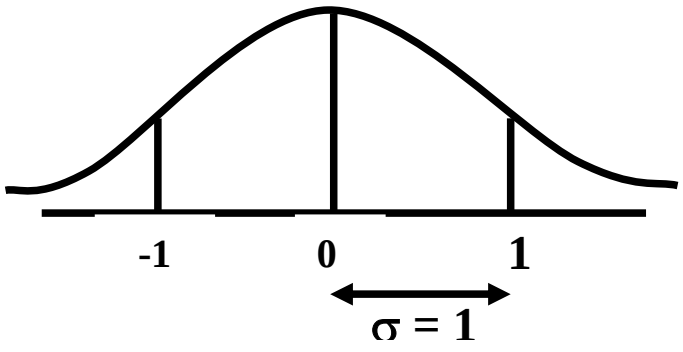
Points of Inflection



Next, we consider the concept of the Standard Normal Distribution:

THE STANDARD NORMAL DISTRIBUTION

A normal distribution whose mean is zero and whose standard deviation is 1 is known as the standard normal distribution.



This distribution has a very important role in *computing areas* under the normal curve. The *reason* is that the mathematical equation of the normal distribution is so complicated that it is not possible to find areas under the normal curve by ordinary integration. Areas under the normal curve have to be found by the more advanced method of *numerical integration*. The point to be noted is that areas under the normal curve have been computed for *that* particular normal distribution whose mean is zero and whose standard deviation is equal to 1, i.e. the standard normal distribution.

Areas under the Standard Normal Curve

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0159	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2083	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2380	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3880
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3990	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4430	0.4441
1.6	0.4452	0.4463	0.4474	0.4485	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4690	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4758	0.4762	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4865	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4980	0.4980	0.4981
2.9	0.4981	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.49865	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.49903	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993

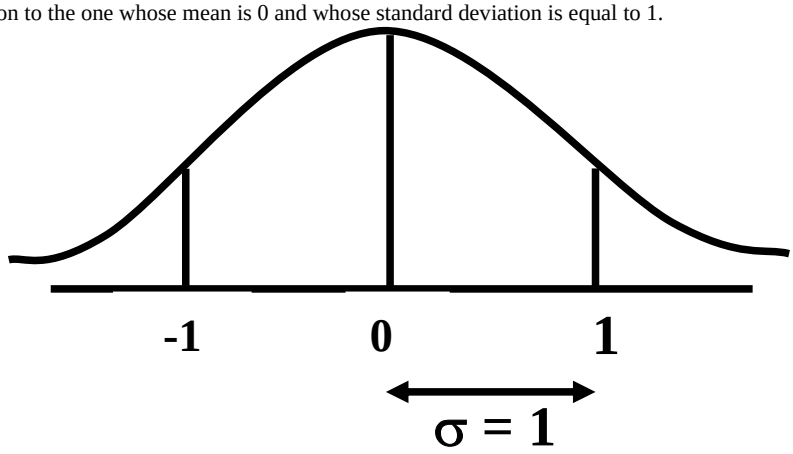
In any problem involving the normal distribution, the generally established procedure is that the normal distribution under consideration is *converted to* the standard normal distribution. This process is called *standardization*. The formula for converting N (μ, σ) to N (0, 1) is:

THE PROCESS OF STANDARDIZATION

The standardization formula is:

$$Z = \frac{X - \mu}{\sigma}$$

If X is N (μ, σ), then Z is N (0, 1). In other words, the standardization formula given above converts our normal distribution to the one whose mean is 0 and whose standard deviation is equal to 1.



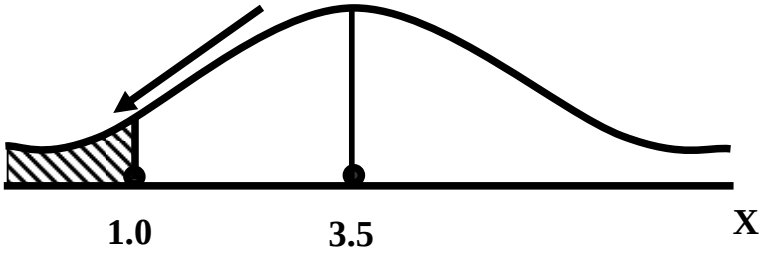
We illustrate this concept with the help of an interesting example:

EXAMPLE

The length of life for an automatic dishwasher is approximately normally distributed with a mean life of 3.5 years and a standard deviation of 1.0 years. If this type of dishwasher is guaranteed for 12 months, what fraction of the sales will require replacement?

SOLUTION

Since 12 months equal one year, hence we need to compute the fraction or *proportion* of dishwashers that will cease to function before a time-span of one year. In other words, we need to find the *probability* that a dishwasher fails before one year.



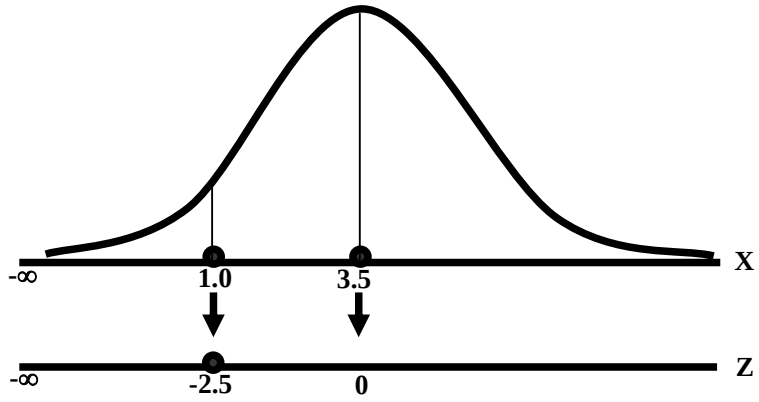
In order to find this area we nee to standardize normal distribution i.e. to convert N(3.5, 1) to N(0, 1):

The method is

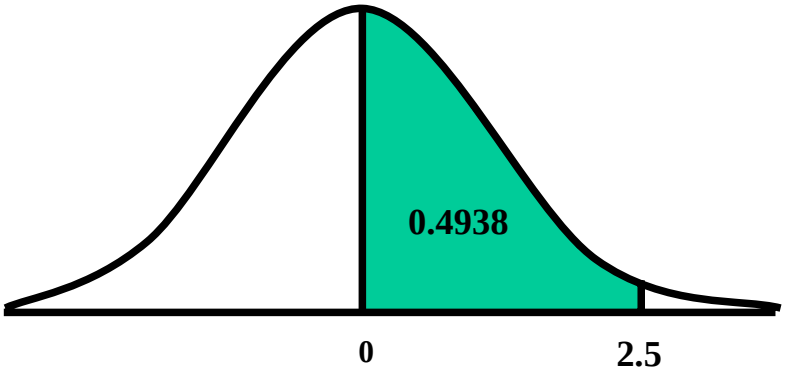
$$Z = \frac{X - \mu}{\sigma} = \frac{X - 3.5}{1.0}$$

The X-value representing the warranty period is 1.0 so

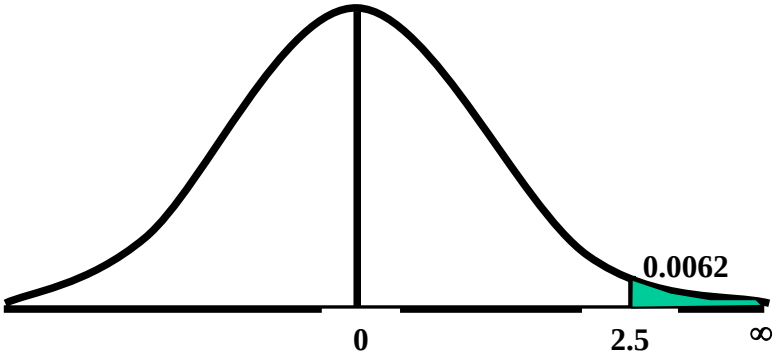
$$Z = \frac{1.0 - 3.5}{1.0} = \frac{-2.5}{1} = -2.5$$



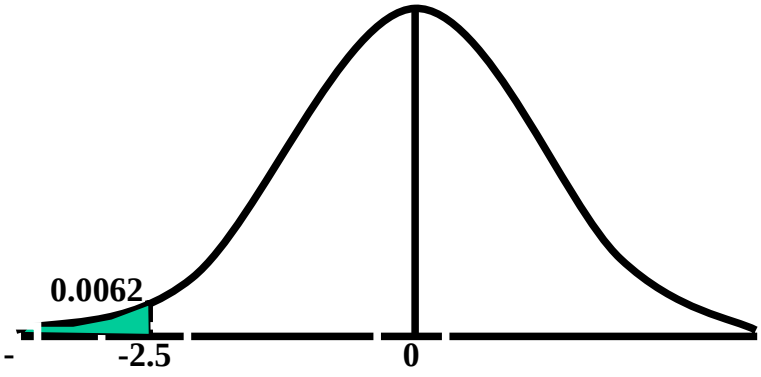
Now we need to find the area under the normal curve from z= -∞ to Z = -2.5. Looking at the area table of the standard normal distribution, we find that Area from 0 to 2.5 = 0.4938



Hence: The area from X = 2.5 to ∞ is 0.0062



But, this means that the area from -∞ to -2.5 is *also* 0.0062, as shown in the following figure:



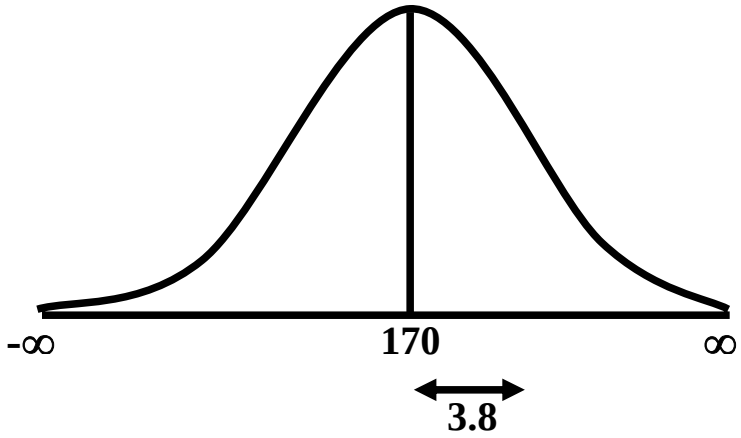
This means that the probability of a dishwasher lasting less than a year is 0.0062 i.e. 0.62% --- even less than 1%.Hence, the owner of the factory should be quite happy with the decision of placing a twelve-month guarantee on the dishwasher! Next, we discuss the Inverse use of the Table of Areas under the Normal Curve. In the above example, we were required to find a certain area against a given x-value. In some situations, we are confronted with just the opposite --- we are given certain areas, and we are required to find the corresponding x-values. We illustrate this point with the help of the following example:

EXAMPLE

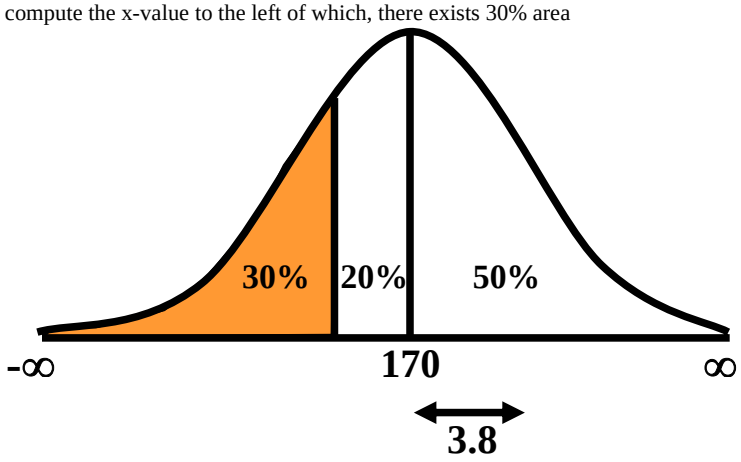
The heights of applicants to the police force in a certain country are normally distributed with mean 170 cm and standard deviation 3.8 cm. If 1000 persons apply for being inducted into the police force, and it has been decided that not more than 70% of these applicants will be accepted, (and the shortest 30% of the applicant are to be rejected), what is the minimum acceptable height for the police force?

SOLUTION:

We have:



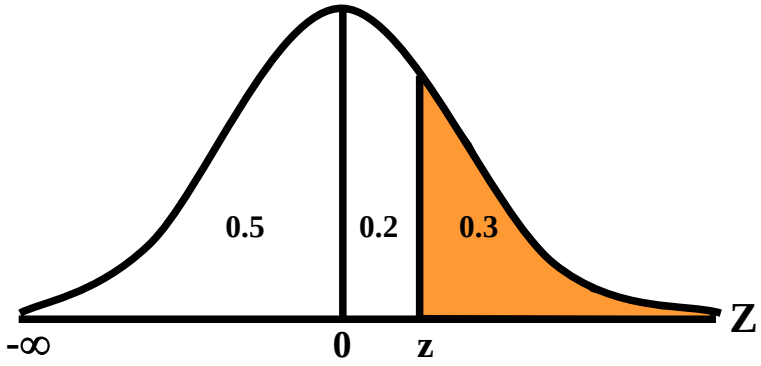
We need to compute the x-value to the left of which, there exists 30% area



The standardization formula can be re-written as

$$Z = \frac{X - \mu}{\sigma}$$

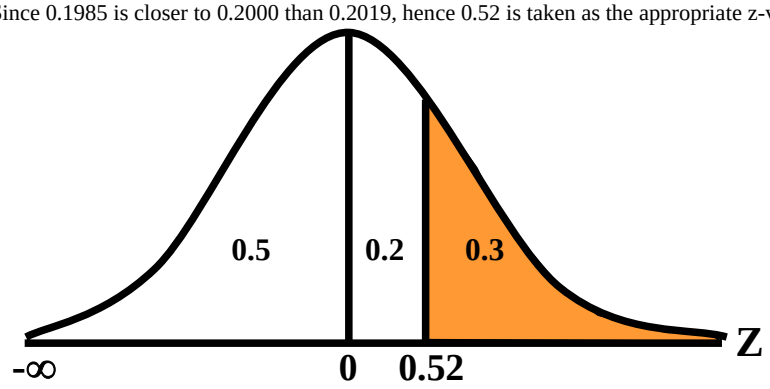
The Z value to the left of which there exists 30% area is obtained as follows.



By studying the figures inside the body of the area table of the standard normal distribution, we find that:

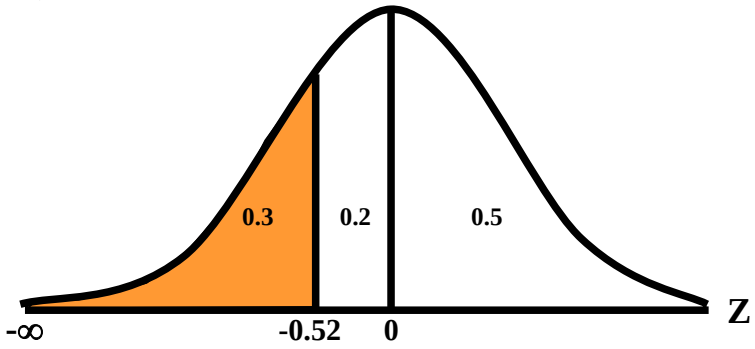
- The area between z = 0 and z = 0.52 is 0.1985, and
- The area between z = 0 and z = 2.53 is 0.2019

Since 0.1985 is closer to 0.2000 than 0.2019, hence 0.52 is taken as the appropriate z-value.



But, we are interested not in the upper 30% but the lower 30% of the applicants.

Hence, we have:



Since the normal distribution is absolutely symmetrical, hence the z-value to the left of which there exists 30% area (on the left-hand-side of the mean) will be at exactly the same distance from the mean as the z-value to the right of which there exists 30% area (on the right-hand-side of the mean).

Substituting z = -0.52 in the standardization formula, we obtain:

X = 170 + 3.8 Z
= 170 + 3.8 (-0.52)
= 170 - 1.976
= 168.024 168 cm

Hence, the minimum acceptable height for the police force is 168 cm. Just as binomial, Poisson and other discrete distributions can be *fitted* to *real-life* data; similarly, the normal distribution can also be *FITTED* to real data. This can be done by equating μ to $\bar{X}$, the mean computed from the observed frequency distribution (based on sample data), and σ to S, the standard deviation of the observed frequency distribution. Of course, this should be done only if

we are reasonably sure that the shape of the observed frequency distribution is quite similar to that of the normal distribution. (As indicated in the case of the fitting of the binomial distribution to real data), in order to *decide* whether or not our fitted normal distribution is a *reasonably good fit*, the *proper* statistical procedure is the Chi-square Test of Goodness of Fit.

NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

The probability for a binomial random variable X to take the value x is

$$f(x) = \binom{n}{x} p^x q^{n-x},$$

$$\text{for } 0 \le x \le n \text{ and } q + p = 1.$$

The above formula becomes cumbersome to apply if n is LARGE. In such a situation, *as long as neither p nor q is close to zero*, we can compute the required probabilities by applying the normal approximation to the binomial distribution. The binomial distribution can be quite closely approximated by the normal distribution *when n is sufficiently large and neither p nor q is close to zero.* As a rule of *thumb,* the normal distribution provides a reasonable approximation to the binomial distribution *if both np and nq are equal to or greater than 5,* i.e.

np > 5 and nq > 5

EXAMPLE:

Suppose that a past record indicate that, in a particular province of an under-developed country, the death rate from Malaria is 20%. Find the probability that in a particular village of that particular province, the number of deaths is between 70 and 80 (inclusive) out of a total of 500 patients of Malaria.

SOLUTION:

Regarding 'death from Malaria' as success, we have

n = 500

and p = 0.20.

It is obvious that it is very cumbersome to apply the binomial formula in order to compute P(70 < X < 80).
In this problem,

np = 500(0.2) = 100 > > > 5, and nq = 500(0.8) = 400 > > > 5,

therefore we can *happily* apply the normal approximation to the binomial distribution. In order to apply the normal approximation to the binomial, we need to keep in mind the following two points:
1) The first point is: The mean and variance of the binomial distribution valid in our problem will be regarded as the mean and variance of the normal distribution that will be used to approximate the binomial distribution.
In this problem, we have:

and $\mu = np = 500 \times 0.20 = 100$

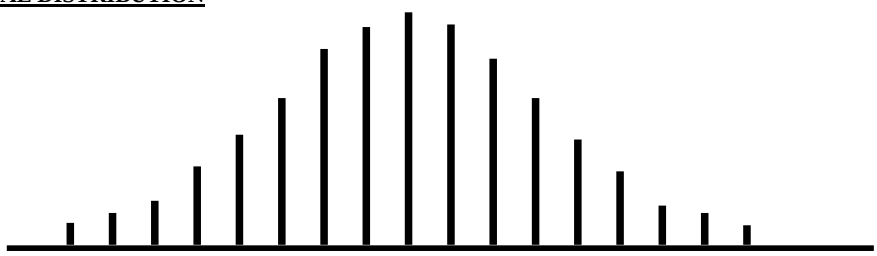
$\sigma^2 = npq = 500 \times 0.20 \times 0.80 = 80$

Hence $\sigma = \sqrt{npq} = \sqrt{80} = 8.94$

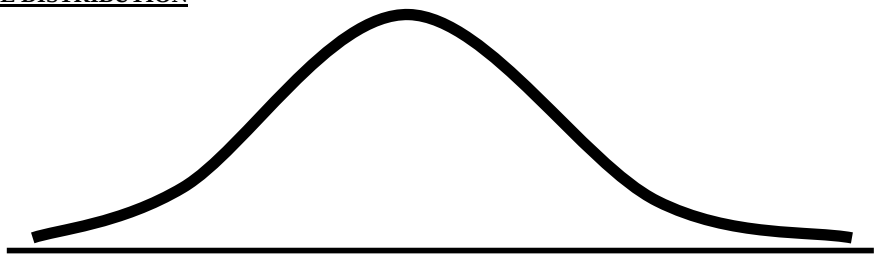
2) The second important point is:

We need to apply a correction that is known as the Continuity Correction. The rationale for this correction is as follows: The binomial distribution is essentially a discrete distribution whereas the normal distribution is a continuous distribution i.e.:

BINOMIAL DISTRIBUTION

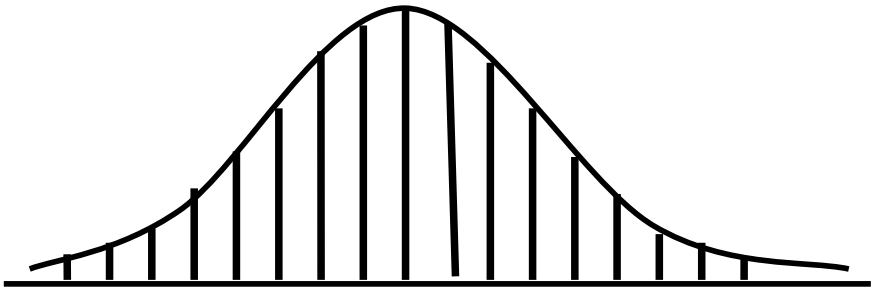


NORMAL DISTRIBUTION



In applying the normal approximation to the binomial, we have the following situation:

THE NORMAL DISTRIBUTION *SUPERIMPOSED* ON THE BINOMIAL DISTRIBUTION



But, the question arises: "How can a set of distinct vertical lines be replaced by a continuous curve?"
In order to overcome this problem, what we do is to replace every integral value x of our binomial random variable by an interval x - 0.5 to x + 0.5. By doing so, we will have the following situation. The x-value 70 is replaced by the interval 69.5 - 70.5, The x-value 71 is replaced by the interval 70.5 - 71.The x-value 72is replaced by the interval 71.5 - 72.5 ……..The x-value 80 is replaced by the interval 79.5 - 80.5

Hence:

Applying the continuity correction,

$P(70 < X < 80)$

is replaced by

$P(69.5 < X < 80.5)$.

Accordingly, the area that we need to compute is the area under the normal curve between the values 69.5 and 80.5.

It is left to the *students* to compute this area, and thus determine the required probability. (This computation involves a few steps.)

By doing so, the students will find that, in that particular village of that province, the probability that the number of deaths from Malaria in a sample of 500 lies between 70 and 80 (inclusive) is 0.0145 i.e. 1½%.

This brings us to the end of the second part of this course i.e. *Probability Theory*.

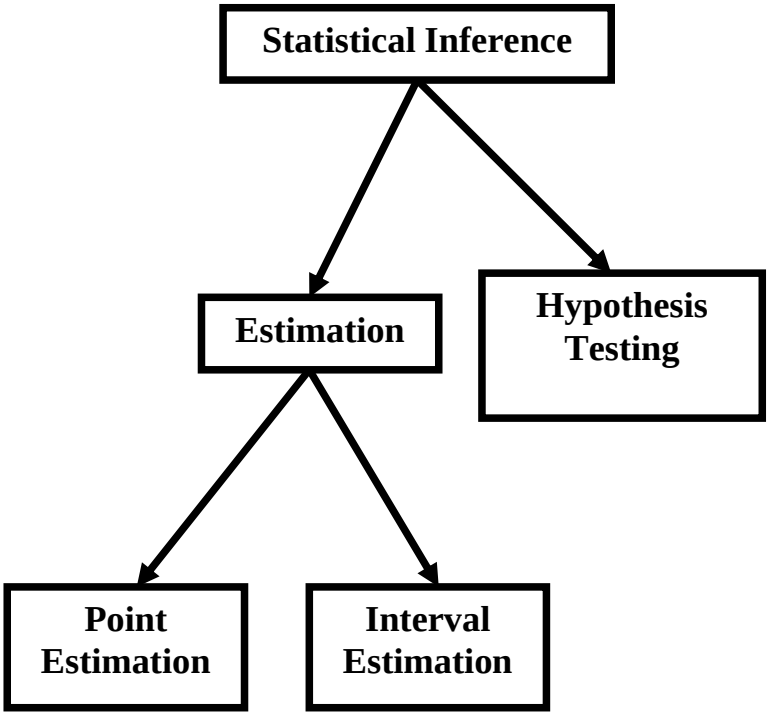
In the next lecture, we will begin the third and last portion of this course i.e. *Inferential Statistics* --- that area of Statistics which enables us to draw conclusions about various phenomena on the basis of data collected on sample basis.

LECTURE NO. 31

- Sampling Distribution of $\overline{X}$
- Mean and Standard Deviation of the Sampling Distribution of $\overline{X}$
- Central Limit Theorem

INFERENTIAL STATISTICS

That branch of Statistics which enables us to draw conclusions or inferences about various phenomena on the basis of real data collected on sample basis. In this regard, the first point to be noted is that statistical inference can be divided into two main branches --- estimation, and hypothesis-testing. Estimation itself can be further divided into two branches --- point estimation, and interval estimation



The second important point is that the concept of sampling distributions forms the basis for both estimation and hypothesis-testing,

SAMPLING DISTRIBUTION

The probability distribution of any statistic (such as the mean, the standard deviation, the proportion of successes in a sample, etc.) is known as its sampling distribution. In this regard, the first point to be noted is that there are two ways of sampling --- sampling with replacement, and sampling without replacement. In case of a finite population containing N elements, the total number of possible samples of size n that can be drawn from this population with replacement is Nn.
In case of a finite population containing N elements, the total number of possible samples of size n that can be drawn from this population without replacement.
We illustrate the concept of the sampling distribution of $\binom{N}{n}$. with the help of the following example:

EXAMPLE

Let us examine the case of an annual Ministry of Transport test to which all cars, irrespective of age, have to be submitted. The test looks for faulty breaks, steering, lights and suspension, and it is discovered after the first year that approximately the same numbers of cars have 0, 1, 2, 3, or 4 faults.
The above situation is equivalent to the following:
Let X denotes the number of faults in a car. Then X can take the values 0, 1, 2, 3, and 4, the probability of each of these X values is 1/5. Hence, we have the following probability distribution:

No. of Faulty Items (X)	Probability f(x)
0	1/5
1	1/5
2	1/5
3	1/5
4	1/5
Total	1

In order to compute the mean and standard deviation of this probability distribution, we carry out the following computations,

MEAN AND VARIANCE OF THE POPULATION DISTRIBUTION

$$\mu = E(X) = \sum xf(x) = 2$$

$$\sigma^2 = Var(X) = E(X)^2 - [E(X)]^2$$

$$= \sum x^2 f(x) - [\sum x\, f(x)]^2$$

$$= 6 - 2^2 = 6 - 4 = 2$$

Practically speaking, only a sample of the cars will be tested at any one occasion, and, as such, we are interested in considering the results that would be obtained if a sample of vehicles is tested. Let us consider the situation when only two cars are tested after being selected at the roadside by a mobile testing station. The following table gives all the possible situations:

NO. OF FAULTY ITEMS

Second Car / First Car	0	1	2	3	4
0	(0,0)	(0,1)	(0,2)	(0,3)	(0,4)
1	(1,0)	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,0)	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,0)	(3,1)	(3,2)	(3,3)	(3,4)
4	(4,0)	(4,1)	(4,2)	(4,3)	(4,4)

The above situation is equivalent to drawing all possible samples of size 2 from this probability distribution (i.e. the population) WITH REPLACEMENT. From the above list of 25 samples, we can work out all the possible sample means. These are indicated in the following table:

SAMPLE MEANS

Second Car / First Car	0	1	2	3	4
0	0.0	0.5	1.0	1.5	2.0
1	0.5	1.0	1.5	2.0	2.5
2.	1.0	1.5	2.0	2.5	3.0
3	1.5	2.0	2.5	3.0	3.5
4	2.0	2.5	3.0	3.5	4.0

It is immediately evident that some of these possible samples mean occur several times. In view of this, it would seem reasonable and sensible to construct a frequency distribution from the sample means. This is given in the following table:

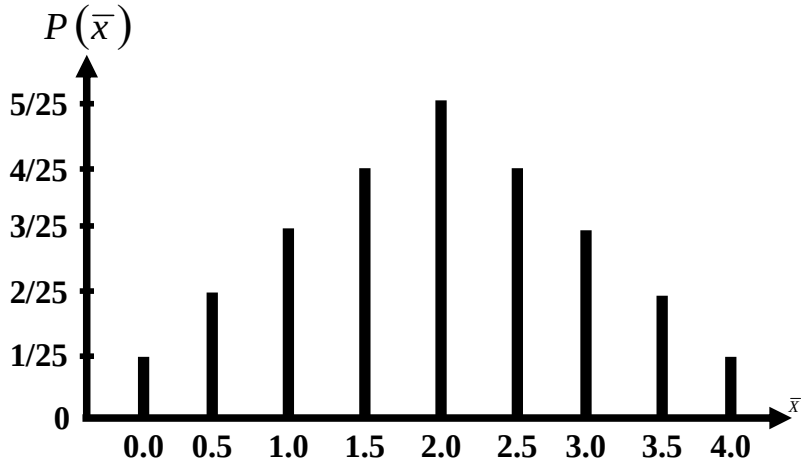
Sample Mean	No. of Samples
$\bar{x}$	f
0.0	1
0.5	2
1.0	3
1.5	4
2.0	5
2.5	4
3.0	3
3.5	2
4.0	1
Total	25

If we divide each of the above frequencies by the total frequency 25, we obtain the probabilities of the various values of $\bar{X}$.(This is so because every one of the 25 possible situations is equally likely to occur, and hence the probabilities of the various possible values of $\bar{X}$ can be computed using the classical definition of probability i.e. m/n --- number of favorable outcomes divided by total number of possible outcomes). Hence, we obtain the following probability distribution:

Sample Mean	No. of Samples	Probability
$\bar{x}$	f	$P(\bar{X} = \bar{x})$
0.0	1	1/25
0.5	2	2/25
1.0	3	3/25
1.5	4	4/25
2.0	5	5/25
2.5	4	4/25
3.0	3	3/25
3.5	2	2/25
4.0	1	1/25
Total	25	25/25=1

The above is referred to as the SAMPLING DISTRIBUTION of the mean. The visual picture of the sampling distribution is as follows:

Sampling Distribution of $\bar{X}$ for n = 2



Next, we wish to compute the mean and standard deviation of this distribution.

As we are already aware, for the probability distribution of a random variable X, the mean is given by

$\mu = E(X) = \sum x f(x)$ and the variance is given by $\sigma^2 = Var(X) = E(X2) - [E(X)]2$

The point to be noted is that, in case of the sampling distribution of $\bar{X}$, our random variable is not X but $\bar{X}$.

Hence, the mean and variance of our sampling distribution are given by

MEAN AND VARIANCE OF THE SAMPLING DISTRIBUTION OF $\bar{X}$

$$\mu_{\bar{x}} = E(\bar{X}) = \sum \bar{x} f(\bar{x})$$

$$\sigma^2_{\bar{x}} = Var(\bar{X}) = E(\bar{X})^2 - [E(\bar{X})]^2$$
$$= \sum \bar{x}^2 f(\bar{x}) - [\sum \bar{x} f(\bar{x})]^2$$

The square root of the variance is the standard deviation, and the standard deviation of a sampling distribution is termed as its standard error. In order to find the mean and standard error of the sampling distribution of $\bar{X}$ in this example, we carry out the following computations:

In order to find the mean and standard error of the sampling distribution of $\bar{X}$ in this example, we carry out the following computations:

Sample Mean	**Probability**		
$\bar{x}$	$f(\bar{x})=P(\bar{X}=\bar{x})$	$\bar{x} f(\bar{x})$	$(\bar{x})^2 f(\bar{x})$
0.0	1/25	0	0
0.5	2/25	1/25	1/50
1.0	3/25	3/25	6/50
1.5	4/25	6/25	18/50
2.0	5/25	10/25	40/50
2.5	4/25	10/25	50/50
3.0	3/25	9/25	54/50
3.5	2/25	7/25	49/50
4.0	1/25	4/25	32/50
Total	25/25=1	50/25=2	250/50=5

Hence, in this example, we have:

$$\mu_{\bar{x}} = E(\bar{X}) = \sum \bar{x} f(\bar{x})$$
$$= 50/25 = 2$$

And

$$\sigma^2_{\bar{x}} = Var(\bar{X}) = E(\bar{X})^2 - [E(\bar{X})]^2$$
$$= \sum \bar{x}^2 f(\bar{x}) - [\sum \bar{x} f(\bar{x})]^2$$
$$= 5 - 2^2 = 5 - 4 = 1$$
$$\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}} = \sqrt{1} = 1$$

These computations lead to the following two very important properties of the sampling distribution of $\bar{X}$

Property No.1

In the case of sampling with replacement as well as in the case of sampling without replacement, we have:

In this example: $\mu_{\bar{x}} = \mu$

$$\mu = 2$$

and

$$\mu_{\bar{x}} = 2$$

Hence

Property No.2 $\mu_{\bar{x}} = \mu$

In case of sampling with replacement:

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

In this example:

$$\sigma = \sqrt{2}$$

$$\therefore \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$$

$$and\ \sigma_{\bar{x}} = 1$$

$$Hence\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

NOTE:

In case of sampling without replacement from a finite population:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

The factor $\sqrt{\frac{N-n}{N-1}}$

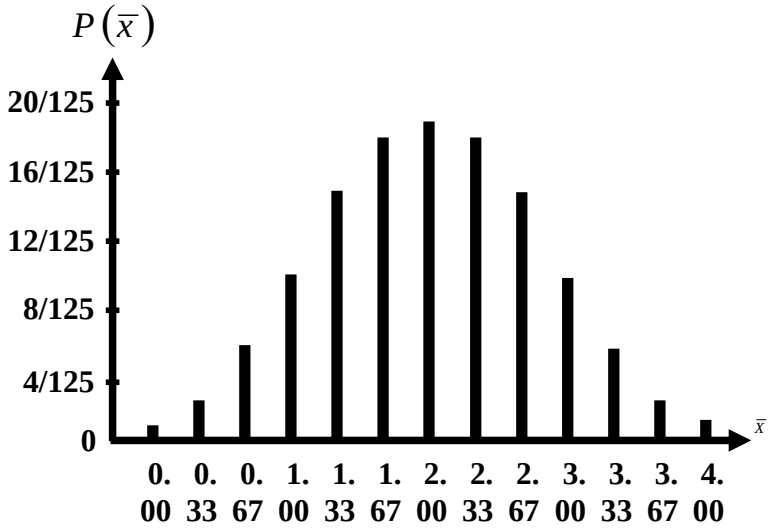
is known as the finite population correction (fpc). The point to be noted is that, if the sample size n is much smaller than the population size N, then is approximately equal to 1, and, as such, the fpc is not required. Hence, in sampling from a finite population, we apply the fpc only if the sample size is greater than 5% of the population size. Next, we consider the shape of the sampling distribution of $\bar{X}$. As indicated by the line chart, the above sampling distribution is absolutely symmetric and triangular. But let us consider what will happen to the shape of the sampling distribution with if the sample size is increased. If in the car tests instead of taking samples of 2 we had taken all possible samples of size 3, our sampling distribution would contain 53 = 125 sample means, and it would be in the following form:

SAMPLING DISTRIBUTION
FOR SAMPLES OF SIZE **3**

$\bar{x}$	No. of Samples	$f(\bar{x})$
0.00	1	1/125
0.33	3	3/125
0.67	6	6/125
1.00	10	10/125
1.33	15	15/125
1.67	18	18/125
2.00	19	19/125
2.33	18	18/125
2.67	15	15/125
3.00	10	10/125
3.33	6	6/125
3.67	3	3/125
4.00	1	1/125
	125	1

The graph of this distribution is as follows:

Sampling Distribution of $\bar{X}$ for n = 3

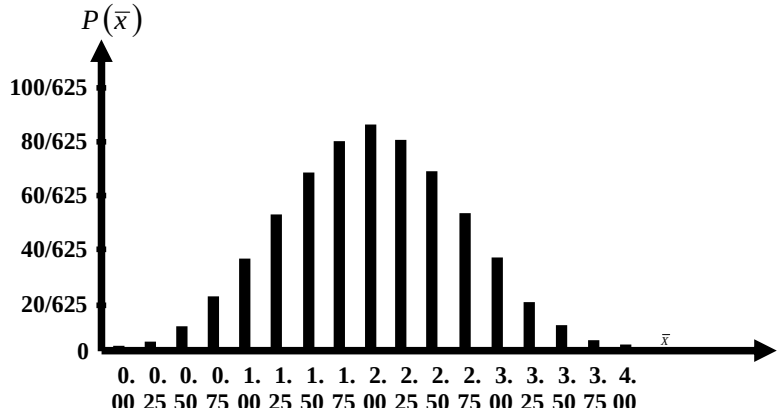


If in the car tests instead of taking samples of 2 we had taken all possible samples of size 4, our sampling distributions would contain 54 = 625 sample means, and it would be in the following form:

SAMPLING DISTRIBUTION
FOR SAMPLES OF SIZE 4

$\bar{x}$	No. of Samples	$f(\bar{x})$
0.00	1	1/625
0.25	4	4/625
0.50	10	10/625
0.75	20	20/625
1.00	35	35/625
1.25	52	52/625
1.50	68	68/625
1.75	80	80/625
2.00	85	85/625
2.25	80	80/625
2.50	68	68/625
2.75	52	52/625
3.00	35	35/625
3.25	20	20/625
3.50	10	10/625
3.75	4	4/625
4.00	1	1/625
	625	1

The graph of this distribution is as follows, Sampling Distribution of $\bar{X}$ for n = 4



As in the case of the sampling distribution of $\bar{X}$ based on samples of size 2, each of these two distributions has a mean of 2 defective items. It is clear from the above figures that as larger samples are taken, the shape of the sampling distribution undergoes discernible changes.

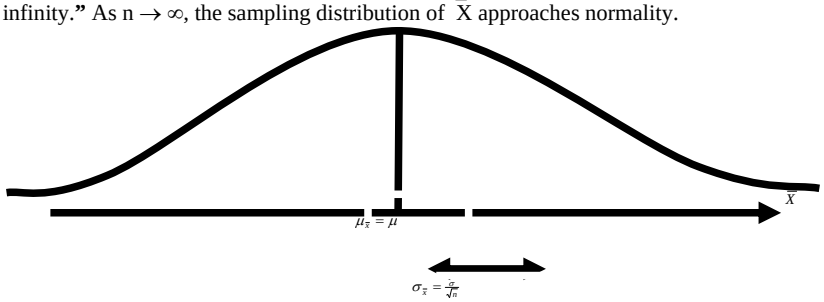
In all three cases the line charts are symmetrical, but as the sample size increases, the overall configuration changes from a triangular distribution to a bell-shaped distribution. When relatively large samples are taken, this bell-shaped distribution assumes the form of a 'normal' distribution (also called the 'Gaussian' distribution), and this happens irrespective of the form of the parent population. (For example, in the problem currently under consideration, the population of defective items in a car is rectangular.)

This leads us to the following fundamentally important theorem:

CENTRAL LIMIT THEOREM

The theorem states that:

"If a variable X from a population has mean μ and finite variance σ2, then the sampling distribution of the sample mean $\bar{X}$ approaches a normal distribution with mean μ and variance σ2/n as the sample size n approaches infinity.**"** As $n \to \infty$, the sampling distribution of $\bar{X}$ approaches normality.



Due to the Central Limit Theorem, the normal distribution has found a central place in the theory of statistical inference.(Since, in many situations, the sample is large enough for our sampling distribution to be approximately normal, therefore we can utilize the mathematical properties of the normal distribution to draw inferences about the variable of interest). The rule of thumb in this regard is that if the sample size, n, is greater than or equal to 30, then we can assume that the sampling distribution of $\bar{X}$ is approximately normally distributed. On the other hand, If the POPULATION sampled is normally distributed, then the sampling distribution of $\bar{X}$ will also be normal regardless of sample size. In other words, $\bar{X}$ will be normally distributed with mean μ and variance σ2/n.

LECTURE NO. 32

- Sampling Distribution of $\hat{p}$
- Sampling Distribution of $\overline{X}_1 - \overline{X}_2$

We discussed the mean and the standard deviation of the sampling distribution, and, towards the end of the lecture, we consider the very important theorem known as the Central Limit Theorem. Let us now consider the *real-life* application of this concept with the help of an example:

EXAMPLE

A construction company has 310 employees who have an average annual salary of Rs.24,000.The standard deviation of annual salaries is Rs.5,000.
Suppose that the employees of this company launch a demand that the government should institute a law by which their average salary should be at least Rs. 24500, and, suppose that the government decides to check the validity of this demand by drawing a random sample of 100 employees of this company, and acquiring information regarding their present salaries. What is the probability that, in a random sample of 100 employees, the average salary will exceed Rs.24,500 (so that the government decides that the demand of the employees of this company is unfounded, and hence does not pay attention to the demand(although, in reality, it was justified))?

SOLUTION

The sample size (n = 100) is large enough to assume that the sampling distribution of $\overline{X}$ is approximately *normally* distributed with the following mean and standard deviation:
and standard deviation

$$\mu_{\bar{x}} = \mu = Rs.24{,}000.$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{5000}{\sqrt{100}}\sqrt{\frac{310-100}{310-1}}$$

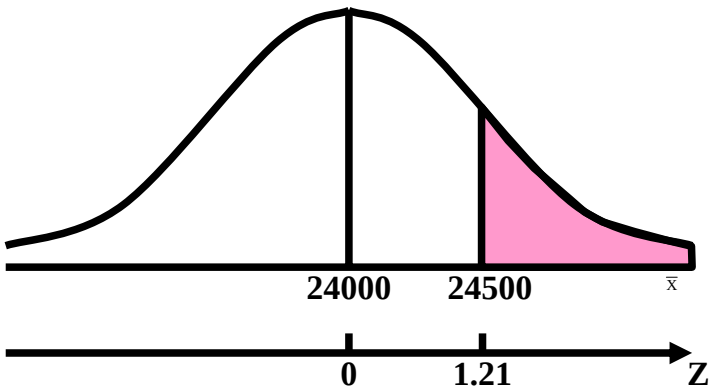
$$= \text{Rs.}412.20$$

NOTE: Here we have used finite population correction factor (fpc), because the sample size n = 100 is *greater than 5 percent* of the population size N = 310. Since $\overline{X}$ is approximately N (24000, 412.20), therefore

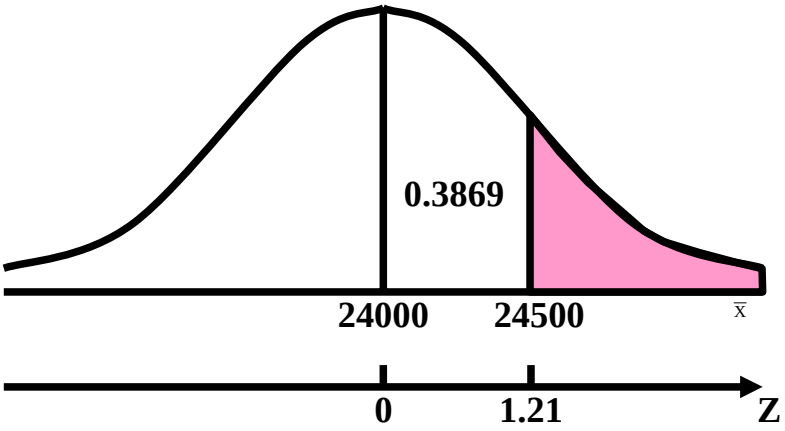
$$Z = \frac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\overline{X} - 24000}{412.20}$$

is approximately N(0, 1).We are required to evaluate P($\overline{X}$ > 24,500).
At $\bar{x}$ = 24,500, we find that

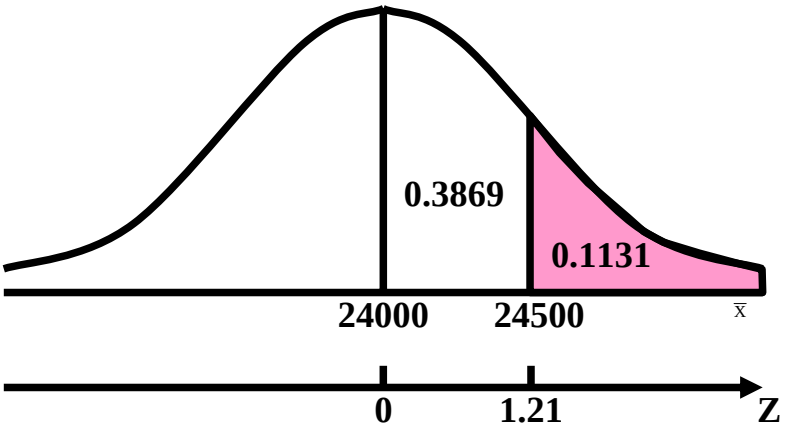
$$z = \frac{24500 - 24000}{412.20} = 1.21$$



Using the table of areas under the standard normal curve, we find that the area between z = 0 and z = 1.21 is 0.3869.



Hence,
P($\bar{X}$ > 24,500)
= P(Z > 1.21)
= 0.5 – P(0 < Z < 1.21)
= 0.5 – 0.3869 = 0.1131.



Hence, the chances are only 11% that in a random sample of 100 employees from this particular construction company, the average salary will exceed Rs.24,500.In other words, the chances are 89% that, in such a sample, the average salary will *not* exceed Rs.24,500.

Hence, the chances are considerably *high* that the government *might* pay attention to the employees' demand.

SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

In this regard, the first point to be noted is that, whenever the elements of a population can be classified into *two* categories, technically called "success" and "failure", we may be interested in *the proportion of "successes"* in the population. If X denotes the number of successes in the population, then the proportion of successes in the population is given by

$$p = \frac{X}{N}.$$

Similarly, if we draw a sample of size n from the population, the proportion of successes in the sample is given by

$$\hat{p} = \frac{X}{n},$$

where X represents the number of successes in the sample.
It is interesting to note that X is a *binomial* random variable and the binomial parameter p is being called a proportion of successes here. The sample proportion has different values in different samples. It is obviously a random variable and has a probability distribution.
This probability distribution of the proportions of successes in all possible random samples of size n, is called the sampling distribution of $\hat{p}$.
We illustrate this sampling distribution with the help of the following examples:

EXAMPLE-1

A population consists of six values 1, 3, 6, 8, 9 and 12.Draw all possible samples of size n = 3 *without* replacement from the population and find the proportion of even numbers in each sample. Construct the sampling distribution of sample proportions and verify that

i) $\mu_{\hat{p}} = p$

ii) $\text{Var}(\hat{p}) = \frac{pq}{n} \cdot \frac{N-n}{N-1}.$

SOLUTION

The number of possible samples of size n = 3 that could be selected without replacement from a population of size N is

$$\binom{6}{3} = 20.$$

Let $\hat{p}$ represent the proportion of even numbers in the sample. Then the 20 possible samples and the proportion of even numbers are given as follows:

Sample No.	Sample Data	Sample Proportion $(\hat{p})$
1	1, 3, 6	1/3
2	1, 3, 8	1/3
3	1, 3, 9	0
4	1, 3, 12	1/3
5	1, 6, 8	2/3
6	1, 6, 9	1/3
7	1, 6, 12	2/3
8	1, 8, 9	1/3
9	1, 8, 12	2/3
10	1, 9, 12	1/3
11	3, 6, 8	2/3
12	3, 6, 9	1/3
13	3, 6, 12	2/3
14	3, 8, 9	1/3
15	3, 8, 12	2/3
16	3, 9, 12	1/3
17	6, 8, 9	2/3
18	6, 8, 12	1
19	6, 9, 12	2/3
20	8, 9, 12	2/3

The sampling distribution of sample proportion is given below;

SAMPLING DISTRIBUTION OF $\hat{p}$:

$(\hat{p})$	No. of Samples	Probability $f(\hat{p})$	$\hat{p}\, f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
0	1	1/20	0	0
1/3	9	9/20	3/20	1/20
2/3	9	9/20	6/20	4/20
1	1	1/20	1/20	1/20
$\sum$	20	1	10/20	6/20

Now

$$\mu_{\hat{p}} = \sum \hat{p}\, f(\hat{p}) = \frac{10}{20} = 0.5, \text{ and}$$

$$\sigma^2_{\hat{p}} = \sum \hat{p}^2 f(\hat{p}) - [\sum \hat{p}\, f(\hat{p})]^2$$

$$= \frac{2}{60} - \left(\frac{10}{20}\right)^2 = \frac{1}{20} = 0.05.$$

To verify the given relations, we first calculate the population proportion p.Thus:

$p = \frac{X}{N}$;Where X represents the number of even numbers in the population. In other words,

$p = \frac{3}{6} = 0.5$

Hence, we find that

$\mu_{\hat{p}} = 0.5 = p$,

$$\frac{pq}{n} \cdot \frac{N-n}{N-1} = \frac{0.25}{3} \cdot \frac{6-3}{6-1}$$

$$= \frac{0.25}{5} = 0.05 = \text{Var}(\hat{p})$$

Hence, two properties of the sampling distribution of $\hat{p}$ are verified.

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\frac{N-n}{N-1}},$$

The sampling distribution of $\hat{p}$ has the following important properties.

PROPERTIES OF THE SAMPLING DISTRIBUTION OF $\hat{p}$

Property No. 1

The mean of the sampling distribution of proportions, denoted by $\mu_{\hat{p}}$, is equal to the population proportion p, that is

$$\mu_{\hat{p}} = p.$$

Property No. 2

The standard deviation of the sampling distribution of proportions, called the *standard error* of and denoted by $\hat{p}$ is given as: $\sigma_{\hat{p}}$,

- a)

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}},$$

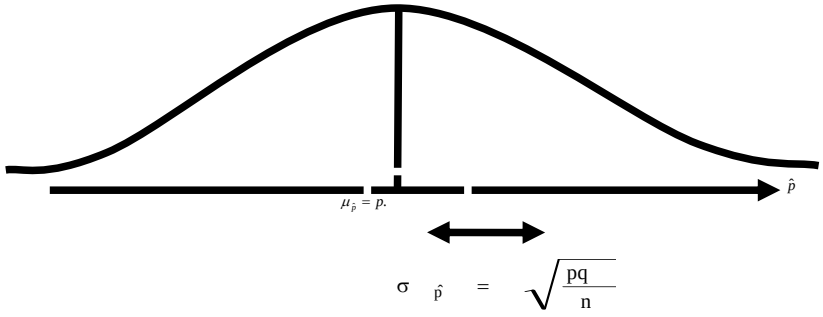
- when the sampling is performed *with* replacement
- b) when sampling is done *without* replacement from a *finite* population
  (As in the case of the sampling distribution of $\bar{X}$,is known as the finite population correction factor (fpc)

$$\sqrt{\frac{N-n}{N-1}},$$

Property No. 3

SHAPE OF THE DISTRIBUTION

The sampling distribution of $\hat{p}$ is the *binomial* distribution. However, for sufficiently *large* sample sizes, the sampling distribution of is approximately normal. As $n \to \infty$, the sampling distribution of approaches normality.



As a rule of thumb, the sampling distribution of will be approximately *normal* whenever both np and nq are equal to or greater than 5.Let us apply this concept to a real-world situation: $\hat{p}$

EXAMPLE-2

Ten percent of the 1-kilogram boxes of sugar in a large warehouse are underweight. Suppose a retailer buys a random sample of 144 of these boxes. What is the probability that at least 5 percent of the sample boxes will be underweight?

SOLUTION

Here the statistic is the sample proportion, the sample size (n = 144) is large enough to assume that the sample proportion is approximately normally distributed with mean

Mean of the sampling distribution of $\hat{p}$

$$\mu_{\hat{p}} = p = 0.10,$$

Standard Error of $\hat{p}$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.10)(0.90)}{144}}$$

$$= \frac{0.3}{12} = 0.025.$$

Therefore, the sampling distribution of is approximately N(0.10, 0.025); and hence

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

$$= \frac{\hat{p} - 0.10}{0.025}$$

is approximately N(0, 1).

We are required to find the probability that the proportion of underweight boxes in the sample is equal to or greater than 5% i.e., we require

$$P(\hat{p} \geq 0.05).$$

In this regard, a very important point to be noted is that, just as we use a *continuity correction* of + ½ whenever we consider the normal approximation to the binomially distributed random variable X, in *this* situation, since

$$\hat{p} = \frac{X}{n},$$

therefore, we need to use the following continuity correction; We need to use a *continuity correction* of $\pm \frac{1}{2n}$ in the case of the sampling distribution of $\hat{p}$.

Applying the continuity correction in this problem, we have:

$$P(\hat{p} \geq 0.05) \Rightarrow P\left(\hat{p} \geq 0.05 - \frac{1}{(2)(144)}\right)$$

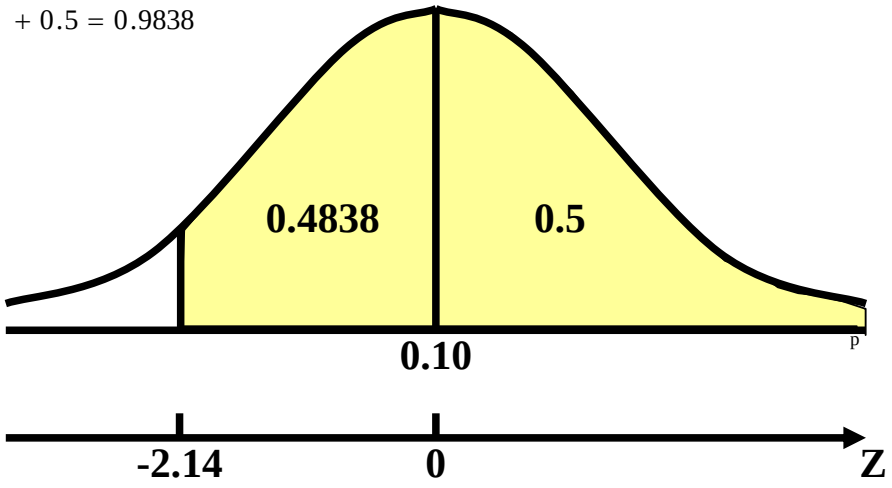
$$= P\left(\hat{p} \geq 0.05 - \frac{1}{288}\right)$$

$$= P\left(\frac{\hat{p} - 0.10}{0.025} \geq \frac{(0.05 - 1/288) - 0.10}{0.025}\right)$$

$$= P(Z \geq -2.14)$$

$$= P(-2.14 \leq Z \leq 0) + P(0 \leq Z \leq \infty)$$

$$= 0.4838 + 0.5 = 0.9838$$



Hence, the probability that at *least* 5% of the sample boxes are under-weight *is as high as* 98% .

The sampling distributions of $\bar{X}$ and $\hat{p}$ pertain to the situation when we are drawing all possible samples of a

particular size from one particular population. Next, we will discuss the case when we are dealing with all possible samples drawn from *two* populations, such that the samples from the two populations are *independent*.
In this regard, we will consider the sampling distributions of $\bar{X}_1 - \bar{X}_2$ and $\hat{p}_1 - \hat{p}_2$:

We begin with the sampling distribution of $\bar{X}_1 - \bar{X}_2$:

SAMPLING DISTRIBUTION OF DIFFERENCES BETWEEN MEANS

Suppose we have two distinct populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. Let *independent* random samples of sizes n_1 and n_2 be selected from the respective populations, and the differences $\bar{x}_1 - \bar{x}_2$ between the means of all possible pairs of samples be computed.
Then, a probability distribution of the differences $\bar{X}_1 - \bar{X}_2$ can be obtained. Such a distribution is called the sampling distribution of the differences of sample means $\bar{X}_1 - \bar{X}_2$. We illustrate the sampling distribution of $\bar{X}_1 - \bar{X}_2$ with the help of the following example.

EXAMPLE

Draw all possible random samples of size n1 = 2 *with replacement* from a finite population consisting of 4, 6, similarly, draw all possible random samples of size n = 2 *with replacement* from another finite population consisting of 1, 2, 3.
a) Find the possible differences between the sample means of the two populations
b) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance
c) Verify that

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}.$$

SOLUTION

Whenever we are sampling with replacement from a finite population, the total number of possible samples is Nn (where N is the population size, and n is the sample size).Hence, in this example, there are (3)2 = 9 possible samples which can be drawn with replacement from each population. These two sets of samples and their means are given below:

From Population 1			From Population 2		
Sample No.	Sample Value	$\bar{x}_1$	Sample No.	Sample Value	$\bar{x}_2$
1	4, 4	4	1	1, 1	1.0
2	4, 6	5	2	1, 2	1.5
3	4, 8	6	3	1, 3	2.0
4	6, 4	5	4	2, 1	1.5
5	6, 6	6	5	2, 2	2.0
6	6, 8	7	6	2, 3	2.5
7	8, 4	6	7	3, 1	2.0
8	8, 6	7	8	3, 2	2.5
9	8, 8	8	9	3, 3	3.0

a) Since there are 9 samples from the first population as well as 9 from the second, hence, there are 81 possible combinations of $\bar{x}1$ and $\bar{x}2$
The 81 possible differences $\bar{x}1 - \bar{x}2$ are presented in the following table:

$\bar{x}_2$	$\bar{x}_2$								
	4	5	6	5	6	7	6	7	8
1.0	3.0	4.0	5.0	4.0	5.0	6.0	5.0	6.0	7.0
1.5	2.5	3.5	4.5	3.5	4.5	5.5	4.5	5.5	6.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
1.5	2.5	3.5	4.5	3.5	4.5	5.5	4.5	5.5	6.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
2.5	1.5	2.5	3.5	2.5	3.5	4.5	3.5	4.5	5.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
2.5	1.0	2.5	3.5	2.5	3.5	4.5	3.5	4.5	5.5
3.0	1.0	2.0	3.0	2.0	3.0	4.0	3.0	4.0	5.0

b)The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is as follows:

$\bar{x}_1 - \bar{x}_2 = d$	Tally	f	Probability $f(\bar{x}_1 - \bar{x}_2) = f(d)$	df (d)	d^2 f(d)
1.0	|	1	1/81	1/81	1.0/81
1.5	||	2	2/81	3/81	4.5/81
2.0	𝍸	5	5/81	10/81	20.0/81
2.5	𝍸 |	6	6/81	15/81	37.5/81
3.0	𝍸 𝍸	10	10/81	30/81	90.0/81
3.5	𝍸 𝍸	10	10/81	35/81	122.5/81
4.0	𝍸 𝍸|||	13	13/81	52/81	208.0/81
4.5	𝍸 𝍸	10	10/81	45/81	202.5/81
5.0	𝍸 𝍸	10	10/81	50/81	250.0/81
5.5	𝍸 |	6	6/81	33/81	181.5/81
6.0	𝍸	5	5/81	30/81	180.0/81
6.5	||	2	2/81	13/81	84.5/81
7.0	|	1	1/81	7/81	49.0/81
Total	---	81	1	324/81	1431/81

Thus the mean and the variance are

$$\mu_{\bar{x}_1-\bar{x}_2} = \sum (\bar{x}_1 - \bar{x}_2)\, f(\bar{x}_1 - \bar{x}_2)$$

$$= \sum df(d) = \frac{324}{81} = 4, \quad \textit{and}$$

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \sum d^2 f(d) - [\sum df(d)]^2$$

$$= \frac{1431}{81} - \left(\frac{324}{81}\right)^2 = \frac{53}{3} - 16 = \frac{5}{3} = 1.67$$

c) In order to verify the properties of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ we first need to compute the mean and variance of the first population:

The mean and standard deviation of the first population are:

$$\mu_1 = \frac{4+6+8}{3} = 6 \text{ , and}$$

$$\sigma_1^2 = \frac{(4-6)^2 + (6-6)^2 + (8-6)^2}{3} = \frac{8}{3}.$$

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{8}{3}.\frac{1}{2} + \frac{2}{3}.\frac{1}{2}$$

$$= \frac{4}{3} + \frac{1}{3} = \frac{5}{3}$$

$$= 1.67$$

$$= \sigma^2_{\bar{x}_1 - \bar{x}_2}$$

The mean and variance of the second population are:

$$\mu_2 = \frac{1+2+3}{3} = 2 \text{ , and}$$

$$\sigma_2^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}.$$

Now $\mu_{\bar{x}_1 - \bar{x}_2} = 4 = 6 - 2 = \mu_1 - \mu_2$, *and*

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{8}{3}.\frac{1}{2} + \frac{2}{3}.\frac{1}{2}$$

$$= \frac{4}{3} + \frac{1}{3} = \frac{5}{3}$$

$$= 1.67$$

$$= \sigma^2_{\bar{x}_1 - \bar{x}_2}$$

Hence, two properties of the sampling distribution of $\overline{X}_1 - \overline{X}_2$ are satisfied. The sampling distribution of the differences $\overline{X}_1 - \overline{X}_2$ has the following properties:

PROPERTIES OF THE SAMPLING DISTRIBUTION OF $\overline{X}_1 - \overline{X}_2$

Property No. 1:

The mean of the sampling distribution of $\overline{X}_1 - \overline{X}_2$, denoted by $\mu_{\overline{X}_1 - \overline{X}_2}$, is equal to the difference between population means, that is

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

Property No. 2:

In case of sampling with or without replacement from two *infinite* populations, the standard deviation of the sampling distribution of $\overline{X}_1 - \overline{X}_2$ (i.e. *standard error* of $\overline{X}_1 - \overline{X}_2$), denoted by $\sigma_{\overline{X}_1 - \overline{X}_2}$, is given by

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The above expression for the Standard Error of $\overline{X}_1 - \overline{X}_2$ also holds for finite population when sampling is performed *with* replacement. In case of sampling without replacement from a finite population, the formula for the standard error of will be suitably modified.

Property No. 3:

Shape of the distribution:

a) If the POPULATIONS are normally distributed, the sampling distribution of $\overline{X}_1 - \overline{X}_2$, regardless of sample sizes, will be *normal* with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

In other words, the variable

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is normally distributed with zero mean and unit variance.

b) If the POPULATIONS are *non-normal* and if *both* sample sizes are *large*, (i.e., greater than or equal to 30), then the sampling distribution of the differences between means is approximately a *normal* distribution by the Central Limit Theorem.

In this case too, the variable

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

will be approximately *normally* distributed with mean zero and variance one.

LECTURE NO. 33

- Sampling Distribution of (continued)
- Point Estimation
- Desirable Qualities of a Good Point Estimator
  - Unbiasedness
  - Consistency

We illustrate the real-life application of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ with the help of the following example:

EXAMPLE

Car batteries produced by company A have a mean life of 4.3 years with a standard deviation of 0.6 years. A similar battery produced by company B has a mean life of 4.0 years and a standard deviation of 0.4 years. What is the probability that a random sample of 49 batteries from company A will have a mean life of at least 0.5 years *more* than the mean life of a sample of 36 batteries from company B?

SOLUTION

We are given the following data:
Population A:
$\mu 1 = 4.3$ years, $\sigma 1 = 0.6$ years,
Sample size: n1 = 49
Population B:
$\mu 2 = 4.0$ years, $\sigma 2 = 0.4$ years,
Sample size: n2 = 36
Both sample sizes (n1 = 49, n2 = 36) are large enough to assume that the sampling distribution of the differences is approximately a normal such that: $\bar{X}_1 - \bar{X}_2$

MEAN

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 4.3 - 4.0 = 0.3 \text{ years}$$

and standard deviation:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{0.36}{49} + \frac{0.16}{36}}$$

Thus the variable $= 0.1086$ *years.*

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 0.3}{0.1086}$$

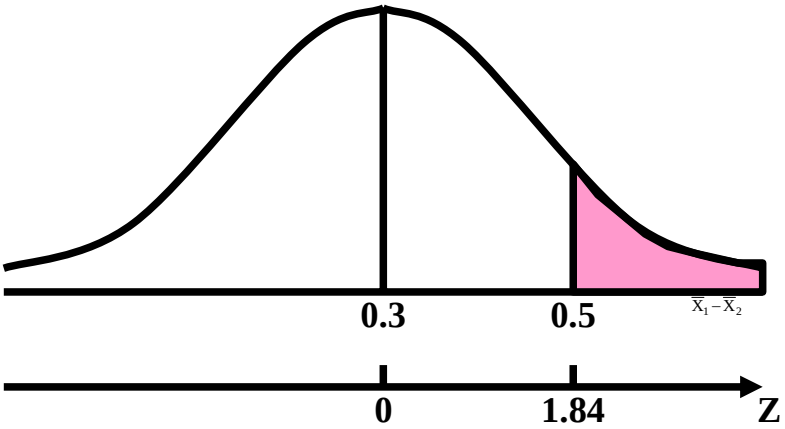
is approximately N (0, 1)
We are required to find the probability that the mean life of 49 batteries produced by company A will have a mean life of at least 0.5 years *longer* than the mean life of 36 batteries produced by company B, i.e.
We are required to find:

$$P\left(\bar{X}_1 - \bar{X}_2 \geq 0.5\right).$$

Transforming $\bar{X}_1 - \bar{X}_2 = 0.5$

to z-value, we find that:

$$z = \frac{0.5 - 0.3}{0.1086} = 1.84$$



Hence, using the table of areas under normal curve, we find:

$$\begin{aligned} P(\bar{X}_1 - \bar{X}_2 \geq 0.5) &= P(Z \geq 1.84) \\ &= 0.5 - P(0 < Z < 1.84) \\ &= 0.5 - 0.4671 \\ &= 0.0329 \end{aligned}$$

In other words, (given that the *real* difference between the mean lifetimes of batteries of company A and batteries of company B is
4.3 - 4.0 = *0.3* years), the probability that a *sample* of 49 batteries produced by company A will have a mean life of at least 0.5 years *longer* than the mean life of a *sample* of 36 batteries produced by company B, is only 3.3%.

SAMPLING DISTRIBUTION OF THE DIFFERENCES BETWEEN PROPORTIONS

Suppose there are two *binomial* populations with proportions of successes p1 and p2 respectively. Let *independent* random samples of sizes n_1 and n_2 be drawn from the respective populations, and the differences $\hat{p}_1 - \hat{p}_2$ between the proportions of *all possible* pairs of samples be computed. Then, a probability distribution of the differences $\hat{p}_1 - \hat{p}_2$ can be obtained. Such a probability distribution is called the *sampling distribution* of the differences between the proportions $\hat{p}_1 - \hat{p}_2$ .We illustrate the sampling distribution of $\hat{p}_1 - \hat{p}_2$ with the help of the following example:

EXAMPLE

It is claimed that 30% of the households in Community A and 20% of the households in Community B have at least one teenager. A simple random sample of 100 households from each community yields the following results:
What is the probability of observing a difference *this* large or *larger* if the claims are true?

$$\hat{p}_A = 0.34,\ \hat{p}_B = 0.13.$$

SOLUTION

We assume that if the claims are true, the sampling distribution of $\hat{p}_A - \hat{p}_B$ is approximately normally distributed (as, in this example, both the sample sizes are large enough for us to apply the normal approximation to the binomial distribution).Since we are reasonably confident that our sampling distribution is approximately normally distributed, hence we will be finding any required probability by computing the relevant areas under our normal curve, and, in order to do so, we will first need to convert our variable $\hat{p}_A - \hat{p}_B$ to Z. In order to convert $\hat{p}_A - \hat{p}_B$ to Z, we need the values of $\mu_{\hat{p}_A - \hat{p}_B}$ as well as $\sigma_{\hat{p}_A - \hat{p}_B}$.
It can be mathematically proved that:

PROPERTIES OF THE SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$

Property No. 1:

The mean of the sampling distribution of $\hat{p}_1 - \hat{p}_2$, denoted by $\mu_{\hat{p}_1 - \hat{p}_2}$, is equal to the difference between the population proportions, that is $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$.

Property No. 2:

The standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$, (i.e. the standard error of $\hat{p}_1 - \hat{p}_2$) denoted by

$\sigma_{\hat{p}_1 - \hat{p}_2}$ is given by

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}},$$

where q = 1 – p
Hence, in this example, we have:

$$\mu_{\hat{p}_A - \hat{p}_B} = 0.30 - 0.20 = 0.10$$

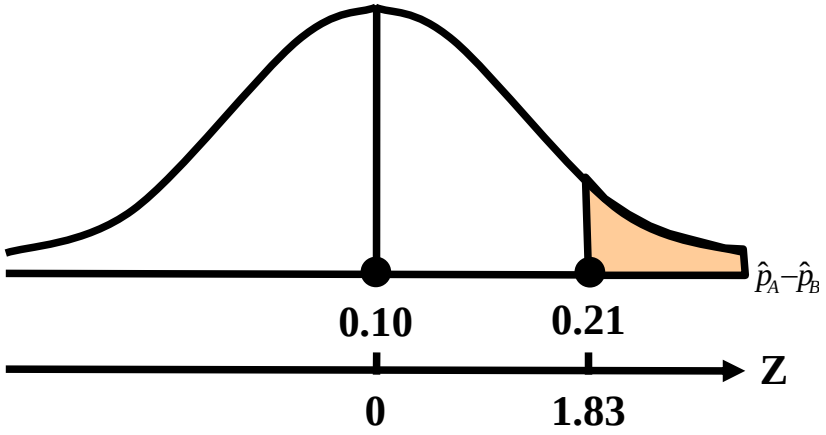
$$\sigma^2_{\hat{p}_A - \hat{p}_B} = \frac{(0.30)(0.70)}{100} + \frac{(0.20)(0.80)}{100} = 0.0037$$

The observed difference in sample proportions is

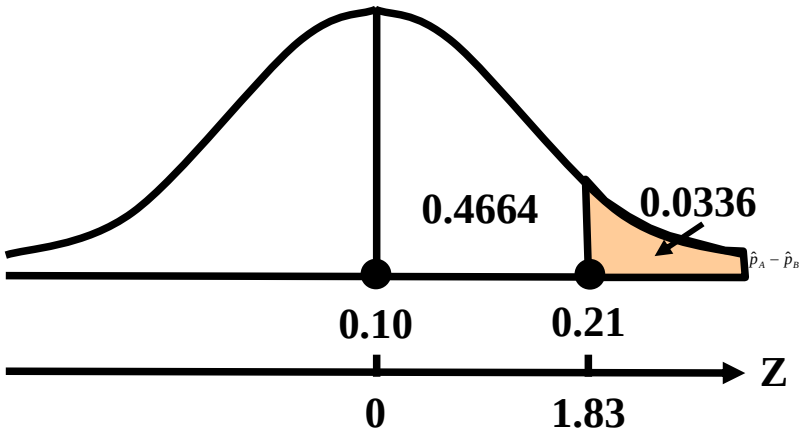
$$\hat{p}_A - \hat{p}_B = 0.34 - 0.13 = 0.21$$

The probability that we wish to determine is represented by the area to the right of 0.21 in the sampling distribution of $\hat{p}_A - \hat{p}_B$ .To find this area, we compute

$$z = \frac{0.21 - 0.10}{\sqrt{0.0037}} = \frac{0.11}{0.06} = 1.83$$



By consulting the Area Table of the standard normal distribution, we find that the area between z = 0 and z = 1.83 is 0.4664. Hence, the area to the right of z = 1.83 is 0.0336.
This probability is shown in following figure:

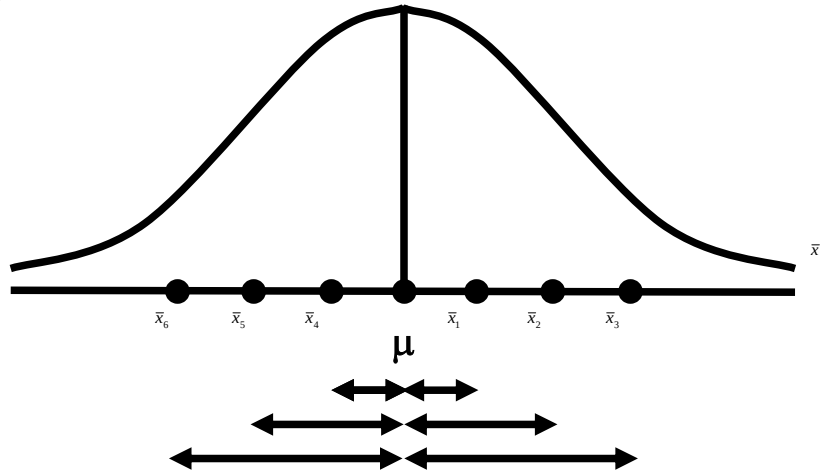


Thus, if the claim is true, the probability of observing a difference as larger as or larger than the actually observed is only 0.0336 i.e. 3.36%. The students are encouraged to try to *interpret* this result with reference to the situation at hand, as, in attempting to solve a statistical problem, it is very important not just to apply various formulae and obtain numerical results, but to *interpret* the results with reference to the problem under consideration. Does the result indicate that at least *one* of the two claims is untrue, or does it imply something *else*? Before we close the basic discussion regarding sampling distributions, we would like to draw the students' attention to the following two important points:

- We have discussed various sampling distributions with reference to the simplest technique of random sampling, i.e. *simple random sampling*.
- And, with reference to simple random sampling, it should be kept in mind that this technique of sampling is appropriate in that situation when the population is *homogeneous*.
- Let us consider the reason why the standard deviation of the sampling distribution of any statistic is known as its standard *error*:

To answer this question, consider the fact that any statistic, considered as an *estimate* of the corresponding population parameter, should be as *close* in magnitude to the parameter as possible. The difference between the value of the statistic and the value of the parameter can be regarded as an *error* --- and is called 'sampling error'. Geometrically, each one of these errors can be represented by *horizontal line segment* below the X-axis, as shown below

Sampling Distribution of



The above diagram clearly indicates that there are various magnitudes of this error, depending on how far or how close the values of our statistic are in different samples.

The standard deviation of $\bar{X}$ gives us a '*standard*' value of this error, and hence the term '*Standard* Error'.

Having presented the basic ideas regarding sampling distributions, we now begin the discussion regarding POINT ESTIMATION:

POINT ESTIMATION

Point estimation of a population parameter provides as an estimate a *single* value calculated from the sample that is likely to be close in magnitude to the unknown parameter.

DIFFERENCE BETWEEN 'ESTIMATE' AND 'ESTIMATOR'

An *estimate* is a numerical value of the unknown parameter obtained by applying a rule or a formula, called an *estimator*, to a sample X1, X2, …, Xn of size n, taken from a population. In other words, an estimator stands for the *rule or method* that is used to estimate a parameter whereas an estimate stands for the *numerical value* obtained by substituting the sample observations in the rule or the formula.

For instance:

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population with mean μ, then $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of μ, and $\bar{x}$, the *numerical value* of $\bar{X}$, is an estimate of μ (i.e. a point estimate of μ).

In general, the (the Greek letter θ) is customarily used to denote an unknown parameter that could be a mean, median, proportion or standard deviation, while an estimator of θ is commonly denoted by $\hat{\theta}$, or sometimes by T.

It is important to note that *an estimator* is always a *statistic* which is a *function* of the sample observations and hence *is a random variable* as the sample observations are likely to vary from sample to sample.

In other words:

In *repeated* sampling, an estimator is a *random variable*, and has a *probability distribution*, which is known as its *sampling distribution*. Having presented the basic definition of a point estimator, we now consider some *desirable qualities* of a good point estimator. In this regard, the point to be understood is that a point estimator is considered a good estimator if it satisfies *various criteria*. Three of these criteria are:

DESIRABLE QUALITIES OF A GOOD POINT ESTIMATOR

- unbiasedness
- consistency
- efficiency

UNBIASEDNESS

An estimator is defined to be unbiased if the statistic used as an estimator has its expected value equal to the true value of the population parameter being estimated. In other words, let $\hat{\theta}$ be an estimator of a parameterθ. Then $\hat{\theta}$ will be called an unbiased estimator if $E(\hat{\theta}) = \theta.$ If $E(\hat{\theta}) \neq \theta,$ the statistic is said to be a biased estimator

EXAMPLE

Let us consider the sample mean $\bar{X}$ as an estimator of the population meanμ. Then we have $\theta = \mu$

and $\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$

Now, we know that $E(\overline{X}) = \mu$

i.e. $E(\hat{\theta}) = \theta.$

Hence, $\overline{X}$ is an *unbiased* estimator ofμ. Let us illustrate the concept of unbiasedness by considering the example of the annual Ministry of Transport test that was presented in the last lecture:

EXAMPLE

Let us examine the case of an annual Ministry of Transport test to which all cars, irrespective of age, have to be submitted. The test looks for faulty breaks, steering, lights and suspension, and it is discovered after the first year that approximately the *same number* of cars have 0, 1, 2, 3, or 4 faults. The above situation is equivalent to the following: If we let X denote *the number of faults* in a car, then X can take the values 0, 1, 2, 3, and 4, and the *probability* of each of these X values is 1/5. Hence, we have the following probability distribution:

No. of Faulty Items (X)	Probability f(x)
0	1/5
1	1/5
2	1/5
3	1/5
4	1/5
Total	1

MEAN OF THE POPULATION DISTRIBUTION

$$\mu = E(X) = \sum xf(x) = 2$$

We are interested in considering the results that would be obtained if a *sample* of only two cars is tested.
You will recall that we obtained 52 = 25 different possible samples, and, computing the mean of each possible sample, we obtained the following sampling distribution of $\bar{X}$:

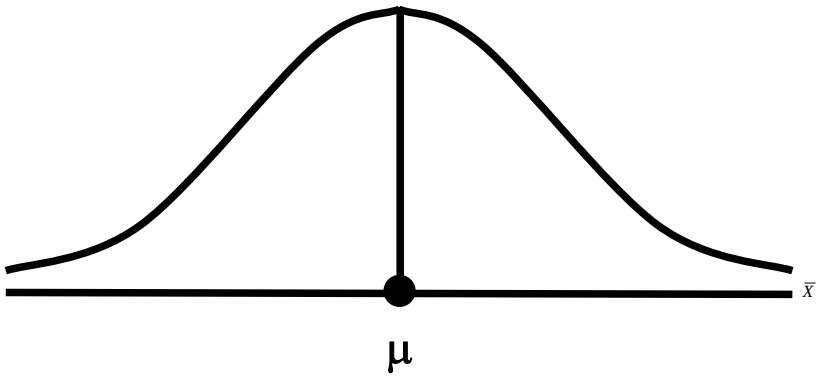
Sample Mean	Probability
$\bar{x}$	P($\bar{X}$ = $\bar{x}$)
0.0	1/25
0.5	2/25
1.0	3/25
1.5	4/25
2.0	5/25
2.5	4/25
3.0	3/25
3.5	2/25
4.0	1/25
Total	25/25=1

We computed the mean of this sampling distribution, and found that the mean of the sample means i.e. comes out to be equal to 2 --- exactly the same as the mean of the population. We find that:

$$\mu_{\bar{x}} = \sum \bar{x}\ f(\bar{x}) = \frac{50}{25} = 2 = \mu$$

i.e the mean of the sampling distribution of $\bar{X}$ is equal to the population mean. By virtue of this property, we say that the sample mean is an *UNBIASED* estimate of the population mean. It should be noted that this property, *always* holds $\mu_{\bar{x}} = \mu,$ *regardless* of the sample size. Unbiasedness is a property that requires that the probability distribution of $\hat{\theta}$ be necessarily *centered* at the parameterθ, irrespective of the value of n.

VISUAL REPRESENTATION OF THE CONCEPT OF UNBIASEDNESS



$E(\bar{X}) = \mu$ implies that the distribution of $\bar{X}$ is *centered* at μ.What this means is that, although many of the individual sample means are either *under-estimates* or *over-estimates* of the true population mean, in the long run, the over-estimates *balance* the under-estimates so that the *mean value* of the sample means comes out to be equal to the population mean.

Let us now consider some other estimators which possess the desirable property of being unbiased: The sample median is also an unbiased estimator of μ when the population is normally distributed (i.e. If X is normally distributed, then Also, as far as p, the *proportion of successes* in the sample is concerned, we have considering the binomial random variable X (which denotes the number of successes in n trials), we have:

$$E(\tilde{X}) = \mu.)$$

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X)$$

$$= \frac{np}{n} = p$$

Hence, the sample proportion is an *unbiased* estimator of the population parameter p. But As far as the sample variance S2 is concerned; it can be mathematically proved that E(S2) ≠ σ2.Hence, the sample variance S2 is a *biased* estimator of σ2.For any population parameter θ and its estimator $\hat{\theta}$, the quantity $E(\hat{\theta}) - \theta$ is known as the amount of *bias.* This quantity is positive if $E(\hat{\theta}) > \theta,$ and is negative if $E(\hat{\theta}) < \theta,$ and, hence, the estimator is said to be positively biased when $E(\hat{\theta}) > \theta$ and negatively biased when $E(\hat{\theta}) < \theta$.Since unbiasedness is a desirable quality, we would like the sample variance to be an *unbiased* estimator of σ2.In order to achieve this end, the formula of the sample variance is *modified* as follows:

Modified formula for the sample variance:

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

Since E(s2) = σ2, hence s2 is an unbiased estimator of σ2.Why is unbiasedness consider a *desirable* property of an estimator? In order to obtain an answer to this question, consider the following: With reference to the estimation of the population mean μ, we note that, in an *actual* study, the probability is very high that the mean of our sample i.e. $\bar{X}$ will either be less than μ or more than μ.

Hence, in an actual study, we can never guarantee that our $\bar{X}$ will coincide with μ.

Unbiasedness implies that, although in an actual study, we *cannot* guarantee that our sample mean will coincide with μ, our estimation *procedure* (i.e. formula) is such that, in *repeated* sampling, the *average* value of our statistic *will* be equal to μ.

The next desirable quality of a good point estimator is *consistency*:

CONSISTENCY

An estimator $\hat{\theta}$ is said to be a consistent estimator of the parameter θ if, for any arbitrarily small positive quantity e,

$$\lim_{n \to \infty} P\left[\left|\hat{\theta} - \theta\right| \le e\right] = 1.$$

In other words, an estimator $\hat{\theta}$ is called a consistent estimator of θ if the probability that $\hat{\theta}$ is very close to θ, approaches unity with an increase in the sample size. It should be noted that consistency is a *large sample* property. Another point to be noted is that a consistent estimator *may or may not* be unbiased.

The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is an unbiased estimator of μ, is a consistent estimator of the mean μ.The sample proportion $\hat{p}$ is also a consistent estimator of the parameter p of a population that has a binomial distribution. The median is *not* a consistent estimator of μ when the population has a skewed distribution. The sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2,$$

though a *biased* estimator, is a consistent estimator of the population variance σ2. Generally speaking, it can be proved that a statistic whose STANDARD ERROR *decreases* with an *increase* in the sample size, will be consistent.

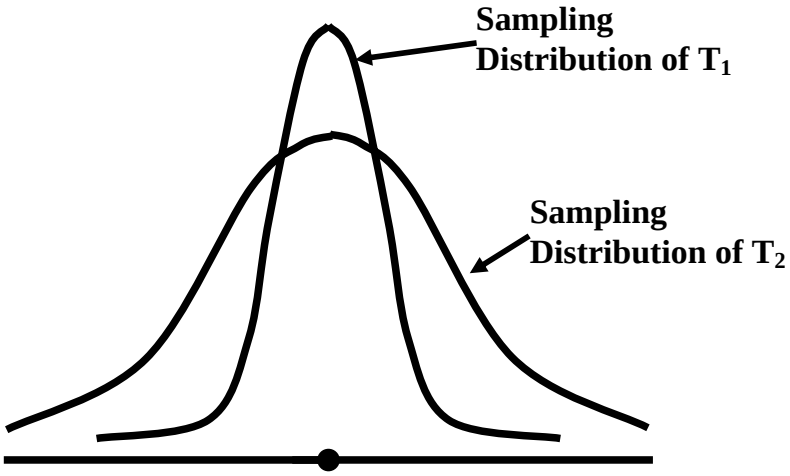
LECTURE NO. 34

- Desirable Qualities of a Good Point Estimator:
  - Efficiency
- Methods of Point Estimation:
  - The Method of Moments
  - The Method of Least Squares
  - The Method of Maximum Likelihood
- Interval Estimation:
  - Confidence Interval for μ

As a sample is only a *part* of the population, it is obvious that the *larger* the sample size, the *more* representative we expect it to be of the population from which it has been drawn. In agreement with the above argument, we will expect our estimator to be close to the corresponding parameter if the sample size is *large*. Hence, we will naturally be happy if the probability of our estimator being close to the parameter *increases* with an increase in the sample size. As *such*, consistency is a desirable property.
Another important desirable quality of a good point estimator is *EFFICIENCY*:

EFFICIENCY

An unbiased estimator is defined to be *efficient* if the variance of its sampling distribution is *smaller than* that of the sampling distribution of *any* other unbiased estimator of the same parameter. In other words, suppose that there are two unbiased estimators T_1 and T_2 of the same parameterθ.Then, the estimator T_1 will be said to be *more efficient* than T_2 if $\text{Var}(T_1) < \text{Var}(T_2)$.
In the following diagram, since Var (T1) < Var (T2),
hence T1 is *more efficient* than T2 :



The *relative efficiency* of T_1 compared to T_2 (where both T_1 and T_2 are unbiased estimators) is given by the ratio

$$E_f = \frac{\text{Var}(T_2)}{\text{Var}(T_1)}.$$

And, if we multiply the above expression by 100, we obtain the relative efficiency in *percentage* form. It thus provides a criterion for *comparing* different unbiased estimators of a parameter. Both the sample mean and the sample median for a population that has a *normal* distribution, are unbiased and consistent estimators of μ but the variance of the sampling distribution of sample means is *smaller than* the variance of the sampling distribution of sample medians. Hence, the sample mean is more *efficient* than the sample median as an estimator of μ.The sample mean may therefore be *preferred* as an estimator.
Next, we consider various *methods of point estimation*. A point estimator of a parameter can be obtained by several methods. We shall be presenting a brief account of the following three methods:

METHODS OF POINTESTIMATION

- The Method of Moments
- The Method of Least Squares
- The Method of Maximum Likelihood

These methods give estimates which may *differ* as the methods are based on different *theories* of estimation.

THE METHOD OF MOMENTS

The method of moments which is due to Karl Pearson (1857-1936), consists of calculating a few *moments* of the sample values and *equating* them to the corresponding moments of a population, thus getting *as many* equations as are needed to solve for the unknown parameters. The procedure is described below:
Let X1, X2, …, Xn be a random sample of size n from a population. Then the rth sample moment about zero is

$$m'_r = \frac{\sum X_i^r}{n},\ r = 1,2,...$$

and the corresponding rth population moment is μ'_r. We then *match* these moments and get *as many* equations as we need to solve for the unknown parameters. The following examples illustrate the method:

EXAMPLE-1

Let X be uniformly distributed on the interval (0, θ). Find an estimator of θ by the method of moments.

SOLUTION

The probability density function of the given uniform distribution is

$$f(x) = \frac{1}{\theta},\ \ 0 \le x \le \theta$$

Since the uniform distribution has only one parameter, (i.e. θ), therefore, in order to find the maximum likelihood estimator of θ by the method of moments, we need to consider only *one* equation.
The first sample moment about zero is

$$m'_1 = \frac{\sum X_i}{n}.$$

And, the first population moment about zero is

$$\mu'_1 = \int_0^\theta x.f(x)dx = \int_0^\theta x.\frac{1}{\theta}\,dx = \frac{1}{\theta}\left[\frac{x^2}{2}\right]_0^\theta = \frac{\theta}{2}$$

Matching these moments, we obtain:

$$\frac{\sum X_i}{n} = \frac{\theta}{2} \text{ or } \theta = 2\overline{X}.$$

Hence, the moment estimator of θ is equal to $2\overline{X}$
i.e. $\hat{\theta} = 2\overline{X}.$

In other words, the moment estimator of θ is just twice the sample mean. It should be noted that, for the above uniform distribution, the mean is given by

$$\mu = \frac{\theta}{2}.$$

(This is so due to the *absolute* symmetry of the uniform distribution around the value $\frac{\theta}{2}$.)

Now, $\mu = \frac{\theta}{2}$ implies that $\theta = 2\mu$.

In other words, if we wish to have the *exact* value of θ, all we need to do is to multiply the population mean μ by 2. Generally, it is not possible to determineμ, and all we can do is to draw a sample from the probability distribution, and compute the *sample* mean $\overline{X}$. Hence, naturally, the equation will be replaced by the equation
(As $2\overline{x}$ provides an *estimate* of θ, hence a 'hat' is placed on top of θ.)
It is interesting to note hat is *exactly* the same quantity as what we obtained as an estimate of θ by the method of moments!(The result obtained by the method of moments *coincides* with what we obtain through simple logic $2\overline{x}$

EXAMPLE-2

Let X1, X2… Xn be a random sample of size n from a normal population with parameters μ and σ2. Find these parameters by the method of moments.

SOLUTION

Here we need two equations as there are *two* unknown parameters, μ and σ2. The first two sample moments about zero are

$$m'_1 = \frac{1}{n}\sum X_i = \overline{X} \text{ and } m'_2 = \frac{1}{n}\sum X_i^2.$$

The *corresponding* two moments of a normal distribution are

′μ1 = μ and ′μ2 = σ2 + μ2.

(σ2 = ′μ2 – ′μ12 = ′μ2 – μ2)

To get the desired estimators by the method of moments, we *match* them.
Thus, we have :

$$\mu = \frac{1}{n}\sum X_i \text{ and } \sigma^2 + \mu^2 = \frac{1}{n}\sum X_i^2$$

Solving the above equations *simultaneously*, we obtain:

$$\hat{\mu} = \frac{1}{n}\sum X_i = \overline{X}, \text{ and}$$

$$\hat{\sigma}^2 = \frac{\sum X_i^2}{n} - \overline{X}^2 = \frac{1}{n}\sum \left(X_i - \overline{X}\right)^2 = S^2.$$

as the moment estimators for μ and σ^2. A shortcoming of this method is that the moment estimators are, in general, *inefficient*.

THE METHOD OF LEAST SQUARES

The method of Least Squares, which is due to *Gauss* (1777-1855) and *Markov* (1856-1922), is based on the theory of *linear* estimation. It is regarded as one of the *important* methods of point estimation. An estimator found by *minimizing* the sum of squared deviations of the sample values from some *function* that has been hypothesized as a *fit* for the data, is called the least squares estimator. The method of least-squares has already been discussed in connection with regression analysis that was presented in Lecture No. 15.

You will recall that, when fitting a straight line y = a+bx to real data, 'a' and 'b' were determined by *minimizing* the sum of squared deviations between the fitted line and the data-points.

The y-intercept and the slope of the fitted line i.e. 'a' and 'b' are *least-square estimates* (respectively) of the y-intercept and the slope of the *TRUE* line that would have been obtained by considering the entire population of data-points, and not just a sample.

METHOD OF MAXIMUM LIKELIHOOD

The method of maximum likelihood is regarded as the *MOST important* method of estimation, and is the *most* widely used method. This method was introduced in 1922 by Sir Ronald A. Fisher (1890-1962).The mathematical technique of finding Maximum Likelihood Estimators is a bit *advanced*, and involves the concept of the Likelihood Function.

RATIONALE OF THE METHOD OF MAXIMUM LIKELIHOOD (ML)

"To consider *every* possible value that the parameter might have, and for *each* value, compute the *probability* that the given sample would have occurred if that *were* the true value of the parameter. That value of the parameter for which the probability of a given sample is *greatest*, is chosen as an estimate." An estimate obtained by this method is called the maximum likelihood estimate (MLE). It should be noted that the method of maximum likelihood is applicable to *both* discrete and continuous random variables.

EXAMPLES OF MLE's IN CASE OF DISCRETE DISTRIBUTIONS

Example-1:
For the Poisson distribution given by

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \; x = 0,1,2,.....,$$

the MLE of μ is $\bar{X}$ (the sample mean).

EXAMPLE-2

For the geometric distribution given by the MLE of p is Hence, the MLE of p is equal to the reciprocal of the mean.

EXAMPLE-3

For the Bernoulli distribution given by

$$P(X = x) = p^x q^{1-x}, \; x = 0,1 \;,$$

the MLE of p is (the sample mean).

EXAMPLES OF MLE's IN CASE OF *CONTINUOUS* DISTRIBUTIONS

Example-1

For the exponential distribution given by

$$f(x) = \theta e^{-\theta x}, \;\; x > 0, \;\; \theta > 0 \;\;,$$

the MLE of θ is (the reciprocal of the sample mean $\frac{1}{\bar{X}}$.)

EXAMPLE-2

For the *normal* distribution with parameters μ and σ2, the *joint* ML estimators of μ and σ2 is the sample mean and the sample variance S2*(which is not an unbiased estimator of σ2).*As indicated many times earlier, the *normal distribution* is encountered frequently in practice, and, in this regard, it is both interesting and important to note that, in the case of this *frequently* encountered distribution, the *simplest* formulae (i.e. the sample *mean* and the sample *variance*) fulfill the criteria of the relatively *advanced* method of maximum likelihood estimation !The last example among the five presented above (the one on the *normal* distribution) points to *another* important fact --- and that is : The Maximum Likelihood Estimators are consistent and efficient but *not necessarily unbiased*. (As we know, S2 is *not* an unbiased estimator of σ2.)

EXAMPLE

It is well-known that human weight is an approximately normally distributed variable. Suppose that we are interested in estimating the mean and the variance of the weights of adult males in one particular province of a country. A random sample of 15 adult males from this particular population yields the following weights (in pounds):

131.5	136.9	133.8	130.1	133.9
135.2	129.6	134.4	130.5	134.2
131.6	136.7	135.8	134.5	132.7

Find the maximum likelihood estimates for θ1 = μ and θ2 = σ2.

SOLUTION

The above data is that of a random sample of size 15 from N(μ, σ2). It has been mathematically proved that the joint maximum likelihood estimators of μ and σ2 are $\bar{X}$ and S2. We compute these quantities for this particular sample, and obtain $\bar{X}$ = 133.43, and S2 = 5.10 .These are the Maximum Likelihood Estimates of the mean and variance of the population of weights in this particular example. Having discussed the concept of point estimation in some detail, we now begin the discussion of the concept of *interval estimation*:

As stated earlier, whenever a *single* quantity computed from the sample acts as an estimate of a population parameter, we call that quantity a *point* estimate e.g. the sample mean is a point estimate of the population mean μ.

The *limitation* of point estimation is that we have no way of ascertaining how close our point estimate is to the true value (the parameter).

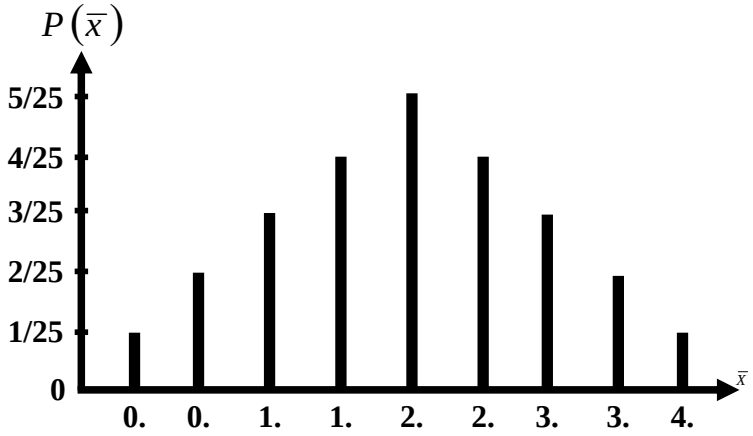
For example, we know that is an unbiased estimator of μ i.e. if we had taken all possible samples of a particular size from the population and calculated the mean of each sample, then the mean of the sample means would have been equal to the population mean (μ), but in an actual survey we will be selecting only one sample from the population and will calculate its mean .

We will have no way of ascertaining how close this particular is toμ. Whereas a point estimate is a single value that acts as an estimate of the population parameter, *interval estimation* is a procedure of estimating the unknown parameter which specifies a *range* of values within which the parameter is expected to lie. A *confidence interval* is an interval computed from the sample observations x1, x2….xn, with a statement of how *confident* we are that the interval *does* contain the population parameter.

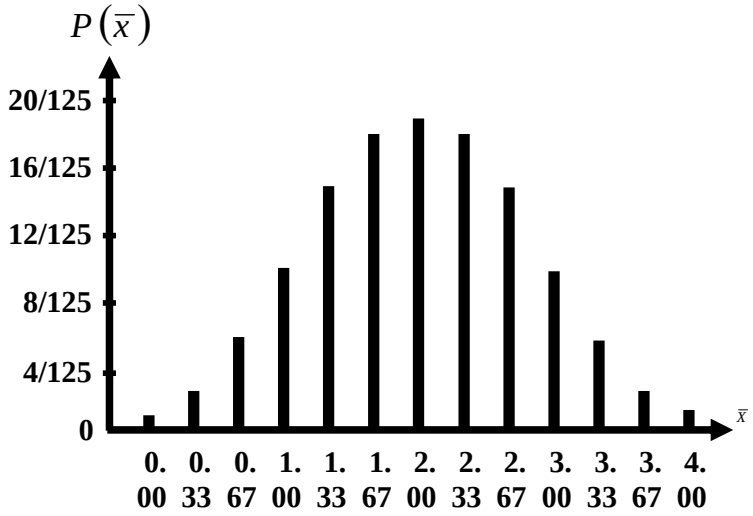
We develop the concept of interval estimation with the help of the example of the Ministry of Transport test to which all cars, irrespective of age, have to be submitted.

EXAMPLE

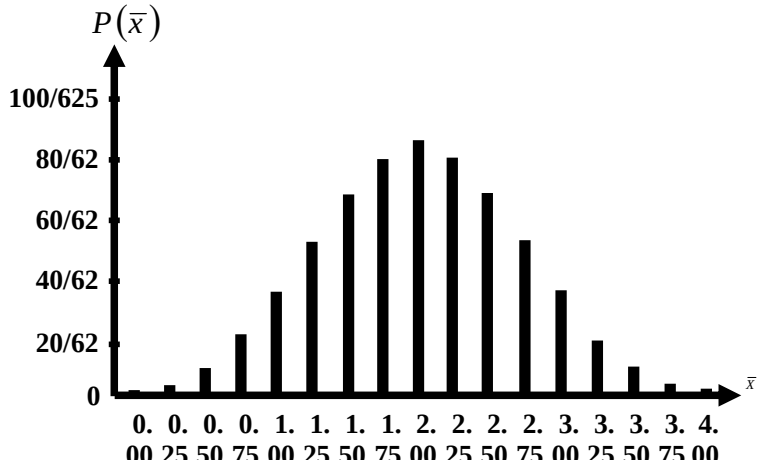
Let us examine the case of an annual Ministry of Transport test to which all cars, irrespective of age, have to be submitted. The test looks for faulty breaks, steering, lights and suspension, and it is discovered after the first year that approximately the *same number* of cars has 0, 1, 2, 3, or 4 faults. You will recall that when we drew all possible samples of size 2 from this uniformly distributed population, the sampling distribution of $\bar{X}$ was *triangular*:
Sampling Distribution of $\bar{X}$ for n = 2



But when we considered what happened to the shape of the sampling distribution with if the sample size is *increased*, we found that it was somewhat like a normal distribution:
Sampling Distribution of $\bar{X}$ for n = 3



And, when we increased the sample size to 4, the sampling distribution resembled a normal distribution even *more* closely, Sampling Distribution of $\bar{X}$ for n = 4



It is clear from the above discussion that as *larger* samples are taken, the shape of the sampling distribution of $\bar{X}$ undergoes *discernible changes*.

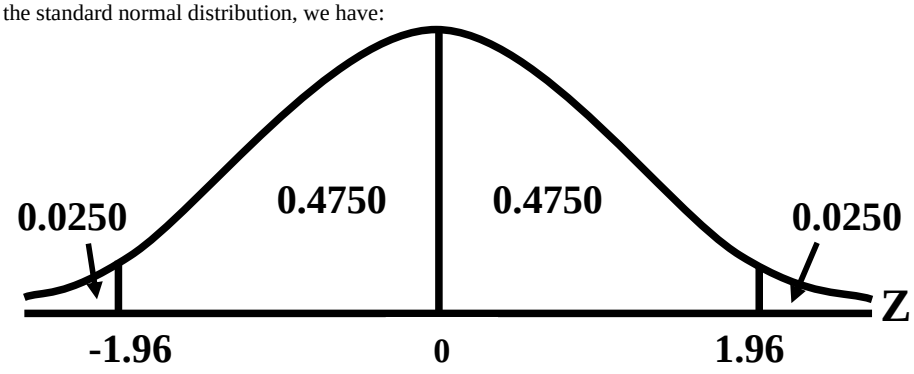
In all three cases the line charts are symmetrical, but as the sample size increases, the overall configuration changed from a triangular distribution to a *bell-shaped* distribution. In other words, for large samples, we are dealing with a *normal* sampling distribution of $\bar{x}$. In other words: When sampling from an infinite population such that the sample size n is large, $\bar{X}$ is *normally distributed* with mean μ and variance $\frac{\sigma^2}{n}$

i.e. $\bar{X}$ is $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Hence, the standardized version of $\bar{X}$ i.e.

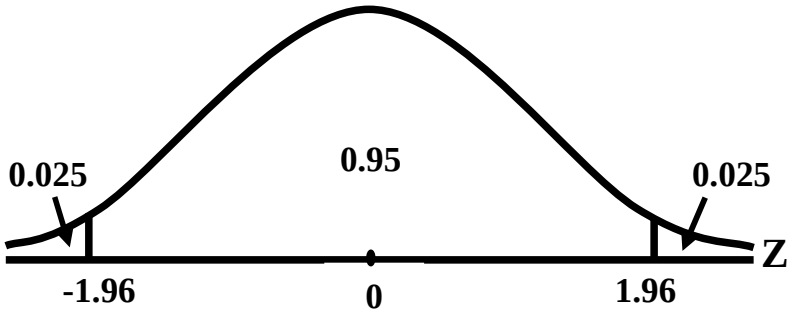
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is normally distributed with mean 0 and variance 1 i.e. Z is N(0, 1). Now, for the standard normal distribution, we have: For the standard normal distribution, we have:



The above is equivalent to P(-1.96 < Z < 1.96)

= 0.4750 + 0.4750 = 0.95



In other words:

$$P\left(-1.96 \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

The above can be re-written as:

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Or

$$P\left(-\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

or

$$P\left(\overline{X}+1.96\frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{X}-1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

or

$$P\left(\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

The above equation yields the 95% confidence interval for μ :
The 95% confidence interval for μ is

$$\left(\overline{X}-1.96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right).$$

In other words, the 95% C.I. for μ is given by

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

In a *real-life* situation, the population standard deviation is usually *not known* and hence it has to be *estimated*.
It can be mathematically proved that the quantity

$$s^2 = \frac{\sum\left(X-\overline{X}\right)^2}{n-1}$$

is an unbiased estimator of σ2 (the population variance). (just as the sample mean is an unbiased estimator of μ).
In this situation, the 95% Confidence Interval for μ is given by:

The points $$P\left(\overline{X}-1.96\frac{s}{\sqrt{n}} < \mu < \overline{X}+1.96\frac{s}{\sqrt{n}}\right) = 95\%$$

$$\overline{X}-1.96\frac{s}{\sqrt{n}} \text{ and } \overline{X}+1.96\frac{s}{\sqrt{n}}$$

are called the lower and upper *limits* of the 95% confidence interval.

LECTURE NO. 35

- Confidence Interval for μ (continued).
- Confidence Interval for μ1-μ2.

In the last lecture, we discussed the construction of the 95% confidence interval regarding the mean of a population i.e. μ.

EXAMPLE-1

Consider a car assembly plant employing something over 25,000 men. In planning its future labour requirements, the management wants an estimate of the number of days lost per man each year due to illness or absenteeism.
A random sample of 500 employment records shows the following situation:

Number of Days Lost	Number of Employees
None	48
1 or 2	43
3 or 4	90
5 or 6	186
7 or 8	78
9 to 12	34
13 to 20	21
Total	500

Construct a 95% confidence interval for the mean number of days lost per man each year due to illness or absenteeism.

SOLUTION

- The point estimate of μ is $\bar{X}$, which in this example comes out to be $\bar{X}$ = 5.38 days
- In order to construct a confidence interval for μ, we need to compute s, which in this example comes out to be s = 3.53 days.

Hence, the 95% confidence interval for μ comes out to be

$$\left(5.38-\frac{1.96\times 3.53}{\sqrt{500}}\,,\,5.38+\frac{1.96\times 3.53}{\sqrt{500}}\right)$$

or 5.38 ± 0.31 days
= 5.07 days to 5.69 days.

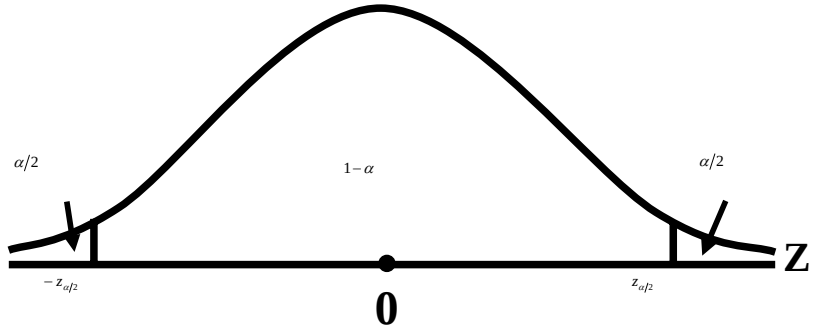
In other words, we can say that the mean number of days lost per man each year due to illness or absenteeism lies somewhere between 5.07 days and 5.69 days, and this statement is being made on the basis of 95% confidence. A very important point to be noted here is that we should be very careful regarding the interpretation of confidence intervals
When we set $1 - \alpha = 0.95$, it means that the probability is 95% that the interval

$$\text{from } \bar{X}-1.96\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X}+1.96\frac{\sigma}{\sqrt{n}}$$

will actually contain the true population mean μ.In other words, if we construct a large number of intervals of this type, corresponding to the large number of samples that we can draw from any particular population, then out of every 100 such intervals, 95 will contain the true population mean μ whereas 5 will not.
The above statement pertains to the overall situation in repeated sampling --- once a sample has actually been chosen from a population, $\bar{X}$ computed and the interval constructed, then this interval either contains μ, or does not contain μ.
So, probability that our interval corresponding to sample values have actually occurred, is either one (i.e. cent per cent), or zero. The statement 95% probability is valid before any sample has actually materialized. In other words, we can say that our procedure of interval estimation is such that, in repeated sampling, 95% of the intervals will containμ. The above example pertained to the 95% confidence interval for μ.In general; the lower and upper limits of the confidence interval for μ are given by

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Where the value of $z\alpha/2$ depends on how much confidence we want to have in our interval estimate.



The above situation leads to the (1-α) 100% C.I. for μ . If (1-α) = 0.95, then zα/2 = 1.96 whereas , If (1-α) = 0.99, then zα/2 = 2.58 and If (1-α) = 0.90, then zα/2 = 1.645 .
(The above values of zα/2 are easily obtained from the area table of the standard normal distribution).An important to note is that, as indicated earlier, the above formula for the conference interval is valid when we are sampling from an infinite population in such a way that the sample size n is large. How large should n be in a practical situation?
The rule of thumb in this regard is that whenever n ≥ 30, we can use the above formula.

CONFIDENCE INTERVAL FOR μ, THE MEAN OF AN INFINITE POPULATION

For large n (n ≥ 30), the confidence interval is given by

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $\bar{x} = \frac{\sum x}{n}$ is the sample mean

and

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

is the sample standard deviation.

EXAMPLE-1

The Punjab Highway Department is studying the traffic pattern on the G.T. Road near Lahore. As part of the study, the department needs to estimate the average number of vehicles that pass the Ravi Bridge each day. A random sample of 64 days gives X = 5410 and s = 680. Find the 90 per cent confidence interval estimate for μ, the average number of vehicles per day.

SOLUTION

The 90% confidence interval for μ is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}},$$

where

$\bar{x}$ = 5410,

s = 680, n = 64 and z0.05 = 1.645.

Substituting these values, we obtain

$$5410 \pm (1.645)\left(\frac{680}{\sqrt{64}}\right)$$

or 5410 ± (1.645) (85)
or 5410 ± 139.8
or 5270.2 to 5549.8
or, rounding the above two figures correct to the nearest whole number, we have
5270 to 5550

Hence, we can say that the average number of vehicles that pass the Ravi bridge each day lies somewhere between 5270 and 5550, and this statement is being made on the basis of 90% confidence.

EXAMPLE-2

Suppose a car rental firm wants to estimate the average number of miles traveled per day by each of its cars rented in one particular city. A random sample of 110 cars rented in this particular city reveals that the mean travel distance per day is 85.5 miles, with a standard deviation of 19.3 miles.
Compute a 99% confidence interval to estimateμ.

SOLUTION

Here, n = 110, $\bar{X}$ = 85.5, and S = 19.3. For a 99% level of confidence, a z-value of 2.575 is obtained.

$$\bar{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

$$85.5 - 2.575\frac{19.3}{\sqrt{110}} \le \mu \le 85.5 + 2.575$$

$$85.5 - 4.7 \le \mu \le 85.5 + 4.7$$

$$80.8 \le \mu \le 90.2$$

The point estimate indicates that the average number of miles traveled per day by a rental car in this particular city is 85.5. With 99% confidence, we estimate that the population mean is somewhere between 80.8 and 90.2 miles per day.
Next, we consider a very interesting and important way of interpreting a confidence interval. An Important Way of Interpreting a Confidence Interval, Because of the fact that

$\sigma_{\bar{x}}$ is equal to $\frac{\sigma}{\sqrt{n}}$,

Hence, $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is equal to

$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$

(where $\sigma_{\bar{x}}$ represents the standard error of $\bar{X}$,Hence The C.I. for μ can be defined as X ± a certain number of standard errors of $\bar{X}$. efining a Confidence Interval as:
"A point estimate plus/minus a few times the standard error of that estimate", The question arises: "How many times?"
The answer is:That depends on the level of confidence that we wish to have. In the case of 99% confidence, zα/2 ~ 2.5, (so that, in this case, we can say that our confidence interval is

$$\bar{x} \pm 2\tfrac{1}{2}\sigma_{\bar{x}});$$

Similarly,
in the case of 95% confidence, zα/2 ~ 2, (so that, in this case, we can say that our confidence interval is and so on.

$$\bar{x} \pm 2\sigma_{\bar{x}});$$

Another important point to be noted is that:
It is a matter of common sense that, in any situation, the narrower our confidence interval, the better.
(Ideally, the width of a confidence interval should be zero --- i.e. we should simply have a point estimate.)
It would be quite unwise to say: "I am 99.999% confident that the mean height of the adult males of this particular city lies somewhere between 4 feet and 12 feet." _!
The important question is: How do we achieve a narrow confidence interval with a high level of confidence?
To answer this question, we should have a closer look at the expression of the confidence interval:

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$

This expression shows clearly that if the quantity $z_{\alpha/2}\sigma_{\bar{x}}$ is small, we will achieve a narrow confidence interval.

This quantity will be small if either $\sigma_{\bar{x}}$ is small or $z_{\alpha/2}$ is small.
Now,

$\sigma_{\bar{x}}$ is equal to $\frac{\sigma}{\sqrt{n}}$,

and hence $\sigma_{\bar{x}}$ will be small if the sample size n is large.

On the other hand, $z_{\alpha/2}$ will be small if the level of confidence 1-α is relatively low. As far as the first point that of n being small is concerned, it should be noted that, in many real-life situations, due to practical constraints, we cannot increase the sample size beyond a certain limit.(We may not have the resources to be able to draw a relatively large sample --- our budget may be limited, the time-period at our disposal may be short, etc. As far as the second point, that of fixing a relatively low level of confidence, is concerned, this is in our own hands, and we can fix our level of confidence as low as we wish --- but, obviously, it will not make much sense to say; "I have estimated that the mean height of adult males of this particular city lies somewhere between 5 feet, 6 inches and 5 feet, 7 inches, and I am saying this with 20% confidence." _!
The gist of the above discussion is that, in any real-life situation, given a particular sample size, we need to strike a compromise between how low a level of confidence can we tolerate, or how wide an interval can we tolerate.
Next, we consider the confidence interval for the difference between two population means i.e. μ1-μ2:

CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN THE MEANS OF TWO POPULATIONS

For large samples drawn independently from two populations, the C.I. for μ1 – μ2 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where
Subscript 1 denotes the first population, and subscript 2 denotes the second population. We illustrate this concept with the help of a few examples:

EXAMPLE-1:

The means and variances of the weekly incomes in rupees of two samples of workers are given in the following table, the samples being randomly drawn from two different factories:

Factory	Sample Size	Mean	Variance
A	160	12.80	64
B	220	11.25	47

Calculate the 90% confidence interval for the real difference in the incomes of the workers from the two factories.

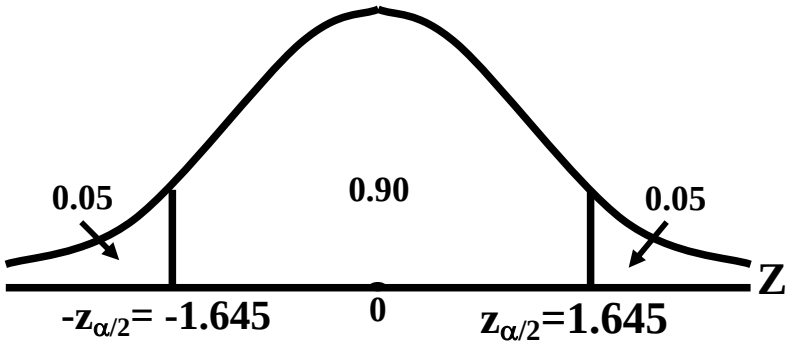
SOLUTION

1. If both n1 and n2 are large, the confidence limits are given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

2. We know that

zα/2 = 1.645 for 90% confidence



3.Hence, Substituting the values in the formula, we obtain

$$(12.80 - 11.25) \pm 1.645 \sqrt{\frac{64}{160} + \frac{47}{220}}$$

$$\sqrt{0.4 + 0.21}$$

or 1.55 ± 1.645

or $1.55 \pm 1.645 \quad \sqrt{0.61}$

or 1.55 ± 1.28

or 0.27 and 2.83

Hence we can say that we are 90% confident that, on the average, the difference in the incomes of the workers from the two factories lies somewhere between Rs.0.27 and Rs.2.83.

EXAMPLE-2

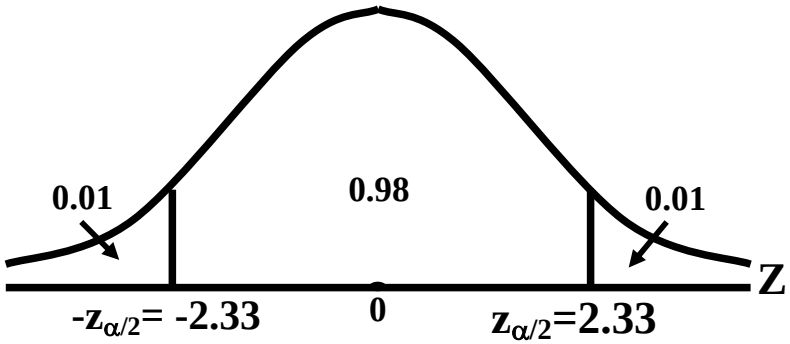
Suppose a study is conducted in a developed country to estimate the difference between middle-income shoppers and low-income shoppers in terms of the average amount saved on grocery bills per week by using coupons. Random samples of 60 middle-income shoppers and 80 low-income shoppers are taken, and their purchases are monitored for 1 week. The average amounts saved with coupons, as well as sample sizes and sample standard deviations are given below:

Middle-Income Shoppers	Low-Income Shoppers
$n_1 = 60$	$n_2 = 80$
$\overline{X}_1 = \$5.84$	$\overline{X}_2 = \$2.67$
$S_1 = \$1.41$	$S_2 = \$0.54$

Use this information to construct a 98% confidence interval to estimate the difference between the mean amounts saved with coupons by middle-income shoppers and low-income shoppers.

SOLUTION:

The value of $z_{\alpha/2}$ associated with a 98% level of confidence is 2.33.



Using this value, we can determine the confidence interval as follows:

$$(5.84 - 2.67) - 2.33\sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}} \le \mu_1 - \mu_2 \le (5.84 - 2.67) + 2.33\sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}}$$

$$3.17 - 0.45 \le \mu_1 - \mu_2 \le 3.17 + 0.45$$

$$2.72 \le \mu_1 - \mu_2 \le 3.62$$

Hence, the 98% confidence interval for the difference between the mean amounts saved with coupons by middle-income shoppers and low-income shoppers is ($2.72, $3.62). The point estimate for the difference in mean savings is $3.17. Note that a zero difference in the population means of these two groups is unlikely, because the number zero is not in the 98% range. The data seems to provide a strong indication that, on the average, the middle income shoppers are saving a little more than the low income shoppers.

LECTURE NO 36

- Large Sample Confidence Intervals for p and p1-p2
- Determination of Sample Size (with reference to Interval Estimation)
- Hypothesis-Testing (An Introduction)

In the last lecture, we discussed the construction and the interpretation of the confidence intervals for μ and μ1 - μ2. We begin today's lecture by focusing on the confidence intervals for p and p1-p2. First, we consider the confidence interval for p, the proportion of successes in a binomial population:

CONFIDENCE INTERVAL FOR A POPULATION PROPORTION (P)

For a large sample drawn from a binomial population, the C.I. for p is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where

$\hat{p}$ = proportion of "successes" in the sample

n = sample size

$z\alpha/2$ = 1.96 for 95% confidence

= 2.58 for 99% confidence

(In a practical situation, the criterion for deciding whether or not n is sufficiently large is that if both np and nq are greater than or equal to 5, then we say that n is sufficiently large).We illustrate this concept with the help of a few examples:

EXAMPLE-1

As a practical illustration, let us look at a survey of teenagers who have appeared in a juvenile court three times or more. A survey of 634 of these shows that 291 are orphans (one or both parents dead). What proportion of all teenagers with three or more appearances in court are orphans? The estimate is to be made with 99% confidence.

SOLUTION

In this problem, we have n = 634, and

$\hat{p}$ =291/634 = 0.459,

$\hat{q} = 1 - \hat{p}$ = 0.541,

Hence, the 99% confidence limits for p are:

$$0.459 \pm 2.58 \sqrt{\frac{0.459 \times 0.541}{634}}$$

= 0.459 ± 0.051

= 0.408 and 0.510

Hence, we estimate that the percentage of teenagers of this type who are orphans lies between 40.8 per cent and 51.0 per cent. It should be noted that, in this problem, happily, the confidence interval has come out to be pretty narrow, and this is happening in spite of the fact that the level of confidence is very high ! This very desirable situation can be ascribed to the fact that the sample size of 634 is pretty large.

EXAMPLE-2

After a long career as a member of the City Council, Mr. Scott decided to run for Mayor.
The campaign against the present Mayor has been strong with large sums of money spent by each candidate on advertisements. In the final weeks, Mr. Scott has pulled ahead according to polls published in a leading daily newspaper. To check the results, Mr. Scott's staff conducts their own poll over the weekend prior to the election. The results show that for a random sample of 500 voters 290 will vote for Mr. Scott. Develop a 95 percent confidence interval for the population proportion who will vote for Mr. Scott. Can he conclude that he will win the election?

SOLUTION

We begin by estimating the proportion of voters who will vote for Mr. Scott. The sample included 500 voters and 290 favored Mr. Scott. Hence, the sample proportion is 290/500 = 0.58. The value 0.58 is a point estimate of the unknown population proportion p.
The 95% Confidence Interval for p is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.58 \pm 1.96 \sqrt{\frac{0.58(1-0.58)}{500}}$$

$$= 0.58 \pm 0.043$$

$$= (0.537,\ 0.623)$$

The end points of the confidence interval are 0.537 and 0.623. The lower point of the confidence interval is greater than 0.50. So, we conclude that the proportion of voters in the population supporting Mr. Scott is greater than 50 percent. He will win the election, based on the polling results.

EXAMPLE-3

A group of statistical researchers surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management-succession plan in place. A spokesman for the group made the statement that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocused a company for long enough to cause it to lose its momentum.

Use the survey-figure to compute a 92% confidence interval to estimate the proportion of all fast-growing small companies that have a management-succession plan.

SOLUTION

The point estimate of the proportion of all fast-growing small companies that have a management-succession plan is the sample proportion found to be 0.51 for that particular sample of size 210 which was surveyed by the group of researchers. Realizing that the point estimate might change with another sample selection, we calculate a confidence interval, as follows:

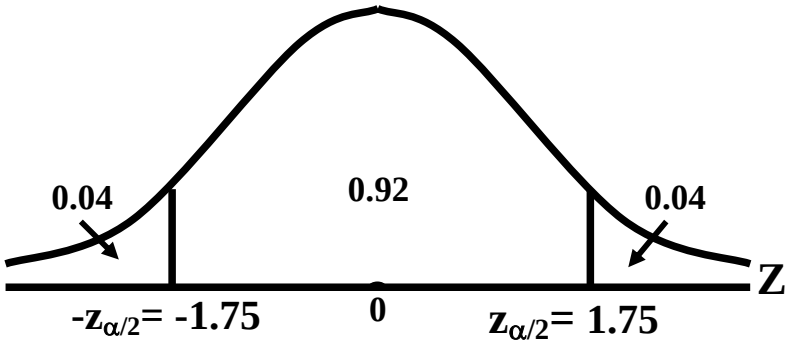
The value of n is 210;

$\hat{p}$ **is 0.51**

and

$$\hat{q} = 1 - \hat{p} = 0.49 .$$

Because the level of confidence is 92%, the value of $Z_{.04} = 1.75$.



The confidence interval is computed as:

$$0.51 - 1.75\sqrt{\frac{(0.51)(0.49)}{210}} \le p$$

$$\le 0.51 + 1.75\sqrt{\frac{(0.51)(0.49)}{210}}$$

$$0.51 - 0.06 \le p \le 0.51 + 0.06$$

$$0.45 \le p \le 0.57$$

$$P(0.45 \le p \le 0.57) = 0.92.$$

CONCLUSION

It is estimated with 92% confidence that the proportion of the population of fast-growing small companies that have a management-succession plan is between 0.45 and 0.57.

Next, we consider the Confidence Interval for the difference in the population proportions (p1 – p2):

CONFIDENCE INTERVAL FOR P1-P2

For large samples drawn independently from two binomial populations, the C.I. for p1-p2 is given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where

subscript 1 denotes the first population, and subscript 2 denotes the second population.

We illustrate this concept with the help of an example:

EXAMPLE

In a poll of college students in a large university, 300 of 400 students living in students' residences (hostels) approved a certain course of action, whereas 200 of 300 students not living in students' residences approved it. Estimate the difference in the proportions favoring the course of action, and compute the 90% confidence interval for this difference.

SOLUTION

Let $\hat{p}_1$ be the proportion of students favouring the course of action in the first sample (i.e. the sample of resident students). And, let $\hat{p}_2$ be the proportion of students favouring the course of action in the second sample (i.e. the sample of students not residing in students' residences).

Then

$$\hat{p}_1 = \frac{300}{400} = 0.75,$$

And

$$\hat{p}_2 = \frac{200}{300} = 0.67.$$

$\therefore$ Difference in proportions

$= \hat{p}_1 - \hat{p}_2 =$ 0.75 – 0.67 = 0.08

The required level of confidence is 0.90. Therefore z0.05 = 1.645, and hence, the 90% confidence interval for p1 – p2 is 90% C.I. for p1-p2:

$$(\hat{p}_1 - \hat{p}_2) \pm (1.645)\sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}}$$

$$or\ \ 0.08 \pm (1.645)\sqrt{\frac{(0.75)(0.25)}{400} + \frac{(0.67)(0.33)}{300}}$$

or 0.08 ± (1.645)

or 0.08 ± (1.645) (0.0347)

or 0.08 ± 0.057

or 0.023 to 0.137

Hence the 90 per cent confidence interval for p1 – p2 is (0.023, 0.137). In other words, on the basis of 90% confidence, we can say that the difference between the proportions of resident students and non-resident students who favor this particular course of action lies somewhere between 2.3% and 13.7%.Evidently, this seems to be a rather wide interval, even though the level of confidence is not extremely high. Hence, it is obvious that, in this example, sample sizes of 400 and 300 respectively, although apparently quite large, are not large enough to yield a desirably narrow confidence interval.

In the last lecture, we discussed the construction and interpretation of confidence intervals. Next, we consider the determination of sample size. In this regard, the first point to be noted is that, in any statistical study based on primary data, the first question is what is going to be the size of the sample that is to be drawn from the population of interest? We present below a method of finding the sample size in such a way that we obtain a desired level of precision with a desired level of confidence, first, we consider the determination of sample size in that situation when we are trying to estimateμ, the population mean:

Sample size for Estimating Population Mean

In deriving the 100(1–α) per cent confidence Interval for μ, we have the expression

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

which implies that the maximum allowable difference between $\bar{X}$ and μ is:

$$\left|\bar{x}-\mu\right| = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where $\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{X}$ when sampling is performed with replacement of population is very large (infinite). The quantity $\left|\bar{x}-\mu\right|$ is also called the error of the estimator $\bar{X}$ and is denoted by e. Thus a 100(1–α) per cent error bound for estimating μ is given by $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. In other words, in order to have a 100(1–α) per cent \confidence that the error is estimating μ with $\bar{X}$ to be less than e, we need n such that

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or $$\sqrt{n} = z_{\alpha/2}\frac{\sigma}{e}$$

or $$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

Hence the desired sample size for being 100(1–α)% confident that the error in estimating μ will be less than e, when sampling is with replacement or the population is very large, is given by

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

It is important to note that the population standard deviation σ is generally not known, and hence, its estimate is found either from past experience or from a pilot sample of size n > 30. In case of fractional result, it is always to be rounded to the next higher integer for the sample size.

EXAMPLE

A research worker wishes to estimate the mean of a population using a sample sufficiently large that the probability will be 0.95 that the sample mean will not differ from the true mean by more than 25 percent of the standard deviation. How large a sample should be taken?

SOLUTION

If the sample mean is not being allowed to differ from the true mean by more than 25% of σ with a probability of 0.95, then

$$e = \left|\bar{x}-\mu\right| = \frac{25\sigma}{100} = \frac{\sigma}{4}, \quad and \quad z_{\alpha/2} = 1.96.$$

Substituting these values in the formula

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2, \text{ we get}$$

$$n = \left(\frac{1.96\times\sigma}{\sigma/4}\right)^2 = 61.4656.$$

Hence the required sample size is 62, (the next higher integer), as the sample size cannot be fractional.

Next, we consider the determination of sample size in that situation when we are trying to estimate p, the proportion of successes in the population:

SAMPLE SIZE FOR ESTIMATING POPULATION PROPORTION

The large sample confidence interval for p is given by

$$\hat{p} = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

This implies that $e = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$

Therefore, solving for n, we obtain

$$n = \frac{(z_{\alpha/2})^2\,\hat{p}\,\hat{q}}{e^2}$$

Since the values of $\hat{p}$ and $\hat{q}$ are not known as the sample has not yet been selected, we therefore use an estimate $\hat{p}$ obtained from pilot sample information.

EXAMPLE

In a random sample of 75 axle shafts, 12 have a surface finish that is rougher than the specification will allow. How large a sample is required if we want to be 95% confident that the error in using $\hat{p}$ to estimate p is less than 0.05?
Solution:

$$e = |\hat{p} - p| = 0.05,$$

Here

$$\hat{p} = \frac{12}{75} = 0.16,$$

$$\hat{q} = 1 - \hat{p} - 0.84 \text{ and } z_{0.025} = 1.96$$
$$(\because \alpha/2 = 0.025)$$

Substituting these values in the formula

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 \hat{p}\hat{q},$$ we obtain

$$n = \left(\frac{1.96}{0.05}\right)^2 \times (0.16)(0.84) = 206.52$$

which, upon rounding upward, yields 207 as the desired sample size. As stated earlier, Inferential Statistics can be divided into two parts, estimation and hypothesis-testing. Having discussed the concepts of point and interval estimation in considerable detail, We now begin the discussion of Hypothesis-Testing:

HYPOTHESIS-TESTING IS A VERY IMPORTANT AREA OF STATISTICAL INFERENCE

It is a procedure which enables us to decide on the basis of information obtained from sample data whether to accept or reject a statement or an assumption about the value of a population parameter. Such a statement or assumption which may or may not be true is called a statistical hypothesis. We accept the hypothesis as being true, when it is supported by the sample data. We reject the hypothesis when the sample data fail to support it. It is important to understand what we mean by the terms 'reject' and 'accept' in hypothesis-testing. The rejection of a hypothesis is to declare it false. The acceptance of a hypothesis is to conclude that there is insufficient evidence to reject it. Acceptance does not necessarily mean that the hypothesis is actually true. The basic concepts associated with hypothesis testing are discussed below:

NULL AND ALTERNATIVE HYPOTHESES

NULL HYPOTHESIS

A null hypothesis, generally denoted by the symbol H0, is any hypothesis which is to be tested for possible rejection or nullification under the assumption that it is true.

A null hypothesis should always be precise such as 'the given coin is unbiased' or 'a drug is ineffective in curing a particular disease' or 'there is no difference between the two teaching methods'. The hypothesis is usually assigned a numerical value. For example, suppose we think that the average height of students in all colleges is 62″. This statement is taken as a hypothesis and is written symbolically as H0 : $\mu = 62''$. In other words, we hypothesize that $\mu = 62''$.

ALTERNATIVE HYPOTHESIS

An alternative hypothesis is any other hypothesis which we are willing to accept when the null hypothesis H0 is rejected. It is customarily denoted by H1 or HA. A null hypothesis H0 is thus tested against an alternative hypothesis H1. For example, if our null hypothesis is H0 : $\mu = 62''$, then our alternative hypothesis may be H1 : $\mu \neq 62''$ or H1 : $\mu < 62''$.

LEVEL OF SIGNIFICANCE

The probability of committing Type-I error can also be called the level of significance of a test. Now, what do we mean by Type-I error? In order to obtain an answer to this question, consider the fact that, as far as the actual reality is concerned, H0 is either actually true, or it is false. Also, as far as our decision regarding H0 is concerned, there are two possibilities --- either we will accept H0, or we will reject H0. The above facts lead to the following table:

		Decision	
		Accept H_0	Reject H_0 (or accept H_1)
True Situation	H_0 is true	Correct decision (No error)	Wrong decision (Type-I error)
	H_0 is false	Wrong decision (Type-II error)	Correct decision (No error)

A close look at the four cells in the body of the above table reveals that the situations depicted by the top-left corner and the bottom right-hand corner are the ones where we are taking a correct decision. On the other hand, the situation depicted by the top-right corner and the bottom left-hand corner are the ones where we are taking an incorrect decision. The situation depicted by the top-right corner of the above table is called an error of the first kind or a Type I-error, while the situation depicted by the bottom left-hand corner is called an error of the second kind or a Type II-error.
In other words:

TYPE-I AND TYPE-II ERRORS

On the basis of sample information, we may reject a null hypothesis H0, when it is, in fact, true or we may accept a null hypothesis H0, when it is actually false. The probability of making a Type I error is conventionally denoted by α and that of committing a Type II error is indicated by β. In symbols, we may write

α = P (Type I error)
= P (reject H0|H0 is true),

β = P (Type II error)
= P (accept H0|Ho is false).

LECTURE NO. 37

- Hypothesis-Testing (continuation of basic concepts)
- Hypothesis-Testing regarding μ (based on Z-statistic)

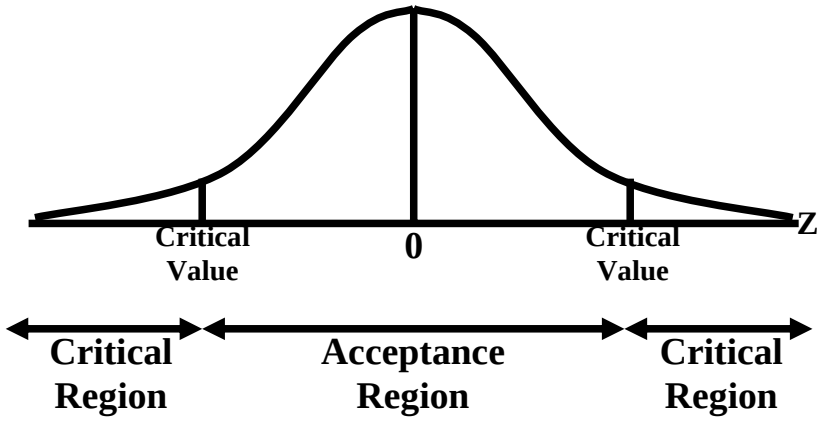
In the last lecture, we commenced the discussion of the concept of Hypothesis-Testing. We introduced the concepts of the Null and Alternative hypotheses as well as the concepts of Type-I and Type-II error. We now continue the discussion of the basic concepts of hypothesis-testing:

TEST-STATISTIC

A statistic (i.e. a function of the sample data not containing any parameters), which provides a basis for testing a null hypothesis, is called a test-statistic. Every test-statistic has a probability distribution (i.e. sampling distribution) which gives the probability that our test-statistic will assume a value greater than or equal to a specified value OR a value less than or equal to a specified value when the null hypothesis is true.

ACCEPTANCE AND REJECTION REGIONS

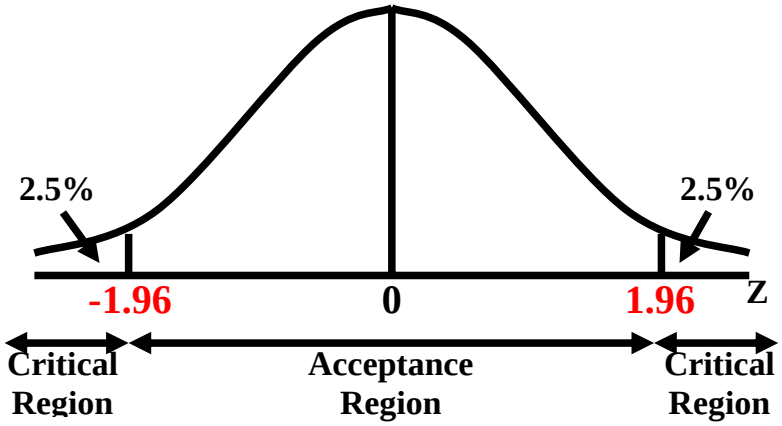
All possible values which a test-statistic may assume can be divided into two mutually exclusive groups: one group consisting of values which appear to be consistent with the null hypothesis (i.e. values which appear to support the null hypothesis), and the other having values which lead to the rejection of the null hypothesis. The first group is called the acceptance region and the second set of values is known as the rejection region for a test. The rejection region is also called the critical region. The value(s) that separates the critical region from the acceptance region, is called the critical value(s):



The critical value which can be in the same units as the parameter or in the standardized units, is to be decided by the experimenter. The most frequently used values of α, the significance level, are 0.05 and 0.01, i.e. 5 percent and 1 percent. By α = 5%, we mean that there are about 5 chances in 100 of incorrectly rejecting a true null hypothesis.

RELATIONSHIP BETWEEN THE LEVEL OF SIGNIFICANCE AND THE CRITICAL REGION

The level of significance acts as a basis for determining the CRITICAL REGION of the test. For example, if we are testing H0: $\mu = 45$ against H1: $\mu \neq 45$, our test statistic is the standard normal variable Z, and the level of significance is 5%, then the critical values are Z = +1.96 Corresponding to a level of significance of 5%, we have:



ONE-TAILED AND TWO-TAILED TESTS

A test, for which the entire rejection region lies in only one of the two tails – either in the right tail or in the left tail – of the sampling distribution of the test-statistic, is called a one-tailed test or one-sided test. A one-tailed test is used when the alternative hypothesis H1 is formulated in the following form:

$H1 : \theta > \theta 0$

or

$H1 : \theta < \theta 0$

For example, if we are interested in testing a hypothesis regarding the population mean, if n is large, and we are conducting a one-tailed test, then our alternative hypothesis will be stated as

$H1 : \mu > \mu 0$

or

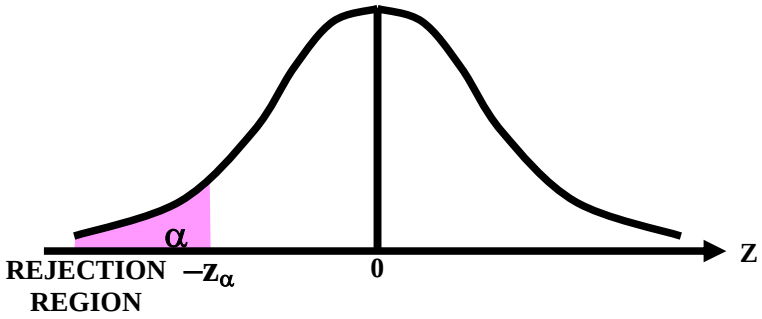
$H1 : \mu < \mu 0$

In this case, the rejection region consists of either all z-values which are greater than $+ z\alpha$ or less than $- z\alpha$ (where α is the level of significance):

If $H0 : \mu > \mu 0$

$H1 : \mu < \mu 0$

Then (in case of large n):



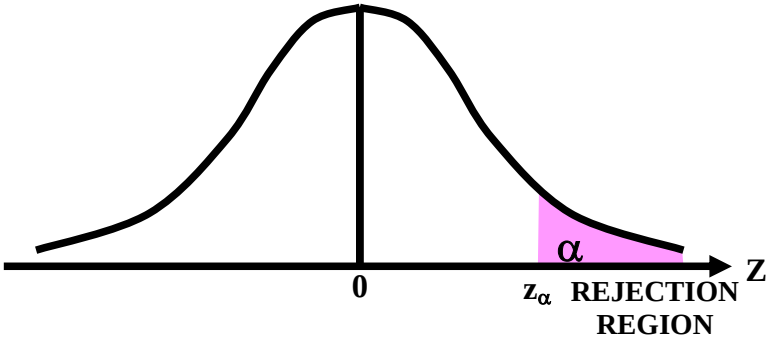
REJECT H0 if $z < -z\alpha$

If

$H0 : \mu < \mu 0$

$H1 : \mu > \mu 0$

Then (in case of large n):



REJECT H0 if $z > z\alpha/2$

If, on the other hand, the rejection region is divided equally between the two tails of the sampling distribution of the test-statistic, the test is referred to as a two-tailed test or two-sided test.

In this case, the alternative hypothesis H1 is set up as:

H1 : $\mu \neq \mu 0$

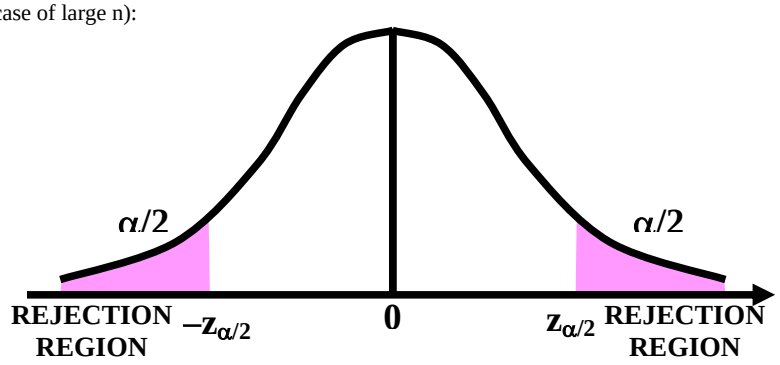
meaning thereby

H1 : $\mu < \mu 0$ or $\mu > \mu 0$

If H0 : $\mu = \mu 0$

H1 : $\mu \neq \mu 0$

Then (in case of large n):



REJECT H0 if $z < -z\alpha/2$ or $z > z\alpha/2$

The location of critical region can be determined only after the alternative hypothesis H1 has been stated. It is important to note that the one-tailed and the two-tailed tests differ only in location of the critical region, not in the size. We illustrate the concept and methodology of hypothesis-testing with the help of an example:

EXAMPLE

A steel company manufactures and assembles desks and other office equipment at several plants in a particular country. The weekly production of the desks of Model A at Plant-I has a mean of 200 and a standard deviation of 16. Recently, due to market expansion, new production methods have been introduced and new employees hired. The vice president of manufacturing would like to investigate whether there has been a change in the weekly production of the desks of Model A. To put it another way, is the mean number of desks produced at Plant-I different from 200 at the 0.05 significance level? The mean number of desks produced last year (50 weeks, because the plant was shut down 2 weeks for vacation) is 203.5. On the basis of the above result, should the vice president conclude that the there has been a change in the weekly production of the desks of Model A.

SOLUTION:

We use the statistical hypothesis-testing procedure to investigate whether the production rate has changed from 200 per month.

Step-1:

Formulation of the Null and Alternative Hypotheses:

The null hypothesis is "The population mean is 200."

The alternative hypothesis is 'The mean is different from 200" or "The mean is not 200."

These two hypotheses are written as follows:

H0 : $\mu = 200$

H1 : $\mu \neq 200$

Note:

This is a two-tailed test because the alternative hypothesis does not state a direction. In other words, it does not state whether the mean production is greater than 200 or less than 200. The vice president only wants to find out whether the production rate is different from 200.

Step-2:

Decision Regarding the Level of Significance (i.e. the Probability of Committing Type-I Error):

Here, the level of significance is 0.05.

This is α, the probability of committing a Type-I error (i.e. the risk of rejecting a true null hypothesis).

Step-3:

Test Statistic (that statistic that will enable us to test our hypothesis):

The test statistic for a large sample mean is

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Transforming the production data to standard units (z values) permits the use of the area table of the standard normal distribution.

Step-4:

Calculations:

In this problem, we have n = 50, $\bar{X}$ = 203.5, and σ = 16.

Hence, the computed value of z comes out to be:

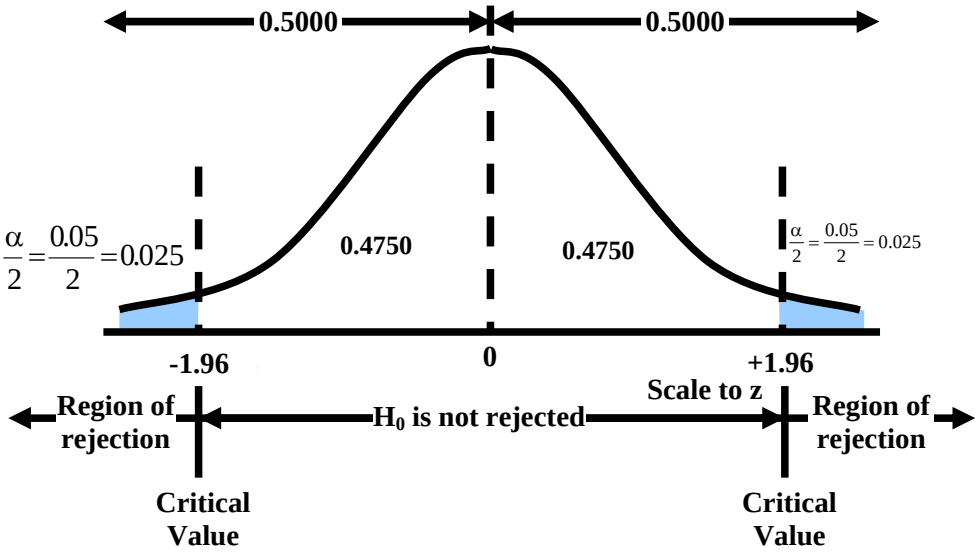
$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{203.5 - 200}{16 / \sqrt{50}} = 1.55$$

Step-5:

Critical Region (that portion of the X-axis which compels us to reject the null hypothesis):Since this is a two-tailed test, half of 0.05, or 0.025, is in each tail.

The area where H0 is not rejected, located between the two critical values, is therefore 0.95.

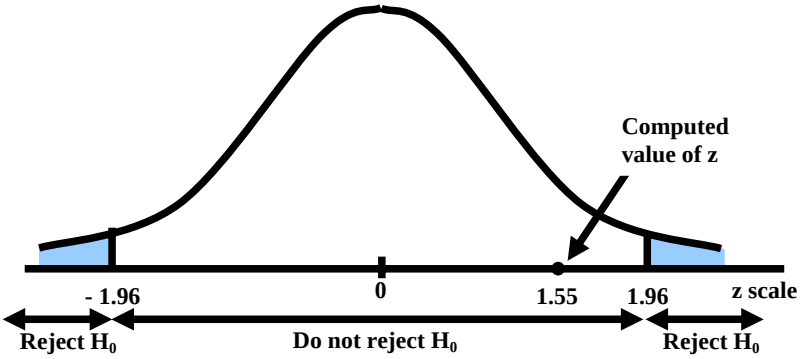
Applying the inverse use of the Area Table, we find that, corresponding to α = 0.05, the critical values are 1.96 and -1.96, as shown below:



DECISION RULE FOR THE 0.05 SIGNIFICANCE LEVEL THE DECISION RULE IS, THEREFORE

Reject the null hypothesis and accept the alternative hypothesis if the computed value of z is not between –1.96 and +1.96. Do not reject the null hypothesis if z falls between –1.96 and + 1.96.

Step-6:
Conclusion:
The computed value of z i.e. 1.55 lies between -1.96 and + 1.96, as shown below:



Because 1.55 lies between -1.96 and + 1.96, therefore, it does not fall in the rejection region, and hence H0 is not rejected. In other words, we conclude that the population mean is not different from 200. So, we would report to the vice president of manufacturing that the sample evidence does not show that the production rate at Plant-I has changed from 200 per week. The difference of 3.5 units between the historical weekly production rate and the production rate of last year can reasonably be attributed to chance. The above example pertained to a two-tailed test. Let us now consider a few examples of one-tailed tests:

EXAMPLE

A random sample of 100 workers with children in day care show a mean day-care cost of Rs.2650 and a standard deviation of Rs.500. Verify the department's claim that the mean exceeds Rs.2500 at the 0.05 level with this information.

SOLUTION
In this problem, we regard the department's claim, that the mean exceeds Rs.2500, as H1,
and regard the negation of this claim as H0.

Thus, we have

i) H0 : $\mu < 2500$
H1 : $\mu > 2500$ (exceeds 2500)

(Important Note: We should always regard that hypothesis as the null hypothesis which contains the equal sign.)

ii) We are given the significance level at $\alpha = 0.05$.

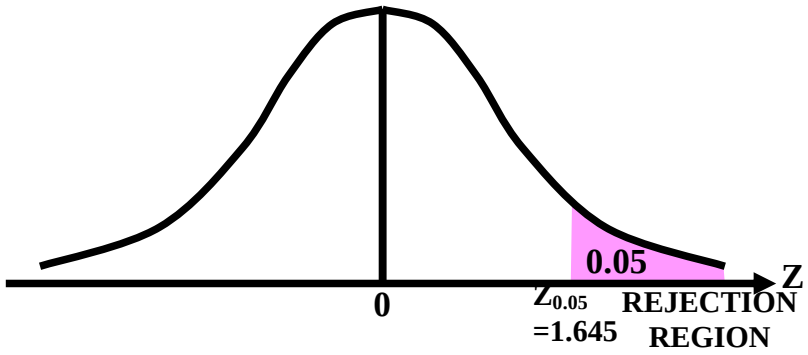
iii) The test-statistic, under H0 is

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

which is approximately normal as n = 100 is large enough to make use of the central limit theorem.

iv) The rejection region is

$Z > Z0.05 = 1.645$



v) Computing the value of Z from sample information, we find

$$z = \frac{2650 - 2500}{500/\sqrt{100}} = \frac{150}{50} = 3$$

vi) Conclusion:

Since the calculated value z = 3 is greater than 1.645, hence it falls in the rejection region, and, therefore, we reject H0, and may conclude that the department's claim is supported by the sample evidence.

An Interesting and Important Point:

For $\alpha = 0.01$, $Z\alpha = 2.33$.

As our computed value of Z i.e. 3 is even greater than 2.33, the computed value of $\bar{X}$ is highly significant. (With only 1% chance of being wrong, the department's claim was correct).

LECTURE NO. 38

- Hypothesis-Testing regarding μ1 - μ2 (based on Z-statistic)
- Hypothesis Testing regarding p (based on Z-statistic)

In the last lecture, we discussed the basic concepts involved in hypothesis-testing. Also, we applied this concept to a few examples regarding the testing of the population mean μ. These examples pointed to the six main steps involved in any hypothesis-testing procedure.

GENERAL PROCEDURE FOR TESTING HYPOTHESES

Testing a hypothesis about a population parameter involves the following six steps:

- State your problem and formulate an appropriate null hypothesis H0 with an alternative hypothesis H1, which is to be accepted when H0 is rejected.
- Decide upon a significance level of the test, α, which is the probability of rejecting the Null Hypothesis if it is true.
- Choose a test-statistic such as the normal distribution, the t-distribution, etc. to test H0.
- Determine the rejection or critical region in such a way that the probability of rejecting the null hypothesis H0, if it is true, is equal to the significance level, α. The location of the critical region depends upon the form of H1 (i.e. whether we are carrying out a one-tailed test or a two-tailed test). The critical value(s) will separate the acceptance region from the rejection region.
- Compute the value of the test-statistic from the sample data in order to decide whether to accept or reject the null hypothesis H0.
- Formulate the decision rule (i.e. draw a conclusion) as follows:
  a) Reject the null hypothesis H0, if the computed value of the test statistic falls in the rejection region.
  b) Accept the null hypothesis H0, otherwise.

IMPORTANT NOTE

It is very important to realize that when applying a hypothesis-testing procedure of the type explained above, we always begin by assuming that the null hypothesis is true.

IMPORTANT NOTE:

As s2 is an unbiased estimator of σ2 whereas S2 is a biased estimator, hence we would like to use this estimator whenever σ2 is unknown. However, when n is large, s2 is approximately equal to S2, as explained below:
We know that

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

whereas

$$S^2 = \frac{\sum (x-\bar{x})^2}{n} \Rightarrow \sum (x-\bar{x})^2 = nS^2.$$

Hence

$$(n-1)s^2 = nS^2 \Rightarrow S^2 = \frac{(n-1)}{n}s^2 = \left(1-\frac{1}{n}\right)s^2$$

Now, as $n \to \infty$, $\frac{1}{n} \to 0.$

Hence, if n is large,

$$S^2 \simeq s^2.$$

Hence, in case of a large sample drawn from a population with unknown variance σ2, we may replace σ2 by S2.We now consider the case when we are interested in testing the equality of two population means.
We illustrate this situation with the help of the following example.

EXAMPLE

A survey conducted by a market-research organization five years ago showed that the estimated hourly wage for temporary computer analysts was essentially the same as the hourly wage for registered nurses. This year, a random sample of 32 temporary computer analysts from across the country is taken. The analysts are contacted by telephone and asked what rates they are currently able to obtain in the market-place A similar random sample of 34 registered nurses is taken. The resulting wage figures are listed in the following table:

Computer Analysts			Registered Nurses		
$ 24.10	$25.00	$24.25	$20.75	$23.30	$22.75
23.75	22.70	21.75	23.80	24.00	23.00
24.25	21.30	22.00	22.00	21.75	21.25
22.00	22.55	18.00	21.85	21.50	20.00
23.50	23.25	23.50	24.16	20.40	21.75
22.80	22.10	22.70	21.10	23.25	20.50
24.00	24.25	21.50	23.75	19.50	22.60
23.85	23.50	23.80	22.50	21.75	21.70
24.20	22.75	25.60	25.00	20.80	20.75
22.90	23.80	24.10	22.70	20.25	22.50
23.20			23.25	22.45	
23.55			21.90	19.10	

Conduct a hypothesis test at the 2% level of significance to determine whether the hourly wages of the computer analysts are still the same as those of registered nurses.

SOLUTION

Hypothesis Testing Procedure:
Step-1:
Formulation of the Null and Alternative Hypotheses:

H0 : $\mu 1 - \mu 2 = 0$
HA : $\mu 1 - \mu 2 \neq 0$
(Two-tailed test)

Step-2:

Level of Significance:
$\alpha = 0.02$

Step-3:

Test Statistic:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step-4:

Calculations:
The sample size, sample mean and sample standard deviation for each of the two samples are given below:
Computer Analysts:

n1	=	32
$\overline{X}$ 1	=	$23.14
S12	=	1.854

Registered Nurses:

n2	=	34
$\bar{X}2$	=	$21.99
S22	=	1.845

Since the sample sizes are larger than 30, hence, the unknown population variances σ12 and σ22 can be replaced by S12 and S22. Hence, our formula becomes:

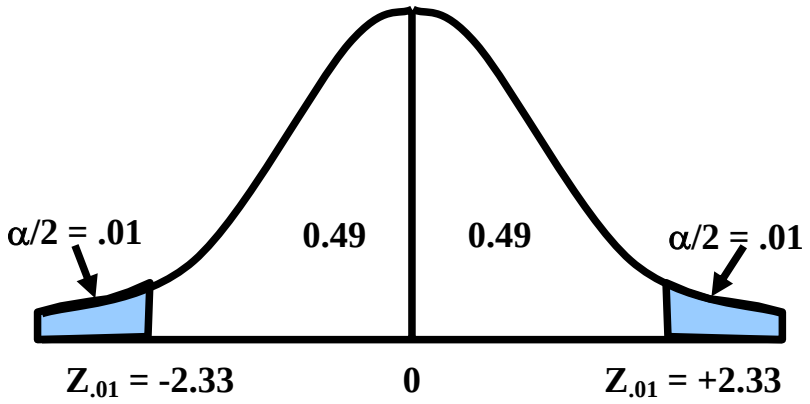
$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Hence, the computed value of Z comes out to be :

$$Z = \frac{(23.14 - 21.99) - (0)}{\sqrt{\frac{1.854}{32} + \frac{1.845}{34}}} = \frac{1.15}{0.335} = 3.43$$

Step-5:

Critical Region:
As the level of significance is 2%, and this is a two-tailed test, hence, we have the following situation:

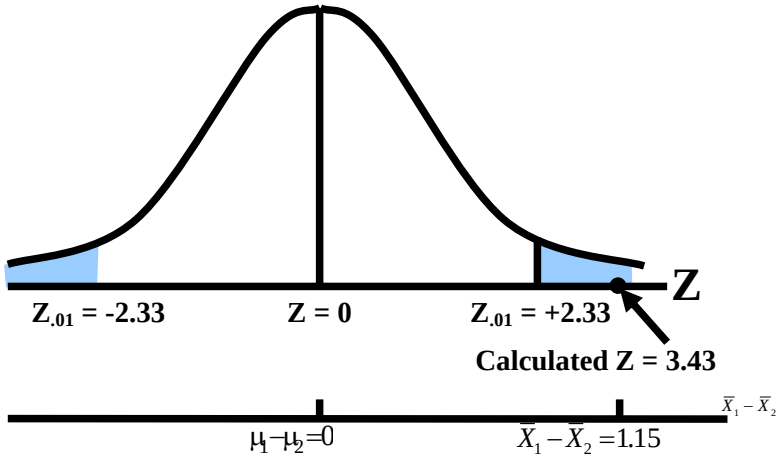


Hence, the critical region is given by

| Z | > 2.33

Step-6:

Conclusion:
As the computed value i.e. 3.43 is greater than the tabulated value 2.33, hence, we reject H0.



The researcher can say that there is a significant difference between the average hourly wage of a temporary computer analyst and the average hourly wage of a temporary registered nurse. The researcher then examines the sample means and uses common sense to conclude that, on the average, temporary computer analyst earn more than temporary registered nurses. Let us consolidate the above concept by considering another example:

EXAMPLE

Suppose that the workers of factory B believe that the average income of the workers of factory A exceeds their average income. A random sample of workers is drawn from each of the two factories, an the two samples yield the following information:

Factory	Sample Size	Mean	Variance
A	160	12.80	64
B	220	11.25	47

Test the above hypothesis?

SOLUTION

Let subscript 1 denote values pertaining to Factory A, and let subscript 2 denote values pertaining to Factory B. Then, we proceed as follows:
Hypothesis-testing Procedure:

Step 1:

H0 : $\mu 1 < \mu 2$ (or $\mu 1 - \mu 2 < 0$)
HA : $\mu 1 > \mu 2$ (or $\mu 1 - \mu 2 > 0$).

Step 2:
Level of significance
= 5%.

Steps 3 & 4:

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{12.80 - 11.25}{\sqrt{\frac{64}{160} + \frac{47}{220}}}$$

$$= \frac{1.55}{\sqrt{0.61}} = \frac{1.55}{0.78} = 1.99$$

Step 5:
Critical Region:

Since it is a right-tailed test, hence the critical region is given by
$Z > Z_{0.05}$
i.e. $Z > 1.645$

Step 6:
Conclusion:
Since 1.99 is greater than 1.645, hence H0 should be rejected in favour of HA. The sample evidence has consolidated the belief of the workers of factory B. Next, we consider the case when we are interested in conducting a test regarding p, the proportion of successes in the population. We illustrate this situation with the help of the following example:

EXAMPLE

A sociologist has a hunch that not more than 50% of the children who appear in a particular juvenile court three times or more are orphans. To test this hypothesis, a sample of 634 such children is taken and it is found that 341 of these children are orphans, (one or both parents dead). Test the above hypothesis using 1% level of significance.

SOLUTION

Hypothesis-testing Procedure:

Step 1:

H0 : p < 0.50
HA : p > 0.50
(one-tailed test)

Step 2:

Level of significance: $\alpha = 1\%$

Step 3:
Test statistic:

$$Z = \frac{X \pm \frac{1}{2} - n\ p_0}{\sqrt{n\ p_0(1-p_0)}}$$

(where + ½ denotes the continuity correction)

Step 4:
Computation:

Here np0 = 634 (0.50) = 317
and X = 341
Hence X > np0 so use X - ½

$$\textbf{So} \quad Z = \frac{341 - \frac{1}{2} - 317}{\sqrt{634(0.50)(0.50)}} = \frac{23.5}{12.59}$$

$$= \mathbf{1.87}$$

Step 5:
Critical region:

Since $\alpha = 0.01$, hence the critical region is given by

$Z > 2.33$

Step 6:
Conclusion:
Since 1.87 < 2.33,
Hence the computed Z does not fall in the critical region. Hence, we conclude that the sociologist's hunch is acceptable.

LECTURE NO. 39

- Hypothesis Testing Regarding p1-p2 (based on Z-statistic)
- The Student's t-distribution
- Confidence Interval for μ based on the t-distribution

In the last lecture, we discussed hypothesis-testing regarding p, the proportion of successes in a binomial population. Next, we consider the case when we are interested in testing the equality of two population proportions. We illustrate this situation with the help of the following example:

EXAMPLE

A leading perfume company in a western country recently developed a new perfume which they plan to market under the name 'Fragrance'. A number of comparison tests indicate that 'Fragrance' has very good market potential. The Sales Departments of the company want to plan their strategy so as to reach and impress the largest possible segments of the buying public. One of the questions is whether the perfume is preferred by younger or older women. These are two independent populations, a population consisting of the younger women and a population consisting of the older women. A standard scent test will be used where each sampled woman is asked to sniff several perfumes, one of which is 'Fragrance', and indicate the one that she likes best.
A total of 100 young women were selected at random, and each was given the standard scent test. Twenty of the 100 young women chose 'Fragrance' as the perfume they liked best. Two hundred older women were selected at random, and each was given the same standard scent test. Of the 200 older women, 100 preferred 'Fragrance
Test the hypothesis that there is no difference between the proportions of younger and older women who prefer 'Fragrance'.

SOLUTION

We designate p1 as the proportion of younger women who prefer 'Fragrance' and p2 as the proportion of older women who prefer 'Fragrance'.
Hypothesis-Testing Procedure:
Step-1:
H0 : p1 = p2 (i.e. p1-p2 = 0)
(There is no difference between the proportions of young women and older women who prefer 'Fragrance'.)
H1 : p1 ≠ p2 (i.e. p1-p2 ≠ 0)
(The two proportions are not equal.)
Step-2:
Level of Significance
$\alpha = 0.05$.
Step-3:
Test Statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where the combined or pooled proportion, $\hat{p}_c$, is given by:

$$\hat{p}_c = \frac{\text{Total number of successes in the two samples combined}}{\text{Total number of observations in the two samples combined}}$$

$$= \frac{X_1 + X_2}{n_1 + n_2}$$

This can also be written as

$$\hat{p}_c = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

which means that $\hat{p}_c$ is the weighted mean of $\hat{p}_1$ and $\hat{p}_2$, n1 and n2 acting as the weights.

Important Note:
In this example, as the hypothesized value of p1 - p2 is equal to zero; therefore both $\hat{p}_1$ *and* $\hat{p}_2$ are estimating the common population proportion p. Hence, we use the pooled proportion of the two samples to estimate p.
(The rationale is that the pooled estimator $\hat{p}_c$ is a better estimator of the common Population proportion p (as compared with $\hat{p}_1$ or $\hat{p}_2$), as it is based on n1 + n2 observations (i.e. based on a greater amount of information).

Step-4:

Calculations:

X1is the number of Preferring 'Fragrance' = 20. n1 is the number is the sample = 100.

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{20}{100} = 0.20$$

X2 is the number of preferring 'Fragrance' = 100. n2 is the number is the sample = 200.

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{100}{200} = 0.50$$

Now, the pooled or weighted proportion $\hat{p}_c$, is computed as follows:

$$\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2} = \frac{20 + 100}{100 + 200} = \frac{120}{300} = 0.40$$

Computation:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

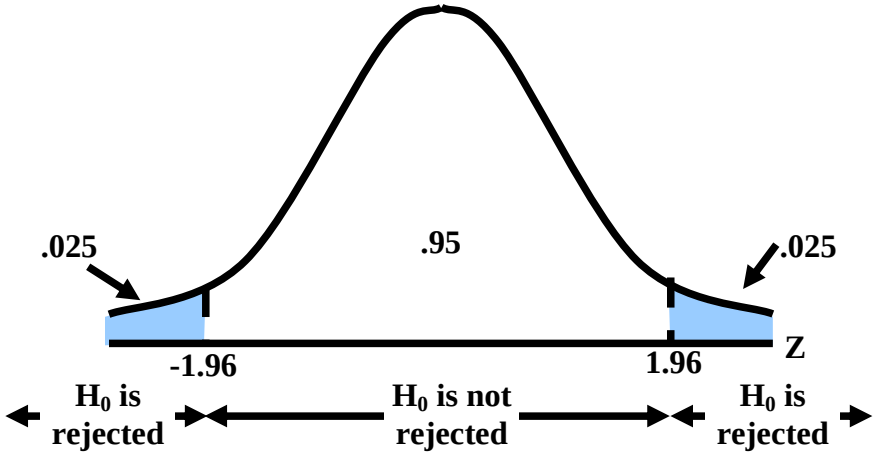
$$= \frac{0.20 - 0.50}{\sqrt{(0.40)(0.60)\left(\frac{1}{100} + \frac{1}{200}\right)}}$$

$$= \frac{-0.30}{0.06} = -5.00$$

Step-5:

Critical Region

Since H1 does not state any direction (such as p1 < p2), the test is two-tailed. Thus, the critical values for the .05 level are –1.96 and + 1.96. Two-Tailed Test, Areas of Rejection and Non-rejection, .05 Level of Significance:



Step-6:

CONCLUSION

The computed z of –5.00 is in the area of rejection, that is, to the left of –1.96. Therefore, the null hypothesis is rejected at the .05 level of significance.

In other words, we conclude that the proportion of young women in the population who prefer 'Fragrance' is not equal to the proportion of older women in the population who prefer 'Fragrance'.(The difference between the two sample proportion i.e. 0.30 is so large that it is highly unlikely that such a large difference could be due to chance (i.e. attributable to sampling fluctuations).)
In fact, the value z = -5.00 is even larger than -2.58, the critical value lying on the left tail of the sampling distribution if $\alpha = 0.01$. As such, we can say that our statistic is highly significant.(In such a situation, the statistic is said to be highly significant because of the fact that we are allowing as small a risk of committing Type-I error as 1%.)
Now, consider another situation:
Suppose that the computed value of our test-statistic comes out to be such that it falls between -1.96 and -2.58. In such a situation, we will reject H0 at the 5% level of significance, but we cannot reject H0 at the 1% level. This means that, if we are willing to allow as much as 5% risk of committing type I error, then we say that we are going to reject H0. But if we are willing to allow only 1% risk of committing type I error, then we conclude that the sample does not provide sufficient evidence to reject H0.Going back to the example of the perfume, obviously, the company would be interested in determining, which category of women prefers this perfume in greater numbers than the other?
The data clearly indicates that the proportion of women who prefer this particular perfume is higher in the population of older women. (This is the reason why the computed value of our test-statistic has come out to be negative.) Let us consolidate the above ideas by considering another example:

EXAMPLE

A candidate for mayor in a large city believes that he appeals to at least 10 per cent more of the educated voters than the uneducated voters. He hires the services of a poll-taking organization, and they find that 62 of 100 educated voters interviewed support the candidate, and 69 of 150 uneducated voters support him at the 0.05 significance level.

Step-1:

The null and alternative hypothesis is
H0 : p1 – p2 > 0.10, and
H1 : p1 – p2 < 0.10, where p1 = proportion of educated voters, and p2 = proportion of uneducated voters.

Step-2:

Level of Significance:
$\alpha = 0.05$.

Step-3:

Test Statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0.10}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

which for large sample sizes, is approximately standard normal.

Important Note
In this example, as the hypothesized value of p1 - p2 is not equal to zero, therefore are note estimating the same quantity, and, as such, we do not use in the formula of the test statistic.

Step-4:

Computation:

$$\text{Here} \quad \hat{p}_1 = \frac{62}{100} = 0.62, \text{ so that } \hat{q}_1 = 0.38,$$

$$\hat{p}_2 = \frac{69}{150} = 0.46, \text{ so that } \hat{q}_2 = 0.54.$$

$$\text{Thus} \quad z = \frac{(0.62 - 0.46) - 0.10}{\sqrt{\frac{(0.62)(0.38)}{100} + \frac{(0.46)(0.54)}{150}}}$$

$$= \frac{0.06}{\sqrt{0.002356 + 0.001656}} = \frac{0.06}{0.063} = 0.95.$$

Step-5:

Critical Region:
As this is a one-tailed test, therefore the critical region is given by

$$Z < -z_{0.05} = -1.645$$

Step-6:
Conclusion:
Since the calculated value z = 0.95 does not fall in the critical region, so we accept the null hypothesis
H0 : p1 – p2 > 0.10.The data seems to support the candidate's view.
Until now, we have discussed in considerable detail interval estimation and hypothesis-testing based on the standard normal distribution and the Z-statistic.
Next, we begin the discussion of interval estimation hypothesis-testing based on the t-distribution.

t-DISTRIBUTION

We begin by presenting the formal definition of the t-distribution and stating some of its main properties:

The Student's t-Distribution:

The mathematical equation of the t-distribution is as follows:

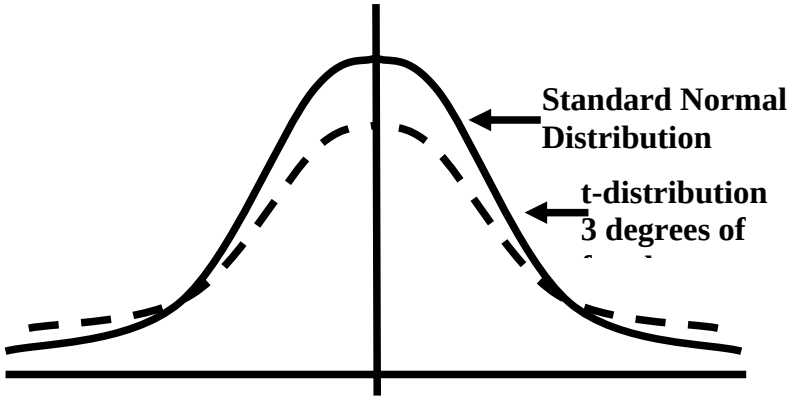
$$f(x) = \frac{1}{\sqrt{\nu}\,\beta\left(\frac{1}{2},\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2} \quad , -\infty < x < \infty$$

This distribution has only one parameter ν, which is known as the degrees of freedom of the t-distribution

PROPERTIES OF STUDENT'S t-DISTRIBUTION

The t-distribution has the following properties:

i) The t-distribution is bell-shaped and symmetric about the value t = 0, ranging from $-\infty$ to ∞.
ii) The number of degrees of freedom determines the shape of the t-distribution.
Thus there is a different t-distribution for each number of degrees of freedom.
As such, it is a whole family of distributions.
The t-distribution, for small values of ν, is flatter than the standard normal distribution which means that the t-distribution is more spread out in the tails than is the standard normal distribution.



As the degrees of freedom increase, the t-distribution becomes narrower and narrower, until, as n tends to infinity, it tends to coincide with the standard normal distribution.
(The t-distribution can never become narrower than the standard normal distribution.)
iii) The t-distribution has a mean of zero, when $\nu \geq 2$. (The mean does not exist when $\nu = 1$.)
iv) The median of the t-distribution is also equal to zero.
v) The t-distribution is unimodal.The density of the distribution reaches its maximum at t = 0 and thus the mode of the t-distribution is t = 0.
(The students will recall that, for any hump-shaped symmetric distribution, the mean, median and mode are equal.)

vi) The variance of the t-distribution is given by $\sigma^2 = \frac{\nu}{\nu - 2}$ for $\nu > 2$.

It is always greater than1, the variance of the standard normal distribution. (This indicates that the t-distribution is more spread out than the standard normal distribution.)
For $\nu \leq 2$, the variance does not exist. Next, we discuss the application of the t-distribution in statistical inference --- those situations where we need to carry out interval estimation and hypothesis - testing on the basis of the t-distribution. (Situations where the t-distribution is the appropriate sampling distribution)
With reference to interval estimation and hypothesis-testing about μ, it has been mathematically proved that, if the population from which the sample has been drawn is normally distributed, the population variance is unknown, and the sample size is small (less than 30), then the statistic

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$\left(\text{where } s = \sqrt{\frac{\sum(X - \overline{X})^2}{n-1}}\right)$$

Follows the t-distribution having n-1 degrees of freedom. First, we discuss the construction of a Confidence Interval for μ based on the t-distribution with the help of an example:

EXAMPLE

The masses, in grams, of thirteen ball bearings seen at random from a batch are
21.4, 23.1, 25.9, 24.7, 23.4, 24.5, 25.0, 22.5, 26.9, 26.4, 25.8, 23.2, 21.9
Calculate a 95% confidence interval for the mean mass of the population, supposed normal, from which these masses were drawn.

SOLUTION

The 95% confidence interval for the mean mass of the population μ, is given by

$$\overline{X} + t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}}$$

(The derivation of the above confidence interval is very similar to that of the confidence interval for μ based on the Z-statistic.)
Now, in this problem, the sample mean $\overline{X}$ and s come out to be:

$$\overline{X} = \frac{\sum X}{n} = \frac{314.7}{13} = 24.21,$$

$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]}$$

$$= \sqrt{\frac{1}{12}[7655.59 - 7618.16]} = \sqrt{\frac{37.43}{12}} = \sqrt{3.12} = 1.77$$

The question is: 'How do we find $t_{\alpha/2(n-1)}$?

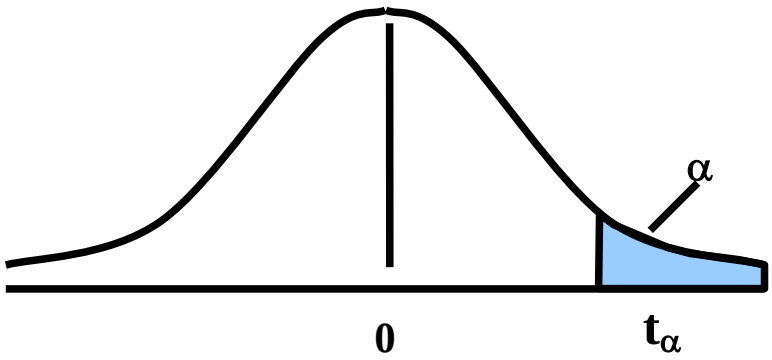
For this purpose, we will need to consult the table of areas under the t-distribution:

TABLE OF AREAS UNDER THE T-DISTRIBUTION

Upper Percentage Points of the t-Distribution							
ν \ α	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	1.000	3.078	6.314	12.706	31.821	63.657	318.310
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.838	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733

Upper Percentage Points of the t-Distribution							
ν \ α	0.25	0.10	0.05	0.025	0.01	0.005	0.001
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

The above table is an abridged version of the table by Fisher and Yates, and the entries in this table are values of $t_{\alpha,(\nu)}$ for which the area to their right under the t-distribution with ν degrees of freedom is equal to α, as shown below:



Now, in this problem, since n – 1 = 12, and the desired level of confidence is 95%, therefore, the right-tail area is 2½%, and, hence, (using the t-table) we obtain

$$t_{0.025\,(12)} = 2.179$$

Substituting these values, we obtain the 95% confidence interval for μ as follows:

$$24.21 \pm 2.179\left(\frac{1.77}{\sqrt{13}}\right)$$

or 24.21 ± 2.179 (0.49)
or 24.21 ± 1.07 or 23.14 to 25.28

Hence, the 95% confidence interval for the mean mass of the ball bearings calculated from the given sample is (23.1, 25.3) grams.

LECTURE NO. 40

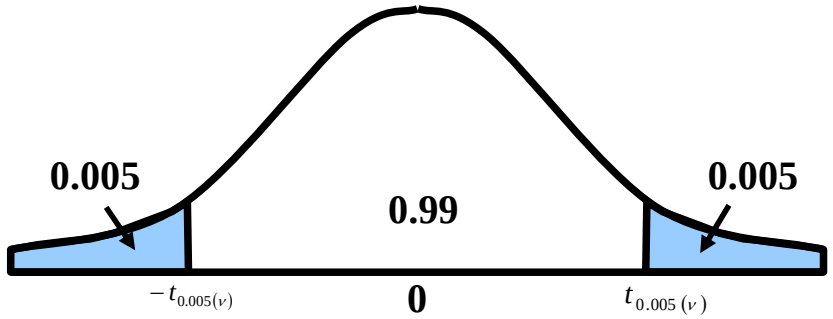
- Tests and Confidence Intervals based on the t-distribution

In the last lecture, we introduced the t-distribution, and began the discussion of statistical inference based on the t-distribution. In particular, we discussed the construction of the confidence interval for μ in that situation when we are drawing a small sample from a normal population having unknown variance σ2. When the parent population is normal, the population variance is unknown, and the sample size n is small (less than 30), then the confidence interval for μ is given by

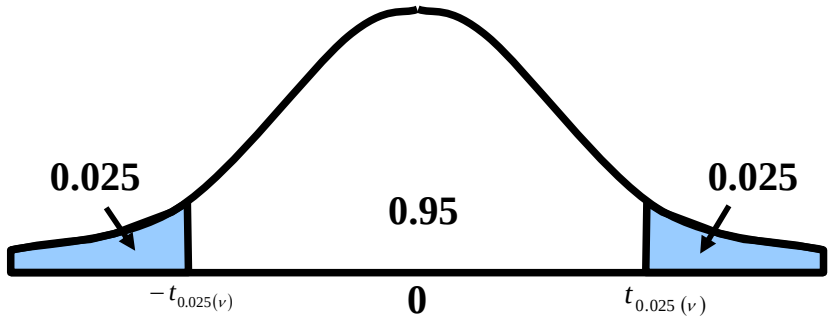
$$\bar{x} \pm t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}}$$

where $\bar{X} = \frac{\sum x}{n}$ is the sample mean $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ is the sample standard deviation n = sample size and $t_{(\alpha/2,\ \nu)}$ is found by looking in the t-table under the appropriate value of α against ν = n – 1;

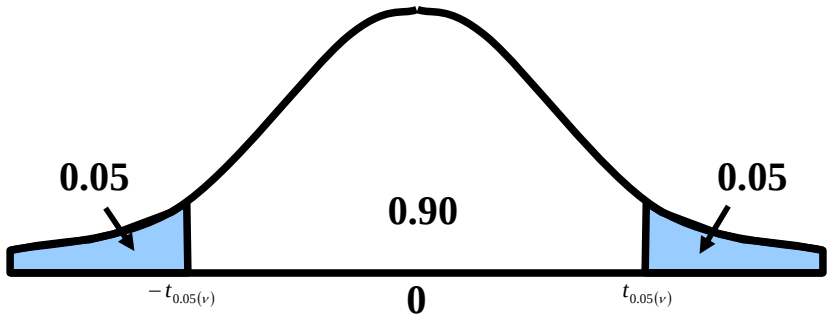
α/2 = **0.005 if we desire 99% confidence:**



α/2 = 0.025 if we desire 95% confidence:



α/2 = 0.05 if we desire 90% confidence:



Next, we discuss hypothesis - testing regarding the mean of a normally distributed population for which σ2 is unknown and the sample size is small (n < 30).

This procedure is illustrated through the following example:

EXAMPLE-1

Just as human height is approximately normally distributed, we can expect the heights of animals of any particular species to be normally distributed. Suppose that, for the past five years, a zoologist has been involved in an extensive research-project regarding the animals of one particular species. Based on his research-experience, the zoologist believes that the average height of the animals of this particular species is 66 centimeters. He selects a random sample of ten animals of this particular species, and, upon measuring their heights, the following data is obtained.
63, 63, 66, 67, 68, 69, 70, 70, 71, 71
In the light of these data, test the hypothesis that the mean height of the animals of this particular species is 66 centimeters.

SOLUTION:

Hypothesis-Testing Procedure:

i) We state our null and alternative hypotheses as
H0 : $\mu = 66$ and H1 : $\mu \neq 66$.

ii) We set the significance level at $\alpha = 0.05$.
iii) Test Statistic:
The test-statistic to be used is

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

which, if H0 is true, has the t-distribution with n – 1 = 9 degrees of freedom.

Important Note:
As indicated in the previous discussion, we always begin by assuming that H0 is true.(The entire mathematical logic of the hypothesis-testing procedure is based on the assumption that H0 is true.)
iv) CALCULATIONS

Individual No.	x_i	x_i^2
1	63	3969
2	63	3969
3	66	4356
4	67	4489
5	68	4624
6	69	4761
7	70	4900
8	70	4900
9	71	5041
10	71	5041
Total	678	46050

Now $\overline{x} = \frac{\sum x_i}{n} = \frac{678}{10} = 67.8$ inches,

and $s^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$

$$= \frac{1}{9}[46050 - 45968.4] = 9.0667,$$

$s = \sqrt{9.0667} = 3.01$ inches.

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

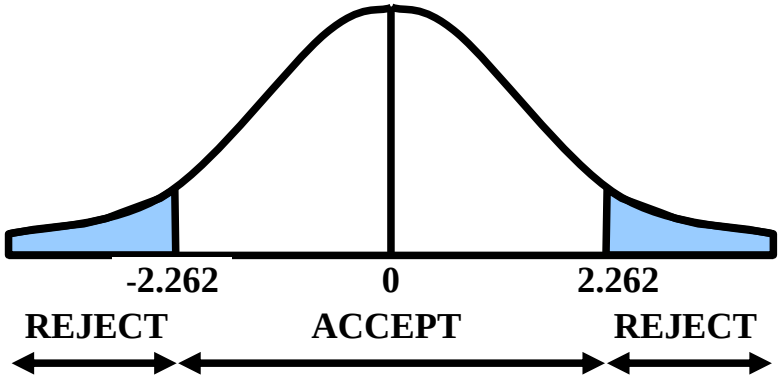
$$= \frac{67.8 - 66}{3.01/\sqrt{10}}$$

$$= \frac{(1.8)(3.1623)}{3.01}$$

$$= 1.89$$

V) Critical Region:

Since this is a two-tailed test, hence the critical region is given by
| t | > t0.025(9) = 2.262.



vi) Conclusion:

Since the computed value of t = 1.89 does not fall in the critical region, we therefore do not reject H0 and may conclude that the mean height of the animals of this particular species is 66 centimeters.

Next, we consider the construction of the confidence interval for μ1-μ2 in that situation when we are drawing small samples from two normally distributed populations having unknown but equal variances:
We illustrate this concept with the help of the following example:

EXAMPLE:

A record company executive is interested in estimating the difference in the average play-length of songs pertaining to pop music and semi-classical music. To do so, she randomly selects 10 semi-classical songs and 9 pop songs.

THE PLAY-LENGTHS (IN MINUTES) OF THE SELECTED SONGS ARE LISTED IN THE FOLLOWING TABLE

Semi-Classic Music	Pop Music
3.80	3.88
3.30	4.13
3.43	4.11
3.30	3.98
3.03	3.98
4.18	3.93
3.18	3.92
3.83	3.98
3.22	4.67
3.38	

Calculate a 99% confidence interval to estimate the difference in population means for these two types of recordings.

SOLUTION:

In this problem, we are dealing with a t-distribution with n1+n2 - 2 = 10 + 9 – 2 = 17 degrees of freedom. The table t-value for a 99% level of confidence and 17 degrees of freedom is t005.17 = 2.898.
Calculations:

Semi-Classical Music	Pop Music
$n_1 = 10$	$n_2 = 9$
$\overline{X}_1 = 3.465$	$\overline{X}_2 = 4.064$
$S_1 = .0.3575$	$S_2 = .0.2417$

Hence:

$$s_p = \sqrt{\frac{(.3575)^2(9)+(.2417)^2(8)}{10+9-2}}$$

$$= \sqrt{\frac{1.1503+0.4674}{17}}$$

$$= \sqrt{\frac{1.6177}{17}} \quad = \sqrt{0.0952}$$

$$= 0.31$$

The confidence interval is

$$(3.465-4.064) \pm (2.898)(0.31)\sqrt{\frac{1}{10}+\frac{1}{9}} = -0.599 \pm 0.411$$

i.e. the C.I. is :

$$-1.010 \le \mu_1 - \mu_2 \le -.188$$

With 99% confidence, the record company executive can conclude that the true difference in population average length of play is between –1.01 minutes and –.188 minute. Zero is not in this interval, so she could conclude that there is a significant difference in the average length of play time between semi-classical music and pop music songs' recordings. Examination of the sample results indicates that pop music songs' recordings are longer. The result and conclusion obtained above can be used in the tactical and strategic planning for programming, marketing, and production of recordings.

EXAMPLE

From an area planted in one variety of guayule (a rubber producing plant), 54 plants were selected at random. Of these, 15 were off types and 12 were aberrant. Rubber percentages for these plants were:

Offtypes	6.21, 5.70, 6.04, 4.47, 5.22, 4.45, 4.84, 5.88, 5.82, 6.09, 6.06, 5.59, 6.74, 5.55
Aberrant	4.28, 7.71, 6.48, 7.71, 7.37, 7.20, 7.06, 6.40, 8.93, 5.91, 5.51, 6.36

Test the hypothesis that the mean rubber percentage of the Aberrants is at least 1 percent more than the mean rubber percentage of off types. Assume the populations of rubber percentages are approximately normal and have equal variances. Let subscript 1 stand for Aberrants, and let subscript 2 stand for off types. Then, we proceed as follows:
i) We formulate our null and alternative hypotheses as
H0 : $\mu 1 - \mu 2 > 1$,
and
H1 : $\mu 1 - \mu 2 < 1$
ii) We set the significance level at $\alpha = 0.05$.
iii) The test-statistic, if H_0 is true, is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a Student's t-distribution with
$\nu = n_1 + n_2 - 2$, i.e. 25 degrees of freedom.

iv) Computations:

We have

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{80.92}{12} = 6.74,$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{84.25}{15} = 5.62,$$

And

$$\sum (x_1 - \bar{x}_1)^2 = \sum x_1^2 - \frac{(\sum x_1)^2}{n_1}$$

$$= 561.6402 - \frac{(80.92)^2}{12}$$

$$= 561.6402 - 545.6705$$

$$= 15.9697$$

$$\sum (x_2 - \bar{x}_2)^2 = \sum x_2^2 - \frac{(\sum x_2)^2}{n_2}$$

$$= 478.9779 - \frac{(84.25)^2}{15}$$

$$= 478.9779 - 473.2042$$

$$= 5.7737$$

$$Now\ s_p^2 = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{5.9697 + 5.7737}{12 + 15 - 2}$$

= 0.8697,

so that

$$s_p = \sqrt{0.8697} = 0.93,$$

Hence, the computed value of our test statistic comes out to be

$$\therefore t = \frac{(6.74 - 5.62) - 1}{0.93\sqrt{\frac{1}{12} + \frac{1}{15}}} = \frac{0.12}{0.36} = 0.33$$

v) Critical Region:

Since this is a left-tailed test, therefore the critical region is given by

$t < -t_{0.05}(25)$ i.e. $t < -1.708$

vi) Conclusion:

Since the computed value of t = 0.33 falls in the acceptance region, therefore we accept H0. We may conclude that the mean rubber percentage of the Aberrants is at least 1 percent more than the mean rubber percentage of Off types.

T-DISTRIBUTION IN THE CASE OF PAIRED OBSERVATIONS

In testing hypotheses about two means, until now we have used independent samples, but there are many situations in which the two samples are not independent. This happen when the observation are found in pairs such that the two observations of a pair are related to each other. Pairing occurs either naturally or by design. Natural pairing occurs whenever measurement is taken on the same unit or individual at two different times. For example, suppose ten young recruits are given a strenuous physical training programme by the Army. Their weights are recorded before they begin and after they complete the training. The two observations obtained for each recruit i.e. the before-and-after measurement constitute natural pairing. The above is natural pairing.

EXAMPLE:
Ten young recruits were put through a strenuous physical training programme by the Army. Their weights were recorded before and after the training with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight before	125	195	160	171	140	201	170	176	195	139
Weight after	136	201	158	184	145	195	175	190	190	145

Using $\alpha = 0.05$, would you say that the programme affects the average weight of recruits?
Assume the distribution of weights before and after to be approximately normal. When the observations from two samples are paired, we find the difference between the two observations of each pair, and the test-statistic in this situation is:

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$$= \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

$$= \frac{\bar{d}}{s_d / \sqrt{n}}$$

LECTURE NO. 41

- Hypothesis-Testing regarding Two Population Means in the Case of Paired Observations (t-distribution)
- The Chi-square Distribution
- Hypothesis Testing and Interval Estimation Regarding a Population Variance (based on Chi-square Distribution)

In the last lecture, we began the discussion of hypothesis-testing regarding two population means in the case of paired observations. It was mentioned that, in many situations, pairing occurs naturally. Observations are also paired to eliminate effects in which there is no interest. For example, suppose we wish to test which of two types (A or B) of fertilizers is the better one. The two types of fertilizer are applied to a number of plots and the results are noted. Assuming that the two types are found significantly different, we may find that part of the difference may be due to the different types of soil or different weather conditions, etc. Thus the real difference between the fertilizers can be found only when the plots are paired according to the same types of soil or same weather conditions, etc.
We eliminate the undesirable sources of variation by taking the observations in pairs. This is pairing by design.
We illustrate the procedure of hypothesis-testing regarding the equality of two population means in the case of paired observations with the help of the same example that we quoted at the end of the last lecture:

EXAMPLE
Ten young recruits were put through a strenuous physical training programme by the Army. Their weights were recorded before and after the training with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight before	125	195	160	171	140	201	170	176	195	139
Weight after	136	201	158	184	145	195	175	190	190	145

Using $\alpha = 0.05$, would you say that the programme affects the average weight of recruits? Assume the distribution of weights before and after to be approximately normal.

SOLUTION

The pairing was natural here, since two observations are made on the same recruit at two different times. The sample consists of 10 recruits with two measurements on each. The test is carried out as below:
Hypothesis-Testing Procedure:

i) We state our null and alternative hypotheses as
H0 : $\mu d = 0$ and
H1 : $\mu d \neq 0$

ii) The significance level is set at $\alpha = 0.05$.

iii) The test statistic under H0 is

$$t = \frac{\bar{d}}{s_d / \sqrt{n}},$$

which has a t-distribution with n – 1 degrees of freedom.

iv) Computations:

Recruit	Weight		Difference, d_i (after minus before)	d_1^2
	Before	After		
1	125	136	11	121
2	195	201	6	36
3	160	158	– 2	4
4	171	184	13	169
5	140	145	5	25
6	201	195	6	36
7	170	175	– 6	25
8	176	190	5	196
9	195	190	14	25
10	139	145	– 5	36
∑	1672	1719	6	673

$$\bar{d} = \frac{\sum d}{n} = \frac{47}{10} = 4.7,$$

$$s_{d^2} = \frac{\sum (d - \bar{d})^2}{n-1} = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$$

$$= \frac{1}{9}\left[673 - \frac{(47)^2}{10}\right] = \frac{673 - 220.9}{9} = 50.23,$$

so that

$$s_d = \sqrt{50.23} = 7.09.$$

Hence, the computed value of our test-statistic comes out to be :

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{4.7}{7.09/\sqrt{10}} = \frac{(4.7)(3.16)}{7.09} = 2.09.$$

v) The critical region is $|t| \geq t_{0.025}(9)$
$= 2.262.$

vi) Conclusion:

Since the calculated value of t = 2.09 does not fall in the critical region, so we accept H0 and may conclude that the data do not provide sufficient evidence to indicate that the programme affects average weight.

From the above example, it is clear that the hypothesis-testing procedure regarding the equality of means in the case of paired observations is very similar to the t-test that is applied for testing
H0 : $\mu = 0$.(The only difference is that when we are testing H0 : $\mu = 0$, our variable is X, whereas when we are testing H0 : $\mu d=0$, our variable is d.)

HYPOTHESIS-TESTING PROCEDURE REGARDING TWO POPULATIONS MEANS IN THE CASE OF PAIRED OBSERVATIONS

When the observations from two samples are paired either naturally or by design, we find the difference between the two observations of each pair. Treating the differences as a random sample from a normal population with mean $\mu d = \mu 1 - \mu 2$ and unknown standard deviation σd, we perform a one-sample t-test on them. This is called a paired difference t-test or a paired t-test.

Testing the hypothesis

H0 : $\mu 1 = \mu 2$ against
HA : $\mu 1 \neq \mu 2$ is equivalent to testing H0 : $\mu d = 0$ against
HA : $\mu d \neq 0$.
Let d = x1 – x2 denote the difference between the two samples observations in a pair. Then the sample mean and standard deviation of the differences are

$$\bar{d} = \frac{\sum d}{n} \text{ and } s_d = \frac{\sum (d - \bar{d})^2}{n-1}$$

where n represents the number of pairs.
Assuming that

1) $d_1, d_2, \ldots, d_n$ is a random sample of differences, and

2) the differences are normally distributed,
the test-statistic

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

follows a t-distribution with $\nu = n - 1$ degrees of freedom. The rest of the procedure for testing the null hypothesis H0 : $\mu d = 0$ is the same

EXAMPLE

The following data give paired yields of two varieties of wheat.

Variety I	45	32	58	57	60	38	47	51	42	38
Variety II	47	34	60	59	63	44	49	53	46	41

Each pair was planted in a different locality.

a) Test the hypothesis that, on the average, the yield of variety-1 is less than the mean yield of variety-2. State the assumptions necessary to conduct this test.

b) How can the experimenter make a Type-I error? What are the consequences of his doing so?

c) How can the experimenter make a Type-II error? What are the consequences of his doing so?

d) Give 90 per cent confidence limits for the difference in mean yield.

Note: The pairing was by design here, as the yields are affected by many extraneous factors such as fertility of land, fertilizer applied, weather conditions and so forth.

SOLUTION:

a) In order to conduct this test, we make the following assumptions:

ASSUMPTIONS

- The differences in yields are a random sample from the population of differences,
- The population of differences is normally distributed.

i) We state our null and alternative hypotheses as

$$H_0 : \mu_d \geq 0 \ \left(\text{or} \ \ \mu_1 \geq \mu_2\right),$$

i.e. the mean yields are equal and

$$H_1 : \mu_d < 0 \left(\text{or} \ \ \mu_1 < \mu_2\right).$$

ii) We select the level of significance at $\alpha = 0.05$.

iii) The test statistic to be used is

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where $\bar{d} = \bar{x}_1 - \bar{x}_2$ and sd2 is the variance of the differences di.

If the populations are normal, this statistic, when H0 is true, has a Student's t-distribution with (n – 1) d. f.

iv) Computations:

Let X1i and X2i represent the yields of Variety I and Variety II respectively. Then the necessary computations are given below:

X_{1i}	X_{2i}	$d_i = X_{1i} - X_{2i}$	d_i^2
45	47	–2	4
32	34	–2	4
58	60	–2	4
57	59	–2	4
60	63	–3	9
38	44	–6	36
47	49	–2	4
51	53	–2	4
42	46	–4	16
38	41	–3	9
Σ	—	–28	94

Now $\bar{d} = \frac{\sum d_i}{n} = \frac{-28}{10} = -2.8$, and

$$s_d{}^2 = \frac{1}{n-1}\left[\sum d_i{}^2 - \frac{(\sum d_i)^2}{n}\right] = \frac{1}{9}\left[94 - \frac{(-28)^2}{10}\right]$$

$$= \frac{15.6}{9} = 1.7333, \text{ so that } s_d = 1.32$$

$$\therefore \quad t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{-2.8}{1.32/\sqrt{10}} = \frac{(-2.8)(3.1623)}{1.32} = -6.71$$

v) As this is a one-tailed test therefore, the critical region is given by

$t < t_{0.05(9)} = -1.833$

vi) Conclusion

Since the calculated value of t = –6.71 falls in the critical region, we therefore reject H0. The data present sufficient evidence to conclude that the mean yield of variety-1 is less than the mean yield of variety-2.

b) The experimenter can make a Type-I error by rejecting a true null hypothesis.

In this case, the Type-I error is make by rejecting the null hypothesis when the mean yield of variety-1 is actually not different from the mean yield of variety-2.

In so doing, the consequences would be that we will be saying that variety-2 is better than variety-1 although in reality they are equally good.

c) The experimenter can make a Type-II error by accepting of false null hypothesis.

In this case, the Type-II error is made by accepting the null hypothesis when in reality the mean yield of variety-1 is less than the mean yield of variety-2 and the consequence of committing this error would be a loss of potential increased yield by the use of variety-2.

d) The 90% confidence limits for the difference in means $\mu 1 - \mu 2$ in case of paired observations, are given by

$$\bar{d} \pm t_{\alpha/2,(n-1)} \cdot \frac{s_d}{\sqrt{n}}$$

Substituting the values, we get

$$-2.8 \pm 1.833 \ \frac{1.32}{\sqrt{10}}$$

or -2.8 + 0.765

or -3.565 to -2.035

Hence the 90% confidence limits for the difference in mean yields, $\mu 1 - \mu 2$, are (-3.6, -2.0) .

Until now, we have discussed statistical inference regarding population means based on the Z-statistic as well as the t-statistic.

Also, we have discussed inference regarding the population proportion based on the Z-statistic.

In certain situations, we would be interested in drawing conclusions about the variability that exists in the population values, and for this purpose, we would like to carry out estimation or hypothesis-testing regarding the population variance $\sigma 2$.

Statistical Inference regarding the population variance is based on the chi-square distribution.

We begin this topic by presenting the formal definition of the Chi-Square distribution and stating some of its main properties:

THE CHI-SQUARE ($\chi 2$) DISTRIBUTION

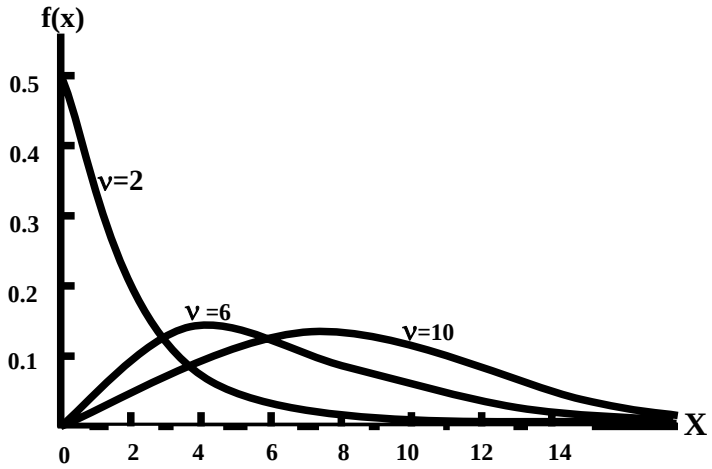
The mathematical equation of the Chi-Square distribution is as follows:

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(x)^{(\nu/2)-1}.e^{-x/2}, \ 0 < x < \infty$$

This distribution has only one parameter ν, which is known as the degrees of freedom of the Chi-Square distribution.

PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The Chi-Square ($\chi 2$) distribution has the following properties:
1. It is a continuous distribution ranging from 0 to $+\infty$.
The number of degrees of freedom determines the shape of the chi-square distribution. Thus there is a different chi-square distribution for each number of degrees of freedom. As such, it is a whole family of distributions.
2. The curve of a chi-square distribution is positively skewed.
The skewness decreases as ν increases.



χ2-distribution for various values of ν

As indicated by the above figures, the chi-square distribution tends to the normal distribution as the number of degrees of freedom approaches infinity.
3. The mean of a chi-square distribution is equal to ν, the number of degrees of freedom.
4. Its variance is equal to 2ν.
5. The moments about the origin are given by

$$\mu'_1 = \nu$$
$$\mu'_2 = \nu(\nu+2)$$
$$\mu'_3 = \nu(\nu+2)(\nu+4)$$
$$\mu'_4 = \nu(\nu+2)(\nu+4)(\nu+6)$$

As such, the moment-ratios come out to be

$$\beta_1 = \frac{8}{\nu}$$
$$\beta_1 = 3 + \frac{12}{\nu}$$

Having discussed the basic definition and properties of the chi-square distribution, we begin the discussion of its role in interval estimation and hypothesis-testing. We begin with interval estimation regarding the variance of a normally distributed population:

EXAMPLE:

Suppose that an aptitude test carrying a total of 20 marks is devised, and administered on a large population of students, and, upon doing so, it was found that the marks of the students were normally distributed. A random sample of size n = 8 is drawn from this population, and the sample values are 9, 14, 10, 12, 7, 13, 11, 12.
Find the 90 percent confidence interval for the population variance $\sigma 2$, representing the variability in the marks of the students.

SOLUTION:

The 90% confidence interval for σ2 is given by

$$\frac{\sum(\mathrm{Xi}-\overline{\mathrm{X}})^2}{\chi^2_{0.05(n-1)}} < \sigma^2 < \frac{\sum(\mathrm{X_i}-\overline{\mathrm{X}})^2}{\chi^2_{0.95(n-1)}}$$

The above formula is linked with the fact that if we keep 90% area under the chi-square distribution in the middle, then we will have 5% area on the left-hand-side, and 5% area on the right-hand-side, as shown below:

χ2(N-1)-DISTRIBUTION

In order to apply the above formula, we first need to calculate the sample mean $\overline{X}$, which is

$$\overline{\mathrm{X}} = \frac{\sum \mathrm{X}}{\mathrm{n}} = \frac{88}{8} = 11$$

Then, we obtain

$$\sum_{i=1}^{8} \left(X_i - \overline{X}\right)^2 = (9-11)^2 + (14-11)^2 + \ldots + (12-11)^2 = 36$$

Next, we need to find :
1) the value of χ2 to the left of which the area under the chi-square distribution is 5%
2) the value of χ2 to the right of which the area under the chi-square distribution is 5%
For this purpose, we will need to consult the table of areas under the chi-square distribution.

THE CHI-SQUARE TABLE

The entries in this table are values of x2α(ν), for which the area to their right under the chi-square distribution with ν degrees of freedom is equal to α.

Upper Percentage Points of the Chi-square Distribution									
α / ν	**0.99**	**0.98**	**0.975**	**0.95**	**0.10**	**0.05**	**0.03**	**0.02**	**0.01**
1	0.0002	0.001	0.001	0.004	2.71	3.84	5.02	5.41	6.64
2	0.020	0.040	0.051	0.103	4.61	5.99	7.38	7.82	9.21
3	0.115	0.185	0.216	0.352	6.25	7.82	9.35	9.84	11.34
4	0.297	0.429	0.484	0.711	7.78	9.49	11.14	11.67	13.28
5	0.554	0.752	0.831	1.145	9.24	11.07	12.83	13.39	15.09
6	0.87	1.13	1.24	1.64	10.64	12.59	14.45	15.03	16.81
7	1.24	1.56	1.69	2.17	12.02	14.07	16.01	16.62	18.48
8	1.65	2.03	2.18	2.73	13.36	15.51	17.54	18.17	20.09
9	2.09	2.53	2.70	3.32	14.68	16.92	19.02	19.68	21.67
10	2.56	3.06	3.25	3.94	15.99	18.31	20.48	21.16	23.21
11	3.05	3.61	3.82	4.58	17.28	19.68	21.92	22.62	24.72
12	3.57	4.18	4.40	5.23	18.55	21.03	23.34	24.05	26.22
13	4.11	4.76	5.01	5.89	19.81	22.36	24.74	25.47	27.69
14	4.66	5.37	5.63	6.57	21.06	23.68	26.12	26.87	29.14
15	5.23	5.98	6.26	7.26	22.31	25.00	27.49	28.26	30.58

Chi-Square Table (continued):

	α								
ν	**0.99**	**0.98**	**0.975**	**0.95**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**
16	5.81	6.61	6.91	7.96	23.54	26.30	28.84	29.63	32.00
17	6.41	7.26	7.56	8.67	24.77	27.59	30.19	31.00	33.41
18	7.02	7.91	8.23	9.39	25.99	28.87	31.53	32.35	34.81
19	7.63	8.57	8.91	10.12	27.20	30.14	32.85	33.69	36.19
20	8.26	9.24	9.59	10.85	28.41	31.41	34.17	35.02	37.57
21	8.90	9.92	10.28	11.59	29.62	32.67	35.48	36.34	38.93
22	9.54	10.60	10.98	12.34	30.81	33.92	36.78	37.66	40.29
23	10.20	11.29	11.69	13.09	32.01	35.17	38.08	38.97	41.64
24	10.86	11.99	12.40	13.85	33.00	36.42	39.36	40.27	42.92
25	11.52	12.70	13.12	14.61	34.38	37.65	40.65	41.57	44.31
26	12.20	13.41	13.84	15.38	35.56	38.88	41.92	42.86	45.64
27	12.88	14.12	14.57	16.15	36.74	40.11	43.19	44.14	46.96
28	13.56	14.85	15.31	16.93	37.92	41.34	44.46	45.42	48.28
29	14.26	15.57	16.05	17.71	39.09	42.56	45.72	46.69	49.59
30	14.95	16.31	16.79	18.49	40.26	43.77	46.98	47.96	50.89

From the $\chi 2$-table, we find that

$\chi^2_{0.05}(7) = 14.07$

and

$\chi^2_{0.95}(7) = 2.17$

Hence the 90 percent confidence interval for $\sigma 2$ is

$$\frac{\sum(Xi - \overline{X})^2}{\chi^2_{0.05,(7)}} < \sigma^2 < \frac{\sum(X_i - \overline{X})^2}{\chi^2_{0.95,(7)}}$$

or $$\frac{36}{14.07} < \sigma^2 < \frac{36}{2.17}$$

or $$2.56 < \sigma^2 < 16.61$$

Thus the 90% confidence interval for σ^2 is (2.56, 16.61).

If we take the square root of the lower limit as well as the upper limit of the above confidence interval, we obtain (1.6, 4.1).

So, we may conclude that, on the basis of 90% confidence, we can say that the standard deviation σ of our population lies between 1.6 and 4.1 .We can obtain a confidence interval for σ by taking the square root of the end points of the interval for $\sigma 2$, but experience has shown that σ cannot be estimated with much precision for small sample sizes.

The formula of the confidence interval for $\sigma 2$ that we have applied in the above example is based on the fact that:

If $\overline{X}$ and S2 are the mean and variance (respectively) of a random sample X1, X2, …, Xn of size n drawn from a normal population with variance $\sigma 2$, then the statistic

$$\frac{\sum(X_i - \overline{X})^2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with (n – 1) degrees of freedom.

Next, we consider hypothesis - testing regarding the population variance $\sigma 2$:

We illustrate this concept with the help of an example:

EXAMPLE

The variability in the tensile strength of a type of steel wire must be controlled carefully. A sample of the wire is subjected to test, and it is found that the sample variance is S2 = 31.5. The sample size was n = 16 observations.

Test the hypothesis that the population variance is 25 against the alternative that the variance is greater than 25. Use a 0.05 level of significance.

SOLUTION

a)i) We have to decide between the hypotheses

H0 : σ2 = 25, and
H1 : σ2 > 25

ii) The level of significance is $\alpha = 0.05$.

iii) The test statistic is $\chi^2 = \frac{nS^2}{\sigma_0^2}$, which under H_0, has a χ^2-distribution with (n–1) degrees of freedom, assuming that the population is normal.

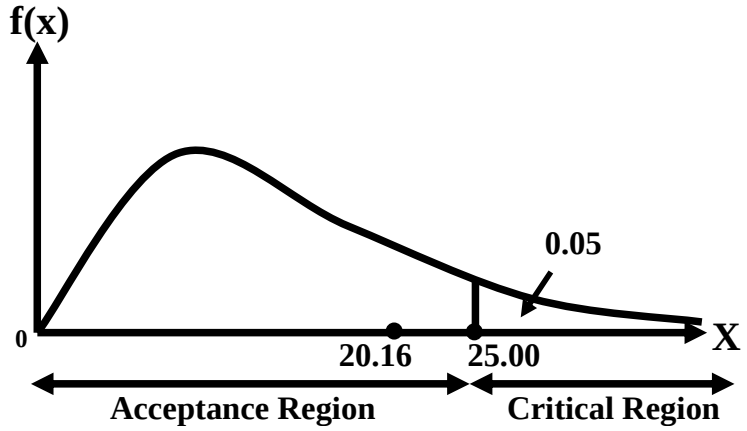
iv) We calculate the value of χ^2 from the sample data as

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{16(31.5)}{25} = 20.16 .$$

v) The critical region is $\chi 2 > \chi^2_{0.05,(15)} = 25.00$ (one tailed test)

vi) Conclusion.

Since the calculated value of χ2 falls in the acceptance region, so we accept our null Hypothesis, i.e. we have reasonable evidence to conclude that σ2 = 25.The Chi-Square Distribution with 15 degrees of Freedom:



The above example points to the following general procedure for testing a hypothesis regarding the population variance σ2: Suppose we desire to test a null hypothesis H0 that the variance σ2 of a normally distributed population has some specified value, say σ02. To do this, we need to draw a random sample $X_1, X_2, \ldots, X_n$ of size n from the normal population and compute the value of the sample variance S^2. If the null hypothesis $H_0 : \sigma^2 = \sigma^2_0$ is true, then the statistic $\chi^2 = \frac{nS^2}{\sigma_0^2}$ has a χ^2-distribution with (n–1) degrees of freedom.

LECTURE NO. 42

- The F-Distribution
- Hypothesis Testing and Interval Estimation in order to Compare the Variances of Two Normal Populations (based on F-Distribution)

Before we describe you statistical inference based on the F-distribution, let us consolidate the idea of hypothesis-testing regarding the population variance with the help of an example:

EXAMPLE

The manager of a bottling plant is anxious to reduce the variability in net weight of fruit bottled. Over a long period, the standard deviation has been 15.2 gm. A new machine is introduced and the net weights (in grams) in 10 randomly selected bottles (all of the same nominal weight) are 987, 966, 955, 977, 981, 967, 975, 980, 953, 972. Would you report to the manager that the new machine has a better performance?

SOLUTION

i) We have to decide between the hypotheses

H0 : $\sigma = 15.2$, i.e. the standard deviation is 15.2gm

H1 : $\sigma < 15.2$ i.e. the standard deviation has been reduced.

ii) We choose the significance level at $\alpha = 0.05$.

iii) The test-statistic is

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2}$$

which under H_0, has a χ^2 -distribution with (n – 1) degrees of freedom, assuming that the weights are normally distributed.

iv) Computations.

n = 10, $\sum Xi = 9713$, $\sum X2i = 9435347$

v) The critical region is $\chi 2 < \chi 20.95\ (9) = 3.32$ (the lower 5% point)

NOW

$nS^2 = \sum(X_i - \bar{X})^2 = \sum X_i^2 - (\sum X_i)^2/n$

$= 9435347 - (9713)^2/10 = 1110.1$

$$\therefore \quad \chi^2 = \frac{1110.1}{(15.2)^2} = \frac{1110.1}{231.04} = 4.81$$

vi) Conclusion:

Since the calculated value of $\chi^2 = 4.81$ does not fall in the critical region, we therefore cannot reject the null hypothesis that the standard deviation is 15.2 gm and hence we would not report to the manager that the new machine has a better performance.

The above example points to the fact that, if we wish to test a null hypothesis H0 that the variance $\sigma 2$ of a normally distributed population has some specified value, say $\sigma 02$, then, (having drawn a random sample X1, X2, …, Xn of size n from the normal population), we will compute the value of the sample variance S2.

The mathematics underlying this hypothesis-testing procedure states that:

If the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ is true, then the statistic $\chi^2 = \frac{nS^2}{\sigma_0^2}$ has a χ^2-distribution with (n–1) degrees of freedom.

A point to be noted is that, since the random variable X is distributed as chi-square, therefore we call it $\chi 2$.

If we do so, our equation of the chi-square distribution can be written as

$$f(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{(\nu/2)-1}.e^{-\chi^2/2},\ 0 < \chi^2 < \infty$$

It should be obvious that the standard deviation of the normal population will be tested in the same way as the population variance is tested.

Next, we begin the discussion of statistical inference regarding the ratio of two population variances.

As this particular inference is based on the F-distribution, therefore we begin with the discussion of the mathematical definition and the main properties of the F-distribution.

THE F-DISTRIBUTION

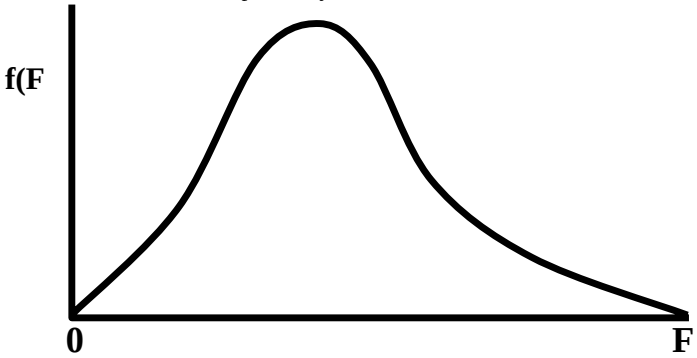
The mathematical equation of the F-distribution is as follows:

$$f(x)=\frac{\Gamma\left[\left(\nu_1+\nu_2\right)/2\right]\left(\nu_1/\nu_2\right)^{\nu_1/2}\ x^{(\nu_1/2)-1}}{\Gamma\left(\nu_1/2\right)\Gamma\left(\nu_2/2\right)\left[1+\nu_1\,x/\nu_2\right]^{(\nu_1+\nu_2)/2}},\quad 0<x<\infty$$

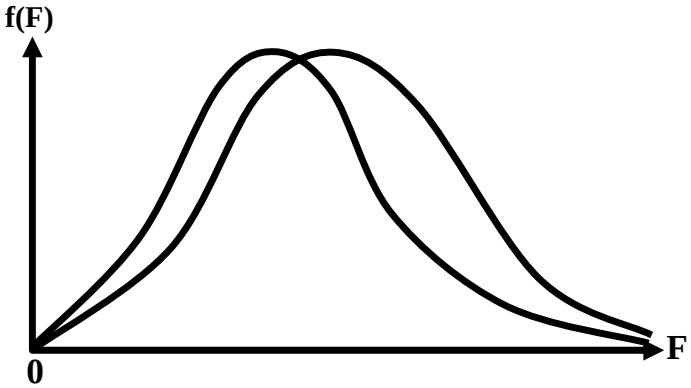
This distribution has two parameters ν1 and ν2, which are known as the degrees of freedom of the F-distribution.The F-distribution having the above equation have ν1 degrees of freedom in the numerator and ν2 degrees of freedom in the denominator. It is usually abbreviated as F(ν1, ν2).

PROPERTIES OF F-DISTRIBUTION

1. The F-distribution is a continuous distribution ranging from zero to plus infinity.
2. The curve of the F-distribution is positively skewed.



But as the degrees of freedom ν1 and ν2 become large, the F-distribution approaches the normal distribution.



3. For ν2 > 2, the mean of the F-distribution is

$$\frac{\nu_2}{\nu_2-2}$$

which is greater than 1.
4.For ν2 > 4, the variance of the F-distribution is

$$\sigma^2=\frac{2\nu_2^2\left(\nu_1+\nu_2-2\right)}{\nu_1\left(\nu_2-2\right)^2\left(\nu_2-4\right)}$$

5. The F-distribution for ν1 > 2, ν2 > 2 is unimodal, and the mode of the distribution with ν1 > 1 is at

$$\frac{\nu_2\left(\nu_1-2\right)}{\nu_1\left(\nu_2+2\right)}$$

which is always less than 1.
6. If F has an F-distribution with ν1 and ν2 degrees of freedom, then the reciprocal has an F-distribution with ν2 and ν1 degrees of freedom. Next, we consider the tables of the F-distribution. As the F-distribution involves two parameters,

ν1 and ν2, hence separate tables have been constructed for 5%, 2½ % and 1% right-tail areas respectively, as shown below:

The F-table pertaining to 5% right-tail areas is as follows:

Upper 5 Percent Points of The F-Distribution i.e., $F_{0.05}(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.73	1.52	1.00

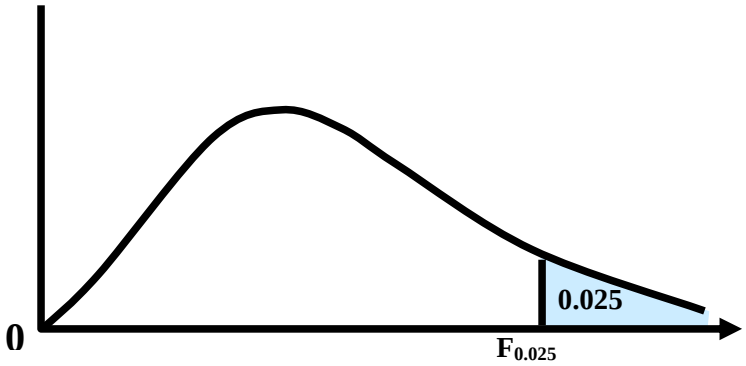
Similarly, the F-table pertaining to 2½% right-tail areas is as follows

Upper 2.5 Percent Points of the F-Distribution i.e. $F_{0.025}(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	956.7	976.7	997.2	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.37	39.41	39.46	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.54	14.34	14.12	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	8.98	8.75	8.51	8.26
5	10.07	8.43	7.76	7.39	7.15	6.98	6.76	6.52	6.28	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.60	5.37	5.12	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.90	4.67	4.42	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.43	4.20	3.95	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.10	3.87	3.61	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.85	3.62	3.37	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.66	3.43	3.17	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.51	3.28	3.02	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.39	3.15	2.89	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.29	3.05	2.79	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.20	2.96	2.70	2.40

Upper 2.5 Percent Points of the F-Distribution i.e. $F_{0.025}(\nu_1, \nu_2)$ (Continued):

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
16	6.12	4.69	4.08	3.73	3.50	3.34	3.12	2.89	2.63	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.06	2.82	2.56	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.01	2.77	2.50	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	2.96	2.72	2.45	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	2.91	2.68	2.41	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.87	2.64	2.37	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.84	2.60	2.33	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.81	2.57	2.30	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.78	2.54	2.27	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.75	2.51	2.24	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.73	2.49	2.22	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.71	2.47	2.19	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.69	2.45	2.17	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.67	2.43	2.15	1.81
30	5.57	4.18	3.59	3.25	3.06	2.87	2.65	2.41	2.14	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.53	2.29	2.01	1.64
60	5.49	3.93	3.34	3.01	2.79	2.63	2.41	2.17	1.88	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.30	2.05	1.76	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.19	1.94	1.64	1.00



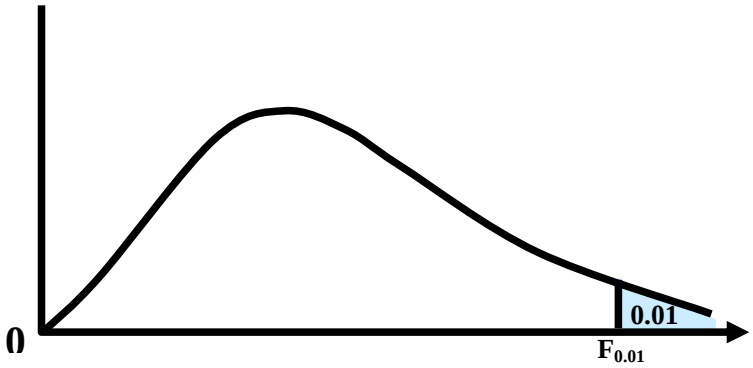
And, the F-table pertaining to 1% right-tail areas is as follows:

Upper 1 Percent Points of the F-Distribution i.e. $F_{0.01}(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	4052	5000	5403	5625	5764	5859	5982	6106	6235	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	10.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.61
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87

Upper 1 Percent Points of the F-Distribution i.e. $F_{0.01}(\nu_1, \nu_2)$ (continued)

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.03	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.94	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00



Having discussed the basic definition and the main properties of the F-distribution, we now begin the discussion of the role of the F-distribution in statistical inference: First, we discuss interval estimation regarding the ratio of two population variances:

CONFIDENCE INTERVAL FOR THE VARIANCE RATIO σ12/σ22

Let two independent random samples of size n1 and n2 be taken from two normal population with variances σ12 and σ22 and let s12 and s22 be the unbiased estimators of σ12 and σ22.
Then, it can be mathematically proved that the quantity

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

has an F-distribution with (n1 – 1, n2 – 1) degrees of freedom.
The confidence interval for σ12/σ22 is given by

$$\left[\frac{s_1^2}{s_2^2} . \frac{1}{F_{\alpha/2}(n_1 - 1,\ n_2 - 1)},\ \frac{s_1^2}{s_2^2} . F_{\alpha/2}(n_2 - 1,\ n_1 - 1) \right]$$

We can also find a confidence interval for σ1/σ2 by taking the square root of the end points of the above interval.
We illustrate this concept with the help of the following example:

EXAMPLE

A random sample of 12 salt-water fish was taken, and the girth of the fish was measured. The standard deviation s1 came out to be 2.3 inches. Similarly, a random sample of 10 fresh-water fish was taken, and the girth of the fish was measured. The standard deviation of this sample i.e. s2 came out to be 1.5 inches. Find a 90% confidence interval for the ratio between the 2 population variances σ12/σ22. Assume that the populations of girth are normal.

SOLUTION

The 90% confidence interval for σ12/σ22 is given by

$$\left[\frac{s_1^2}{s_2^2} . \frac{1}{F_{0.05}(n_1 - 1,\ n_2 - 1)},\ \frac{s_1^2}{s_2^2} . F_{0.05}(n_2 - 1,\ n_1 - 1) \right]$$

Here s12 = (2.3)2 = 5.29,
s22 = (1.5)2 = 2.25,
n1– 1 = 12 – 1 = 11 and n2 – 1 = 10 – 1 = 9
Hence,
F0.05 (n1 – 1, n2 – 1) = F0.05 (11, 9) = 3.1
and
F0.05 (n2 – 1, n1 – 1) = F0.05 (9, 11) = 2.9
With reference to the F-table, it should be noted that if it is an abridged table and the F-values are not available for all possible pairs of degrees of freedom, then the required F-values are obtained by the method of interpolation. In this example, for the lower limit of our confidence interval, we need the value of F0.05(11, 9), but in the above table pertaining to 5% right-tail area, values are available for ν1 = 8 and ν1 = 12, but not for ν1 = 11. Hence, we can find the F-value corresponding to ν1 = 11 by the method of interpolation: The F-value corresponding to ν2 = 9 and ν1 = 8 is 3.23 whereas the F-value corresponding to ν2 = 9 and ν1 = 12 is 3.07.If we wish to find the F-value corresponding to ν2 = 9 and ν1 = 10, we can find the arithmetic mean of 3.23 and 3.07 which is 3.15.If we wish to find the F-value corresponding to ν2 = 9 and ν1 = 11, we can find the arithmetic mean of 3.15 and 3.07 which is 3.11, which, upon

rounding, is equal to 3.1.The above method of interpolation is based on the assumption that the F-values between any two successive F-values (printed in any row of the F-table) are equally spaced between the two given values.
If we do not wish to go through the rigorous procedure of interpolation, we can note that ν1 = 11 is close to ν1 = 12, and hence, we can consider that F-value which corresponds to ν1 = 12 (which in this case is 3.07 ~ 3.1 ----- exactly the same as what we obtained above (correct to one decimal place) by the method of interpolation). Going back to our example, the 90% confidence interval is

$$\left[\frac{5.29}{2.25}.\left(\frac{1}{3.1}\right), \frac{5.29}{2.25}(2.9)\right]$$

or (0.76, 6.81).

Taking the square root of the end points (0.76, 6.81), we obtain the 90% confidence interval for σ1/σ2 as (0.87, 2.61).
Next, we discuss hypothesis - testing regarding the equality of two population variances: Suppose that we have two independent random samples of size n1 and n2 from two normal populations with variances σ12 and σ22, we wish to test the hypothesis that the two variances are equal. The main steps of the hypothesis - testing procedure are similar to the ones that we have been discussing earlier. We illustrate this concept with the help of an example:

EXAMPLE

In two series of hauls to determine the number of plankton organisms inhabiting the waters of a lake, the following results were found:
Series I: 80, 96, 102, 77, 97, 110, 99, 88, 103, 1089
Series II: 74, 122, 92, 81, 104, 92, 92
In series I, the hauls were made in succession at the same place. In series II, they were made in different parts scattered over the lake. Does there appear to be a greater variability between different places than between different times at the same place?

SOLUTION

If X denotes the number of plankton organisms per haul, then for each of the two series, X can be assumed to be normally distributed.

Hypothesis-testing Procedure:

Step 1 :

H0 : $\sigma_1^2 \geq \sigma_2^2$ i.e. $\sigma_2^2 \leq \sigma_1^2$

HA : $\sigma_1^2 < \sigma_2^2$ i.e. $\sigma_2^2 > \sigma_1^2$

Step 2: Level of significance: $\alpha = 0.05$

Step 3: Test-statistic:

Since both the populations are normally distributed, hence, the statistic

$$F = \frac{s_2^2 / \sigma_2^2}{s_1^2 / \sigma_1^2}$$

will follow the F-distribution having (n2 - 1, n1 - 1) degrees of freedom.

Step 4 : Computations:

X_1	X_1^2
80	6400
96	9216
102	10404
77	5929
97	9409
110	12100
99	9801
88	7744
103	10609
108	11664
960	93276

X_2	X_2^2
74	5476
122	14884
92	8464
81	6561
104	10816
92	8464
92	8464
657	63129

Now

$$s_1^2 = \frac{1}{n_1 - 1}\left[\sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n_1}\right]$$

and

$$s_2^2 = \frac{1}{n_2 - 1}\left[\sum X_2^2 - \frac{(\sum X_2)^2}{n_2}\right]$$

So $$s_1^2 = \frac{1}{10-1}\left[9326 - \frac{(960)^2}{10}\right]$$

$$= \frac{1}{9}[93276 - 92160]$$

$$= \frac{1}{9}[1116] = 124$$

Similarly $$s_2^2 = \frac{1}{n_2 - 1}\left[\sum X_2^2 - \frac{(\sum X_2)^2}{n_2}\right]$$

$$= \frac{1}{7-1}\left[63129 - \frac{(657)^2}{7}\right]$$

$$= \frac{1}{6}[63129 - 61664.14]$$

$$= \frac{1}{6}[1464.86] = 244.14$$

Hence $$F = \frac{s_2^2}{s_1^2} = \frac{244.14}{124} = 1.97$$

Step 5 : Critical Region:

F > F0.05 (6, 9) = 3.37

Step 6: Conclusion:

Since 1.97 is less than 3.37, we do not reject H0; our data does not provide sufficient evidence to indicate that there is greater variability (in the number of plankton organisms per haul) between different places than between different times at the same place.

Let us consider another example:

EXAMPLE

Two methods of determining the moisture content of samples of canned corn have been proposed and both have been used to make determinations on proportions taken from each of 21 cans. Method I is easier to apply but appears to be more variable than Method II.

If the variability of Method I were not more than 25 per cent greater than that of Method II, then we would prefer Method I.

The sample results are as follows:

$$n_1 = n_2 = 21;\ \overline{X}_1 = 50;\ \overline{X}_2 = 53$$

$$\sum(X_1 - \overline{X}_1)^2 = 720; \qquad \sum(X_2 - \overline{X}_2)^2 = 340.$$

Based on the above sample results, which method would you recommend?

SOLUTION

In order to solve this problem, the first point to be noted is that, in this problem, our null and alternative hypotheses will be

H0: $\sigma 12 \leq 1.25\ \sigma 22$

and

H1: $\sigma 12 > 1.25\ \sigma 22$.

Null and Alternative Hypotheses:

In this problem, we need to test
H0 : σ12 ≤ 1.25 σ22
against
H1 : σ12 > 1.25 σ22.
This is so, because 1.25 σ22 means 125% of σ22, and this means 25% greater than σ22. You are encouraged to work on this point on their own. The second point to be noted is that, in this problem, our test-statistic is not but is

$$F = \frac{s_1^{\,2}}{1.25\ s_2^{\,2}}.$$

Test Statistic:

$$F = \frac{s_1^{\,2}}{1.25\ s_2^{\,2}}.$$

(Under the null hypothesis, s12 / 1.25 s22 has an F-distribution with ν1 = ν2 = 21-1 = 20 degrees of freedom.)
This is so because, (in accordance with the fact that has an F-distribution with (n1 – 1, n2 – 1) degrees of freedom), it can be shown that:

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

If we have

H0: σ12/ σ22 = k

then

$$F = \frac{s_1^{\,2}}{s_2^{\,2}} \cdot \frac{1}{k}$$

has an F-distribution with (n1 - 1, n2 - 1) degrees of freedom. (In this problem, k = 1.25.)You are encouraged to work on this problem also on their own, and to carry out the rest of the steps of the hypothesis-testing procedure (which are the usual ones), and to decide whether to accept or to reject the null hypothesis.

LECTURE NO. 43

- Analysis of Variance
- Experimental Design

Earlier, we compared two-population means by using a two-sample t-test. However, we are often required to compare more than two population means simultaneously. We might be tempted to apply the two-sample t-test to all possible pairwise comparisons of means. For example, if we wish to compare 4 population means, there will be $\binom{4}{2} = 6$ separate pairs, and to test the null hypothesis that all four population means are equal, we would require six two-sample t-tests. Similarly, to test the null hypothesis that 10 population means are equal, we would need

$$\binom{10}{2} = 45$$

Separate two-sample t-tests. This procedure of running multiple two-sample t-tests for comparing means would obviously be tedious and time-consuming. Thus a series of two-sample t-tests is not an appropriate procedure to test the equality of several means simultaneously. Evidently, we require a simpler procedure for carrying out this kind of a test. One such procedure is the Analysis of Variance, introduced by Sir R.A. Fisher (1890-1962) in 1923:

ANALYSIS OF VARIANCE (ANOVA)

It is a procedure which enables us to test the hypothesis of equality of several population means
(i.e.
H0 : $\mu 1 = \mu 2 = \mu 3 = \ldots\ldots = \mu k$
against
HA: not all the means are equal)
The concept of Analysis of Variance is closely related with the concept of Experimental Design:

EXPERIMENTAL DESIGN

By an experimental design, we mean a plan used to collect the data relevant to the problem under study in such a way as to provide a basis for valid and objective inference about the stated problem. The plan usually includes:

- the selection of treatments whose effects are to be studied,
- the specification of the experimental layout, and
- the assignment of treatments to the experimental units.

All these steps are accomplished before any experiment is performed. Experimental Design is a very vast area. In this course, we will be presenting only a very basic introduction of this area. There are two types of designs:

SYSTEMATIC AND RANDOMIZED DESIGNS

In this course, we will be discussing only the randomized designs, and, in this regard, it should be noted that for the randomized designs, the analysis of the collected data is carried out through the technique known as Analysis of Variance.
Two of the very basic randomized designs are:

- The Completely Randomized (CR) Design,
- The Randomized Complete
- Block (RCB) Design

We will consider these one by one. We begin with the simplest design i.e. the Completely Randomized (CR) Design:

THE COMPLETELY RANDOMIZED DESIGN (CR DESIGN)

A completely randomized (CR) design, which is the simplest type of the basic designs, may be defined as a design in which the treatments are assigned to experimental units completely at random, i.e. the randomization is done without any restrictions. This design is applicable in that situation where the entire experimental material is homogeneous (i.e. all the experimental units can be regarded as being similar to each other). We illustrate the concept of the Completely Randomized (CR) Design (pertaining to the case when each treatment is repeated equal number of times) with the help of the following example.

EXAMPLE

An experiment was conducted to compare the yields of three varieties of potatoes. Each variety as assigned at random to equal-size plots, four times. The yields were as follow:

Variety		
A	B	C
23	18	16
26	28	25
20	17	12
17	21	14

Test the hypothesis that the three varieties of potatoes are not different in the yielding capabilities.

SOLUTION

The first thing to note is that this is an example of the Completely Randomized (CR) Design. We are assuming that all twelve of the plots (i.e. farms) available to us for this experiment are homogeneous (i.e. similar) with regard to the fertility of the soil, the weather conditions, etc., and hence, we are assigning the four varieties to the twelve plots totally at random. Now, in order to test the hypothesis that the mean yields of the three varieties of potato are equal, we carry out the six-step hypothesis-testing procedure, as given below:
Hypothesis-Testing Procedure:

i) H0 : $\mu A = \mu B = \mu C$
HA : Not all the three means are equal

ii) Level of Significance:
$\alpha = 0.05$

iii) Test Statistic:

$$F = \frac{MS\ Treatments}{MS\ Error}$$

which, if H0 is true, has an F distribution with $\upsilon 1 = k-1 = 3 - 1 = 2$ and $\upsilon 2 = n-k = 12 - 3 = 9$ degree of freedom

iv) Computations:
The computation of the test statistic presented above involves quite a few steps, including the formation of what is known as the ANOVA Table.
First of all, let us consider what is meant by the ANOVA Table (i.e. the Analysis of Variance Table).
In the case of the Completely Randomized (CR) Design, the ANOVA Table is a table of the type given below:

ANOVA TABLE IN THE CASE OF THE COMPLETELY RANDOMIZED (CR) DESIGN

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Between treatments	k-1	SST	MST	MST/MSE
Within treatments (Error)	n-k	SSE	MSE	--
Total	n-1	TSS	--	--

Let us try to understand this table step by step:
The very first column is headed 'Source of Variation', and under this heading, we have three distinct sources of variation:

'Total' stands for the overall variation in the twelve values that we have in our data-set.

Variety		
A	B	C
23	18	16
26	28	25
20	17	12
17	21	14

As you can see, the values in our data-set are 23, 26, 20, 17, 18, 28, and so on. Evidently, there is a variation in these values, and the term 'Total' in the lowest row of the ANOVA Table stands for this overall variation.
The term 'Variation between Treatments' stands for the variability that exists between the three varieties of potato that we have sown in the plots.
(In this example, the term 'treatments' stands for the three varieties of potato that we are trying to compare)

(The term 'variation between treatments' points to the fact that:
It is possible that the three varieties or, at least two of the varieties are significantly different from each other with regard to their yielding capabilities. This variability between the varieties can be measured by measuring the differences between the mean yields of the three varieties.)
The third source of variation is 'variation within treatments'. This point to the fact that even if only one particular variety of potato is sown more than once, we do not get the same yield every time

Variety		
A	B	C
23	18	16
26	28	25
20	17	12
17	21	14

In this example, variety A was sown four times, and the yields were 23, 26, 20, and 17 --- all different from one another! Similar is the case for variety B as well as variety C. The variability in the yields of variety A can be called 'variation within variety A'.

Similarly, the variability in the yields of variety B can be called 'variation within variety B'. Also, the variability in the yields of variety C can be called 'variation within variety C'. We can say that the term 'variability within treatments' stands for the combined effect of the above-mentioned three variations. The 'variation within treatments' is also known as the 'error variation'. This is so because we can argue that if we are sowing the same variety in four plots which are very similar to each other, then we should have obtained the same yield from each plot!

If it is not coming out to be the same every time, we can regard this as some kind of an 'error'.
The second, third and fourth columns of the ANOVA Table are entitled 'degrees of freedom', 'Sum of Squares' and 'Mean Square'.

ANOVA TABLE IN THE CASE OF THE COMPLETELY RANDOMIZED (CR) DESIGN

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Between treatments	k-1	SST	MST	MST/MSE
Within treatments (Error)	n-k	SSE	MSE	--
Total	n-1	TSS	--	--

The point to understand is that the sources of variation corresponding to treatments and error will be measured by computing quantities that are called Mean Squares, and 'Mean Square' can be defined as:

$$Mean\ Square = \frac{Sum\ of\ Squares}{Degrees\ of\ Freedom}$$

Corresponding to these two sources of variation, we have the following two equations:

$$1)\ 'MS\ Treatment' = \frac{'SS\ Treatment'}{d.f.}$$

AND

$$2)\ 'MS\ Error' = \frac{'SS\ Error'}{d.f.}$$

It has been mathematically proved that, with reference to Analysis of Variance pertaining to the Completely Randomized (CR) Design, the degrees of freedom corresponding to the Treatment Sum of Squares are k-1, and the degrees of freedom corresponding to the Error Sum of Squares are n-k. Hence, the above two equations can be written as:

$$1)\ 'MS\ Treatment' = \frac{'SS\ Treatment'}{k-1}$$

AND

$$2)\ 'MS\ Error' = \frac{'SS\ Error'}{n-k}$$

How do we compute the various sums of squares? The three sums of squares occurring in the third column of the above ANOVA Table are given by:

$$1)\ Total\ SS = TSS = \sum_i \sum_j X_{ij}^2 - C.F.$$

$$2)\ SS\ Treatment = SST = \frac{\sum_j T_{\cdot j}^2}{r} - C.F.$$

where C.F. stands for 'Correction Factor', and is given by

$$C.F. = \frac{T_{\cdot\cdot}^2}{n}$$

and r denotes the number of data-values per column (i.e. the number of rows). (It should be noted that this example pertains to that case of the Completely Randomized (CR) Design where each treatment is being repeated equal number of times, and the above formulae pertain to this particular situation. With reference to the CR Design, it should be noted that, in some situations, the various treatments are not repeated an equal number of times.

For example, with reference to the twelve plots (farms) that we have been considering above, we could have sown variety A in five of the plots, variety B in three plots, and variety C in four plots. Going back to the formulae of various sums of squares, the sum of squares for error is given by

$$3)\ SS\ Error = Total\ SS - SS\ Treatment$$

i.e.

$$SSE = TSS - SST$$

It is interesting to note that,
Total SS = SS Treatment + SS Error
In a similar way, we have the equation:
Total d.f. = d.f. for Treatment + d.f. for Error
It can be shown that the degrees of freedom pertaining to 'Total' are n - 1.
Now,

n-1 = (k-1) + (n-k)

i.e.
Total d.f. = d.f. for Treatment + d.f. for Error
The notations and terminology given in the above equations relate to the following table:

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{\cdot j}$	86	84	67	237	4953
$T_{\cdot j}^2$	7396	7056	4489	18941	↑
$\sum_i X_{ij}^2$	1894	1838	1221	4953	| Check ←

The entries in the body of the table i.e. 23, 26, 20, 17, and so on are the yields of the three varieties of potato that we had sown in the twelve farms. The entries written in brackets next to the above-mentioned data-values are the squares of those values.
For example:

529 is the square of 23,
676 is the square of 26,
400 is the square of 20,

and so on.

Adding all these squares, we obtain:

$$\sum_i \sum_j X_{ij}^2 = 4953$$

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{.j}$	86	84	67	237	4953
$T_{.j^2}$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

The notation T.j stands for the total of the jth column.(You must already be aware that, in general, the rows of a bivariate table are denoted by the letter 'i', whereas the columns of a bivariate table are denoted by the letter 'j'.

In other words, we talk about the 'ith row', and the 'jth column' of a bivariate table.)The 'dot' in the notation T.j indicates the fact that summation has been carried out over i (i.e. over the rows).

In this example, the total of the values in the first column is 86, the total of the values in the second column is 84, and the total of the values in the third column is 67.

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{.j}$	86	84	67	237	4953
$T_{.j^2}$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

Hence, ΣT.j is equal to 237.

ΣT.j is also denoted by T..

i.e.

T.. = ΣT.j The 'double dot' in the notation T.. indicates that summation has been carried out over i as well as over j.

The row below T.j is that of T.j2, and squaring the three values of T.j, we obtain the quantities 7396, 7056 and 4489. Adding these, we obtain ΣT.j2 = 18941.

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{\cdot j}$	86	84	67	237	4953
$T_{\cdot j^2}$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

Now that we have obtained all the required quantities, we are ready to compute SS Total, SS Treatment, and SS Error: We have

$$C.F. = \frac{T_{\cdot\cdot}^2}{n} = \frac{(237)^2}{12} = 4680.75$$

Hence, the total sum of squares is given by

$$\begin{aligned} TSS &= \sum_i \sum_j X_{ij}^2 - C.F. \\ &= 4953 - 4680.75 \\ &= 272.25 \end{aligned}$$

Also, we have

$$SS\ Treatment = SST = \frac{\sum_j T_{\cdot j}^2}{r} - C.F.$$

$$= \frac{18941}{4} - 4680.75$$

$$= 54.50$$

And, hence:
SS Error = SSE = TSS - SST
= 272.25 - 54.50 = 217.75

In this example, we have n = 12, and k = 3, hence:
n-1=11,
k- 1 = 2, and n - k = 9.
Substituting the above sums of squares and degree of freedom in the ANOVA table, we obtain:

ANOVA TABLE

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between treatments (i.e. Between varieties)	2	54.50		
Error	9	217.75		
Total	11	272.25		

Now, the mean squares for treatments and for error are very easily found by dividing the sums of squares by the corresponding degrees of freedom. Hence, we have

ANOVA TABLE

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between treatments (i.e. Between varieties)	2	54.50	27.25	
Error	9	217.75	24.19	
Total	11	272.25	--	

As indicated earlier, the test-statistic appropriate for testing the null hypothesis

H0 : $\mu A = \mu B = \mu C$

versus

HA : Not all the three means are equal is:

$$F = \frac{MS\ Treatments}{MS\ Error}$$

which, if H0 is true, has an F distribution with $\upsilon 1 = k-1 = 3 - 1 = 2$ and $\upsilon 2 = n-k = 12 - 3 = 9$ degree of freedom Hence, it is obvious that F will be found by dividing the first entry of the fourth column of our ANOVA Table by the second entry of the same column i.e.

$$F = \frac{MS\ Treatment}{MS\ Error} = \frac{27.25}{24.19} = 1.13$$

We insert this computed value of F in the last column of our ANOVA table, and thus obtain:

ANOVA TABLE

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between treatments (i.e. Between varieties)	2	54.50	27.25	1.13
Error	9	217.75	24.19	--
Total	11	272.25	--	--

The fifth step of the hypothesis - testing procedure is to determine the critical region. With reference to the Analysis of Variance procedure, it can be shown that it is appropriate to establish the critical region in such a way that our test is a right-tailed test. In other words, the critical region is given by:

Critical Region:

$F > F\alpha\ (k - 1, n - k)$

In this example:

The critical region is $F > F0.05\ (2,9) = 4.26$

vi) Conclusion:

Since the computed value of F = 1.13 does not fall in the critical region, so we accept our null hypothesis and may conclude that, on the average, there is no difference among the yielding capabilities of the three varieties of potatoes.

In this course, we will not be discussing the details of the mathematical points underlying One-Way Analysis of Variance that is applicable in the case of the Completely Randomized (CR) Design. One important point that the students should note is that the ANOVA technique being presented here is valid under the following assumptions:

- The k populations (whose means are to be compared) are normally distributed;
- All k populations have equal variances i.e. $\sigma 12 = \sigma 22 = \ldots = \sigma k2$. (This property is called homoscedasticity)
- The k samples have been drawn randomly and independently from the respective populations.

Next, we begin the discussion of the Randomized Complete Block (RCB) Design:

THE RANDOMIZED COMPLETE BLOCK DESIGN (RCB DESIGN)

A randomized complete block (RCB) design is the one in which

- The experimental material (which is not homogeneous overall) is divided into groups or blocks in such a manner that the experimental units within a particular block are relatively homogeneous.
- Each block contains a complete set of treatments, i.e., it constitutes a replication of treatments.
- The treatments are allocated at random to the experimental units within each block, which means the randomization is restricted.(A new randomization is made for every block.)The object of this type of arrangement is to bring the variability of the experimental material under control.

In simple words, the situation is as follows:

We have experimental material which is not homogeneous overall. For example, with reference to the example that we have been considering above, suppose that the plots which are closer to a canal are the most fertile ones, the ones a little further away are a little less fertile, and the ones still further away are the least fertile.

In such a situation, we divide the experimental material into groups or blocks which are relatively homogeneous. The randomized complete block design is perhaps the most widely used experimental design. Two-way analysis of variance is applicable in case of the randomized complete block (RCB) design.

We illustrate this concept with the help of an example:

EXAMPLE

In a feeding experiment of some animals, four types of rations were given to the animals that were in five groups of four each. The following results were obtained

Groups	Rations			
	A	B	C	D
I	32.3	33.3	30.8	29.3
II	34.0	33.0	34.3	26.0
III	34.3	36.3	35.3	29.8
IV	35.0	36.8	32.3	28.0
V	36.5	34.5	35.8	28.8

The values in the above table represent the gains in weights in pounds. Perform an analysis of variance and state your conclusions. In the next lecture, we will discuss this example in detail, and will analyze the given data to carry out the following test:

H0 : $\mu A = \mu B = \mu C = \mu D$

HA : Not all the treatment-means are equal

LECTURE NO. 44

- Randomized Complete Block Design
- The Least Significant Difference (LSD) Test
- Chi-Square Test of Goodness of Fit

At the end of the last lecture, we introduced the concept of the *Randomized Complete Block (RCB) Design*, and we picked up an example to illustrate the concept. In this lecture, we begin with a detailed discussion of the same example:

EXAMPLE

In a feeding experiment of some animals, four types of rations were given to the animals that were in five groups of four each. The following results were obtained:

Groups	Rations			
	A	B	C	D
I	32.3	33.3	30.8	29.3
II	34.0	33.0	34.3	26.0
III	34.3	36.3	35.3	29.8
IV	35.0	36.8	32.3	28.0
V	36.5	34.5	35.8	28.8

The values in the above table represent the gains in weights in pounds. Perform an analysis of variance and state your conclusions.

SOLUTION

Hypothesis-Testing Procedure:

i a) Our primary interest is in testing:
H0 : $\mu A = \mu B = \mu C = \mu D$
HA : Not all the ration-means (*treatment-means*) are equal

i b) In addition, we can also test:
H′0 : $\mu I = \mu II = \mu III = \mu IV = \mu V$
H′A : Not all the group-means (*block-means*) are equal

ii) Level of significance
$\alpha = 0.05$

iii a) Test Statistic for testing H0 versus HA:

$$F = \frac{MS\ Treatment}{MS\ Error}$$

which, if H0 is true, has an F-distribution with v1 = c-1= 4-1= 3 and v2 = (r-1)(c-1) =(5-1)(4-1) =4 × 3 = 12 degrees of freedom.

iii b) Test Statistic for testing H ′0 versus H ′A:

$$F = \frac{\text{MS Block}}{\text{MS Error}}$$

which, if H0 is true, has an F-distribution with v1 = r-1 = 5-1= 4 and v2 = (r-1)(c-1) = (5-1)(4-1) =4 × 3 = 12 degrees of freedom.

Now, the given data leads to the following table:

iv) Computations:

Groups	Ration				$B_{i.}$	$B_{i.}^2$	$\sum_j X_{ij}^2$
	A	B	C	D			
I	32.3 (10.43.29)	33.3 (1108.89)	30.8 (948.64)	29.3 (858.49)	125.7	15800.49	3959.31
II	34.00 (1156.00)	33.0 (1089.00)	34.3 (1176.49)	26.0 (676.00)	127.3	16205.29	4097.49
III	34.3 (1176.49)	36.3 (1317.69)	35.3 (1246.09)	29.8 (888.04)	135.7	18414.49	4628.31
IV	35.0 (1225.00)	36.8 (1354.24)	32.3 (1043.29)	28.0 (784.00)	132.1	17450.41	4406.53
V	36.5 (1332.25)	34.5 (1190.25)	35.8 (1281.64)	28.8 (829.44)	135.6	18387.36	4633.58
$T_{.j}$	172.1	173.9	168.5	141.9	656.4	86258.04	21725.22
$T_{.j}^2$	29618.41	30241.21	28392.25	20135.61	108387.48		↑
$\sum_i X_{ij}^2$	5933.03	6060.07	5696.15	4035.97	21725.22	←	Check ←—

Hence, we have Total SS $= \sum\sum X_{ij}^2 - \frac{T_{..}^2}{n}$

$$= 21725.22 - \frac{(656.4)^2}{20}$$
$$= 21725.22 - 21543.05$$
$$= 182.17$$

Treatment SS $= \frac{\sum_j T_{.j}^2}{r} - \frac{T_{..}^2}{n}$

$$= \frac{108387.48}{5} - \frac{(656.4)^2}{20}$$
$$= 21677.50 - 21543.05$$
$$= 134.45$$

Block SS $= \frac{\sum_i B_{i.}^2}{c} - \frac{T_{..}^2}{n}$

$$= \frac{86258.04}{4} - \frac{(656.4)^2}{20}$$
$$= 21564.51 - 21543.05$$
$$= \mathbf{21.46}$$

where c represents the number of observations per block (i.e. the number of *columns*)
And
Error SS= Total SS – (Treatment SS + Block SS)
= 182.17 – (134.45 + 21.46)
= 26.26
The degrees of freedom corresponding to the various sums of squares are as follows:

- Degrees of freedom for treatments: c - 1 (i.e. the number of treatments - 1)

- Degrees of freedom for blocks:
  r - 1 (i.e. the number of blocks - 1)
- Degrees of freedom for Total:
  rc - 1 (i.e. the total number of observations - 1)
- Degrees of freedom for error: degrees of freedom for Total minus degrees of freedom for treatments minus degrees of freedom for blocks

i.e. (rc-1) - (r-1) - (c-1)
= rc - r - c + 1
= (r-1) (c-1)

Hence the ANOVA-Table is:

ANOVA-TABLE

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Between Treatments (i.e. Between Rations)	3	134.45	44.82	$F_1 = 20.47$
Between Blocks (i.e. Between Groups)	4	21.46	5.36	$F_2 = 2.45$
Error	12	26.26	2.19	--
Total	19	182.17	--	--

v a) Critical Region for Testing H0 against HA is given by
F > F0.05 (3, 12) = 3.49

v b) Critical Region for Testing H′0 against H′A is given by
F > F0.05 (4, 12) = 3.26

vi a) **Conclusion Regarding Treatment Means**

Since our computed value F1 = 20.47 exceeds the critical value F0.05 (3, 12) = 3.49, therefore we *reject* the null hypothesis, and conclude that there *is* a difference among the means of *at least two* of the treatments (i.e. the mean weight-gains corresponding to at least two of the rations are different).

vi b) Conclusion Regarding Block Means

Since our computed value F2 = 2.45 does not exceed the critical value F0.05(4, 12) = 3.26, therefore we accept the null hypothesis regarding the equality of block means and thus conclude that blocking (i.e. the *grouping* of animals) was actually *not required* in this experiment. As far as the conclusion regarding the block means is concerned, this information can be used when designing a similar experiment in the future.

[If blocking is actually not required, then a future experiment can be designed according to the Completely Randomized design, thus retaining more degrees of freedom for Error. (The more degrees of freedom we have for Error, the better, because an estimate of the error variation based on a greater number of degrees of freedom implies an estimate based on a greater amount of information (which is obviously good).)]

As far as the conclusion regarding the *treatment* means is concerned, the situation is as follows:

Now that we have concluded that there is a significant difference between the treatment means (i.e. we have concluded that the mean weight-gain is *not* the same for all four rations, then it is obvious that we would be interested in finding out, "Which of the four rations produces the greatest weight-gain?"

The answer to this question can be found by applying a technique known as the *Least Significant Difference (LSD) Test*.

THE LEAST SIGNIFICANT DIFFERENCE (LSD) TEST

According to this procedure, we compute the *smallest* difference that would be judged significant, and *compare* the absolute values of all differences of means with it. This smallest difference is called the least significant difference or LSD, and is given by:

LEAST SIGNIFICANT DIFFERENCE (LSD):

$$\text{LSD} = t_{\alpha/2,(\nu)}\sqrt{\frac{2(\text{MSE})}{r}},$$

where MSE is the Mean Square for Error, r is the size of equal samples, and $t\alpha/2\ (\nu)$ is the value of t at $\alpha/2$ level taken against the error degrees of freedom (ν).

The test-criterion that uses the least significant difference is called the LSD test.

Two sample-means are declared to have come from populations with significantly different means, when the absolute value of their difference *exceeds* the LSD.

It is customary to *arrange* the sample means in ascending order of magnitude, and to draw a *line* under any pair of adjacent means (or set of means) that are not significantly different.
The LSD test is applied *only* if the null hypotheses is rejected in the Analysis of Variance. We will not be going into the mathematical details of this procedure, but it is useful to note that this procedure can be regarded as an alternative way of conducting the t-test for the equality of two population means.
If we were to apply the usual two-sample t-test, we would have had to *repeat* this procedure quite a few times!
(The six possible tests are:
H0 : μA = μB
H0 : μA = μC
H0 : μA = μD
H0 : μB = μC
H0 : μB = μD
H0 : μC = μD)
The LSD test is a procedure by which we can compare *all* the treatment means simultaneously.
We illustrate this procedure through the above example:
The Least Significant Difference is given by

$$\begin{aligned} LSD &= t_{\alpha/2,(\nu)}\sqrt{\frac{2(MSE)}{r}} \\ &= t_{0.025(12)}\sqrt{\frac{2(2.19)}{5}} \\ &= 2.179\sqrt{\frac{2(2.19)}{5}} \\ &= 2.179\times 0.936 \\ &= 2.04. \end{aligned}$$

Going back to the given data:

Groups	Rations			
	A	B	C	D
I	32.3	33.3	30.8	29.3
II	34.0	33.0	34.3	26.0
III	34.3	36.3	35.3	29.8
IV	35.0	36.8	32.3	28.0
V	36.5	34.5	35.8	28.8
Total	172.1	173.9	168.5	141.9
Mean	34.42	34.78	33.70	28.38

We find that the four treatment means are:

$$\bar{X}_A = 34.42$$
$$\bar{X}_B = 34.78$$
$$\bar{X}_C = 33.70$$
$$\bar{X}_D = 28.38$$

Arranging the above means in *ascending* order of magnitude, we obtain:

$\bar{X}_D$	$\bar{X}_C$	$\bar{X}_A$	$\bar{X}_B$
28.38	33.70	34.42	34.78

Drawing lines under pairs of adjacent means (or sets of means) that are not significantly different, we have:

$\bar{X}_D$	$\bar{X}_C$	$\bar{X}_A$	$\bar{X}_B$
28.38	33.70	34.42	34.78

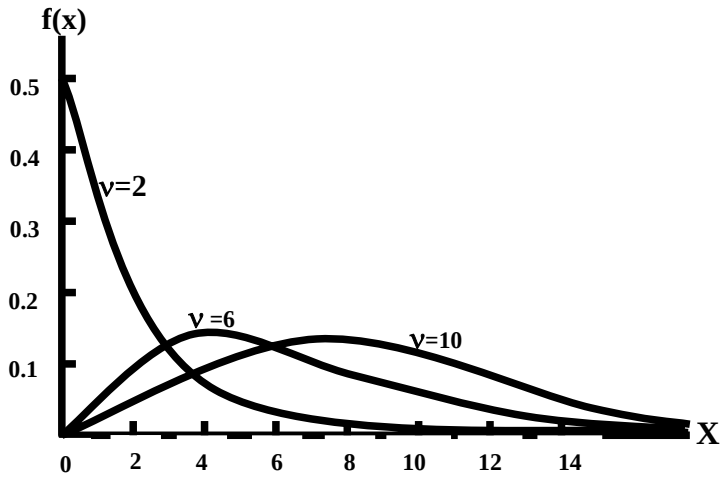
From the above, it is obvious that rations C, A and B are *not* significantly different from each other with regard to weight-gain. The only ration which *is* significantly different from the others is ration D.
Interestingly, ration D has the *poorest* performance with regard to weight-gain. As such, if our primary objective is to *increase* the weights of the animals under study, then we may recommend *any* of the other three rations i.e. A, B and C to the farmers (depending upon availability, price, etc.), but we must *not* recommend ration D.
Next, we will consider two important tests based on the
chi-square distribution. These are:

- The chi-square test of *goodness of fit*
- The chi-square test of *independence*

Before we begin the discussion of these tests, let us review the basic properties of the chi-square distribution:

PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The Chi-Square ($\chi 2$) distribution has the following properties:
1. It is a continuous distribution ranging from 0 to $+\infty$ .The number of degrees of freedom determines the *shape* of the chi-square distribution. (Thus, there is a different chi-square distribution for each number of degrees of freedom. As such, it is a whole *family* of distributions.)
2. The curve of a chi-square distribution is *positively skewed*. The skewness *decreases as* ν *increases*.



$\chi 2$-distribution for various values of ν

As indicated by the above figures, the chi-square distribution tends to the normal distribution as the number of degrees of freedom approaches infinity. Having reviewed the basic properties of the chi-square distribution; we begin the discussion of the Chi-Square Test of Goodness of Fit:

CHI-SQUARE TEST OF GOODNESS OF FIT

The chi-square test of goodness-of-fit is a test of hypothesis concerned with the comparison of observed frequencies of a sample, and the corresponding expected frequencies based on a theoretical distribution. We illustrate this concept with the help of the same example that we considered in Lecture No. 28 --- the one pertaining to the fitting of a binomial distribution to real data:

EXAMPLE:

The following data has been obtained by tossing a *LOADED* die 5 times, and noting the number of times that we obtained a *six*. Fit a binomial distribution to this data.

No. of Sixes (x)	0	1	2	3	4	5	Total
Frequency (f)	12	56	74	39	18	1	200

SOLUTION

To fit a binomial distribution, we need to find n and p.
Here n = 5, the largest x-value.
To find p, we use the relationship $\bar{x} = np$.
We have:

No. of Sixes (x)	0	1	2	3	4	5	Total
Frequency (f)	12	56	74	39	18	1	200
fx	0	56	148	117	72	5	398

Therefore:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{0+56+148+117+72+5}{200}$$

$$= \frac{398}{200} = 1.99$$

Using the relationship $\bar{x} = np$,
we obtain

$p = 1.99$ or $p = 0.398$.

Letting the random variable X represent the number of sixes, the above calculations yield the fitted binomial distribution as

$$b(x;5,\ 0.398) = \binom{5}{x} (0.398)^x (0.602)^{5-x}$$

Hence the *probabilities* and *expected frequencies* are calculated as below:

No. of Sixes (x)	**Probability f(x)**		**Expected frequency**
0	$\binom{5}{0} q^5 = (0.602)^5$	**= 0.07907**	**15.8**
1	$\binom{5}{1} q^5 p = 5.(0.602)^4 (0.398)$	**= 0.26136**	**52.5**
2	$\binom{5}{2} q^3 p^2 = 10.(0.602)^3 (0.398)^2$	**= 0.34559**	**69.1**
3	$\binom{5}{3} q^2 p^3 = 10.(0.602)(0.398)^3$	**= 0.22847**	**45.7**
4	$\binom{5}{4} q p^4 = (0.602)(0.398)^4$	**= 0.07553**	**15.1**
5	$\binom{5}{5} p^5 = (0.398)^5$	**= 0.00998**	**2.0**
Total		**= 1.00000**	**200.0**

Comparing the observed frequencies with the expected frequencies, we obtain:

No. of Sixes x	Observed Frequency o_i	Expected Frequency e_i
0	12	15.8
1	56	52.5
2	74	69.1
3	39	45.7
4	18	15.1
5	1	2.0
Total	200	200.0

The above table seems to indicate that there is not much discrepancy between the observed and the expected frequencies. Hence, in Lecture No.28, we concluded that it was a reasonably *good* fit. But, it was indicated that *proper* comparison of the expected frequencies with the observed frequencies can be accomplished by applying the *chi-square test of goodness of fit*. The Chi-Square Test of Goodness of Fit enables us to determine in a *mathematical* manner whether or not the theoretical distribution fits the observed distribution reasonably well.

The procedure of the chi-square of goodness of fit is very *similar* to the *general* hypothesis-testing procedure:

HYPOTHESIS-TESTING PROCEDURE

Step-1:

H_0 : The fit is good

H_A : The fit is not good

Step-2:

Level of Significance: $\alpha = 0.05$

Step-3: Test-Statistic:

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

which, if H0 is true, follows the chi-square distribution having k - 1 - r degrees of freedom(where k = No. of x-values (*after* having carried out the necessary mergers), and r = number of parameters that we estimate from the sample data).

Step-4: Computations:

No. of Sixes x	Observed Frequency o_i	Expected Frequency e_i	$o_i - e_i$	$(o_i - e_i)^2$	$(o_i - e_i)^2/e_i$
0	12	15.8	-3.8	14.44	0.91
1	56	52.5	3.5	12.25	0.23
2	74	69.1	4.9	24.01	0.35
3	39	45.7	-6.7	44.89	0.98
4	18 } **19**	15.1 } **17.1**	1.9	3.61	0.21
5	1	2.0			
Total	200	200.0			2.69

IMPORTANT NOTE

In the above table, the category x = 4 has been merged with the category x = 5 because of the fact that the expected frequency corresponding to x = 5 was less than 5.

[It is one of the basic requirements of the chi-square test of goodness of fit that the expected frequency of any x-value (or any combination of x-values) should not be less than 5.]Since we have combined the category x = 4 with the category x = 5, hence k = 5. Also, since we have estimated one parameter of the binomial distribution (i.e. p) from the sample data, hence r = 1. (The other parameter n is already known.)

As such, our statistic follows the chi-distribution having k - 1 - r = 5 - 1 - 1 = 3 degrees of freedom. Going back to the above calculations, the computed value of our test-statistic comes out to be $\chi 2 = 2.69$.

Step-5: Critical Region:

Since $\alpha = 0.05$, hence, from the Chi-Square Tables, it is evident that the critical region is:$\chi 2 \geq \chi 20.05\ (3) = 7.82$

Step-6:

Conclusion:

Since the computed value of $\chi 2$ i.e. 2.69 is less than the critical value 7.82, hence we accept H0 and conclude that the fit is good.(In other words, with only 5% risk of committing Type-I error, we conclude that the distribution of our random variable X can be regarded as a binomial distribution with n = 5 and p = 0.398.)

LECTURE NO. 45

- Chi-Square Test of Goodness of Fit (in continuation of the last lecture)
- Chi-Square Test of Independence
- The Concept of Degrees of Freedom
- p-value
- Relationship Between Confidence; Interval and Tests of Hypothesis

An Overview of the Science of Statistics in Today's World (including Latest Definition of Statistics)

The students will recall that, towards the end of the last lecture, we discussed the chi-square test of goodness of fit. We applied the test to the example where we had fitted a binomial distribution to real data, and, since the computed value of our test statistic turned out to be insignificant, therefore we concluded that the fit was good.

Let us consider another example:

EXAMPLE

The platform manager of an airline's terminal ticket counter wants to determine whether customer arrivals can be modelled by using a Poisson distribution. The manager is especially interested in late-night traffic.

Accordingly, data for the time period of interest have been collected, as follows:

Number of Arrivals Per Minute	Frequency
0	84
1	114
2	70
3	60
4	32
5	16
6	15
7	4
8	5
	400

Is the distribution Poisson?

SOLUTION:

First of all, we fit a Poisson distribution to the given data Because a mean is not specified, it must be estimated from the sample data. The mean of the frequency distribution can be found by using the formula

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

where $n = \Sigma f$.

Thus we have the following calculations:

Number of Arrivals x	Frequency f	fx
0	84	0
1	114	114
2	70	140
3	60	180
4	32	128
5	16	80
6	15	90
7	4	28
8	5	40
	400	800

Hence:

$$Mean = \bar{x} = \frac{\sum fx}{n} = \frac{800}{400} = 2$$

Replacing μ by , the formula for the Poisson probabilities is

Hence, we obtain: $f(x) = \frac{e^{-\bar{x}}\bar{x}^x}{x!} = \frac{e^{-2}2^x}{x!}$

Number of Customer Arrivals	Observed Frequencies	Poisson Probabilities f(x)	Expected Frequencies 400 f(x)
0	84	0.1353	54.12
1	114	0.2707	108.28
2	70	0.2707	108.28
3	60	0.1804	72.16
4	32	0.0902	36.08
5	16	0.0361	14.44
6	15	0.0120	4.80
7	4	0.0034	1.36
8	5	0.0009	0.36
9 or more	0	0.0002	0.08
	400	1	400

Next, we apply the chi-square test of goodness of fit according to the following procedure:

HYPOTHESIS-TESTING PROCEDURE

Step-1:
H0 : Arrivals are Poisson-distributed.
H1 : The distribution is not Poisson.

Step-2:
Level of Significance: $\alpha = 0.05$

Step-3: Test-Statistic: $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$

which, if H0 is true, follows the chi-square distribution having k - 1 - r degrees of freedom; (where k = No. of x-values (after having carried out the necessary mergers), and r = number of parameters that we estimate from the sample data)
Step-4: Computations:
The necessary calculations are shown in the following table:

Number of Customer Arrivals	Observed Frequency o_i	Expected Frequency e_i	(0 – e)	$(0-e)^2$	$(0-e)^2/e$
0	84	54.12	29.88	892.81	16.50
1	114	108.28	5.72	32.72	0.30
2	70	108.28	-38.28	1465.36	13.53
3	60	72.16	-12.16	147.87	2.05
4	32	36.08	-4.08	16.65	0.46
5	16	14.44	1.56	2.43	0.17
6	15	4.80			
7	4	1.36	17.40	302.76	45.87
8	5	0.36			
9 or more	0	0.08			
	400	400			χ^2=78.88

With reference to the above, it should be noted that, since some of the expected frequencies are less than the required minimum of 5, it became necessary to combine some of those classes. Combination is best accomplished working from the bottom up.

In order that we obtain a number greater than 5, the last four expected frequencies had to be combined.

Hence, the effective number of categories becomes 7.

Step-5:

Determination of the Critical Region:

Since the effective number of categories becomes 7

Therefore k = 7.

Also, since the one lone parameter of the Poisson distribution has been estimated from the sample data, hence r = 1.

Hence: Our statistic follows the chi-square distribution having

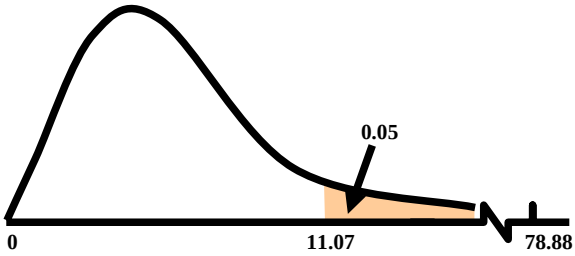
k - 1 - r = 7 - 1 - 1 = 5

degrees of freedom.

The critical region is given by

$\chi 2 \geq \chi 20.05\ (5) = 11.07$

CRITICAL REGION:



Step-6:

Conclusion:

Since the computed value of our test statistic i.e. 78.88 is much larger than the critical value 11.07, therefore, we reject H0 and conclude that the distribution is probably not a Poisson distribution with parameter 2.

(With only 5% risk of committing Type-1 error, we conclude that the fit is not good.)

In fact, the computed value of our test statistic i.e. 78.88 is so large that it is possible that if we had set the level of significance at 1%, even then it would have exceeded the critical value. The students are encouraged to check this up themselves. If the computed value does fall in the critical region corresponding to 1% level of significance, then our result is highly significant

RATIONALE OF THE CHI-SQUARE TEST OF GOODNESS OF FIT

It is clear that $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$ will be a small quantity when all the oi's are close to the corresponding ei's. (In fact, if the observed frequencies are exactly equal to the expected ones, then $\chi 2$ will be exactly equal to zero.)

The $\chi 2$ - statistic will become larger when the differences between the oi's and ei have become larger. Thus, $\chi 2$ measure the amount of deviation (or discrepancy) between the observed and the expected results.

ASSUMPTIONS OF THE CHI-SQUARE TEST OF GOODNESS OF FIT

While applying the chi-square test of goodness of fit, certain requirements must be satisfied, three of which are as follows:

1. The total number of observations (i.e. the sample size) should be at least 50.

2. The expected number ei in any of the categories should not be less than 5. (So, when the expected frequency ei in any category is less than 5, we may combine this category with one or more of the other categories to get $ei \geq 5$.)

3. The observations in the sample or the frequencies of the categories should be independent.

Next, we begin the discussion of the Chi-Square Test of Independence:

In this regard, it is interesting to note that, (since the formula of chi-square in this particular situation is very similar to the formula that we have just discussed), therefore, the chi-square test of independence can also be regarded as a kind of chi-square test of goodness of fit. We illustrate this concept with the help of an example:

EXAMPLE

A random sample of 250 men and 250 women were polled as to their desire concerning the ownership of personal computers. The following data resulted:

	Men	Women	Total
Want PC	120	80	200
Don't Want PC	130	170	300
Total	250	250	500

Test the hypothesis that desire to own a personal computer is independent of sex at the 0.05 level of significance.

SOLUTION

i) H0 : The two variables of classification (i.e. gender and desire for PC) are independent, and
H1 : The two variables of classification are not independent.

ii) The significance level is set at $\alpha = 0.05$.

iii) The test-statistic to be used is $\chi^2 = \sum_i \sum_j (0_{ij} - e_{ij})^2 / e_{ij}$

This statistic, if H0 is true, has an approximate chi-square distribution with (r - 1) (c - 1) = (2 - 1) (2 - 1) = 1 degrees of freedom.

iv) Computations:

In order to determine the value of $\chi 2$, we carry out the following computations:

The first step is to compute the expected frequencies. The expected frequency of any cell is obtained by multiplying the marginal total to the right of that cell by the marginal total directly below that cell, and dividing this product by the grand total.

In this example,

$$e_{11} = \frac{(200)(250)}{500} = 100,$$

$$e_{12} = \frac{(200)(250)}{500} = 100,$$

$$e_{21} = \frac{(300)(250)}{500} = 150,$$

and

$$e_{22} = \frac{(300)(250)}{500} = 150.$$

Hence, we have:

Expected Frequencies:

	Men	Women	Total
Want PC	100	100	200
Don't Want PC	150	150	300
Total	250	250	500

Next, we construct the columns of oij - eij, (oij - eij)2 and (oij - eij)2 eij , as shown below:

Observed Frequency o_{ij}	Expected Frequency e_{ii}	$o_{ij} - e_{ij}$	$(o_{ij} - e_{ij})^2$	$(o_{ij} - e_{ij})^2/e_{ii}$
120	100	20	400	4.00
130	150	-20	400	2.67
80	100	-20	400	4.00
170	150	20	400	2.67
				$\chi^2 = 13.33$

Hence, the computed value of our test-statistic comes out to be $\chi^2 = 13.33.$

v) Critical Region:

$\chi 2 \geq \chi 20.05(1) = 3.84$

vi)

Conclusion:

Since 13.33 is bigger than 3.84, we reject H0 and conclude that desire to own a personal computer set and sex are associated. Now that we have concluded that gender and desire for PC are associated, the natural question is, "Which gender is it where the proportion of persons wanting a PC is higher?" We have:

	Men	Women	Total
Want PC	120	80	200
Don't Want PC	130	170	300
Total	250	250	500

A close look at the given data indicates clearly that the proportion of persons who are desirous of owning a personal computer is higher among men than among women.
And, (since our test statistic has come out to be significant), therefore we can say that the proportion of men wanting a PC is significantly higher than the proportion of women wanting to own a PC.
Let us consider another example:

EXAMPLE

A national survey was conducted in a country to obtain information regarding the smoking patterns of the adults males by marital status. A random sample of 1772 citizens, 18 years old and over, yielded the following data :

MARITAL STATUS	SMOKING PATTERN			Total
	Total Abstinence	Only at times	Regular Smoker	
Single	67	213	74	**354**
Married	411	63	129	**1173**
Widowed	85	51	7	**143**
Divorced	27	60	15	**102**
Total	**590**	**957**	**225**	**1772**

Use this data to decide whether there is an association between marital status and smoking patterns. The students are encouraged to work on this problem on their own, and to decide for themselves whether to accept or reject the null hypothesis.(In this problem, the null and the alternative hypotheses will be:

H0: Marital status and smoking patterns are statistically independent.

HA : Marital status and smoking patterns are not statistically independent.)

This brings us to the end of the series of topics that were to be discussed in some detail for this course on Statistics and Probability. For the remaining part of today's lecture, we will be discussing some interesting and important concepts. First and foremost, let us consider the concept of

DEGREES OF FREEDOM

As you will recall, when discussing the t-distribution, the chi-square distribution, and the F-distribution, it was conveyed to you that the parameters that exists in the equations of those distributions are known as degrees of freedom. But the question is, 'Why these parameters are called degrees of freedom?' Let us try to obtain an answer to this question by considering the following:

Consider the two-dimensional plane, and consider a straight line segment in the plane. If one edge of the line segment is fixed at some point (x0, y0), the line segment can be rotated in the plane such that the fixed edge stays in its place. In other words, we can say that the line segment is free to move in the plane with one restriction. Hence, if we fix one end-point of the line segment, then we are left with one degree of freedom for its movement. Next, consider the case when we fix both end-points of the line segment in the plane. In this case, both degrees of freedom are lost, and therefore the line can no longer move in the plane. But, if we view the above situation with reference to the three-dimensional space --- the one that we live in --- we note that the whole plane (containing the fixed line segment) can move in three dimensions, and hence, we have one degree of freedom for its movement. Let us try to understand this concept in another way: Suppose we have a sample of size n = 6, and suppose that the sum of the sample values is 20. That is, we have the following situation: Our Sample:

Sr. No.	Value
1	
2	
3	
4	
5	
6	
Total	20

Now, the point is that, given this total of 20, if we choose the first 5 values freely, we are not free to choose the sixth value. Hence, one degree of freedom is lost. This point can also be explained in the following alternative way. Given that the sum of the six values is 20, if we have knowledge of the first five values, but the sixth value is missing, then we can re-generate the sixth value. This can also be expressed as follows.

If there are six observations and you find their sum; next, you throw away one of the six observations; then, you can re-generate that observation (because of the fact that you have already computed the sum).
Since, the number of values that can be re-generated is one, hence, the degrees of freedom are n minus one.
(The one which can be re-generated is not the one that we can choose freely.)
Going back to sampling distributions such as the t-distribution, the chi-square distribution and the F-distribution, 'degrees of freedom' can be defined as the number of observations in the sample minus the number of population parameters that are estimated from the sample data (from those observations). For example, in lecture number 39, we noted that the statistic follows the t-distribution having n-1 degrees of freedom.

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

Here n denotes the number of observations in our sample, and since we are estimating one population parameter i.e. σ from the sample data, hence the number of degrees of freedom is n-1.
Similarly, referring to lecture number 42, the students will recall that it was stated that the statistic $\frac{s_1^2}{s_2^2}$

Follows the F-distribution having (n1-1, n2-1) degrees of freedom
Here n1 denotes the number of observations in the first sample, and since we are estimating one parameter of the first population i.e. σ12 from the sample data, hence the number of degrees of freedom for the numerator of our statistic is n1 minus one. Similarly, n2 denotes the number of observations in the second sample, and since we are estimating one parameter of the second population i.e. σ22 from the sample data, hence the number of degrees of freedom for the denominator of our statistic is n2 minus one. In addition, in today's lecture, you learnt that the statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(0_{ij} - e_{ij})^2}{e_{ij}},$$

follows the chi-square distribution having (r-1)(c-1) degrees of freedom. Let us try to understand this point: Consider the 2 × 2 contingency table ,similar to the one that we had in the example regarding the desire for ownership of a personal computer. In this regard, suppose that we have two variables of classification, A and B, and the situation is as follows:

	A_1	A_2	Total
B_1			200
B_2			300
Total	250	250	500

The point is that, given the marginal totals and the grand total, if we choose the frequencies of the first cell of the first row freely, we are not free to choose the frequency of the second cell of the first row. Also, given the frequency of the above-mentioned first cell, we are not even free to choose the frequency of the second cell of the first column.

Not only this, it is interesting to note that, given the above, we are not even free to choose the frequency of the second cell of the second row or the second column !Hence, given the marginal and grand totals, we have only degree of freedom (i.e. 1 = 1 × 1 = (2-1)(2-1) degrees of freedom).A similar situation holds in the case of a 2 x 3 contingency table. The students are encouraged to work on this point on their own, and to realize for themselves that, in the case of a 2 x 3 contingency table, there exist (2 - 1) (3 - 1) = 2 degrees of freedom . Next, let us consider the concept of p-value:

You will recall that, with reference to the concept of hypothesis-testing, we compared the computed value of our test statistic with a critical value. For example, in case of a right-tailed test, we rejected the null hypothesis if our

computed value exceeded the critical value, and we accepted the null hypothesis if our computed value turned out to be smaller than the critical A hypothesis can also be tested by means of what is known as the p-value.

P-VALUE

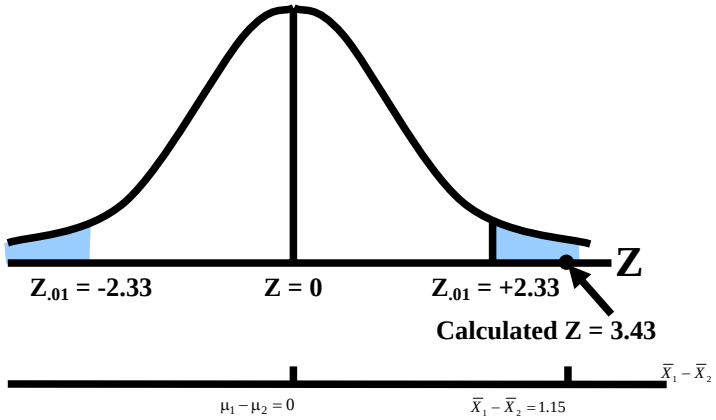
The probability of observing a sample value as extreme as, or more extreme than, the value observed, given that the null hypothesis is true. We illustrate this concept with the help of the example concerning the hourly wages of computer analysts and registered nurses that we discussed in an earlier lecture. The students will recall that the example was as follows:

EXAMPLE

A survey conducted by a market-research organization five years ago showed that the estimated hourly wage for temporary computer analysts was essentially the same as the hourly wage for registered nurses. This year, a random sample of 32 temporary computer analysts from across the country is taken. The analysts are contacted by telephone and asked what rates they are currently able to obtain in the market-place. A similar random sample of 34 registered nurses is taken. The resulting wage figures are listed in the following table.

Computer Analysts			Registered Nurses		
$ 24.10	$25.00	$24.25	$20.75	$23.30	$22.75
23.75	22.70	21.75	23.80	24.00	23.00
24.25	21.30	22.00	22.00	21.75	21.25
22.00	22.55	18.00	21.85	21.50	20.00
23.50	23.25	23.50	24.16	20.40	21.75
22.80	22.10	22.70	21.10	23.25	20.50
24.00	24.25	21.50	23.75	19.50	22.60
23.85	23.50	23.80	22.50	21.75	21.70
24.20	22.75	25.60	25.00	20.80	20.75
22.90	23.80	24.10	22.70	20.25	22.50
23.20			23.25	22.45	
23.55			21.90	19.10	

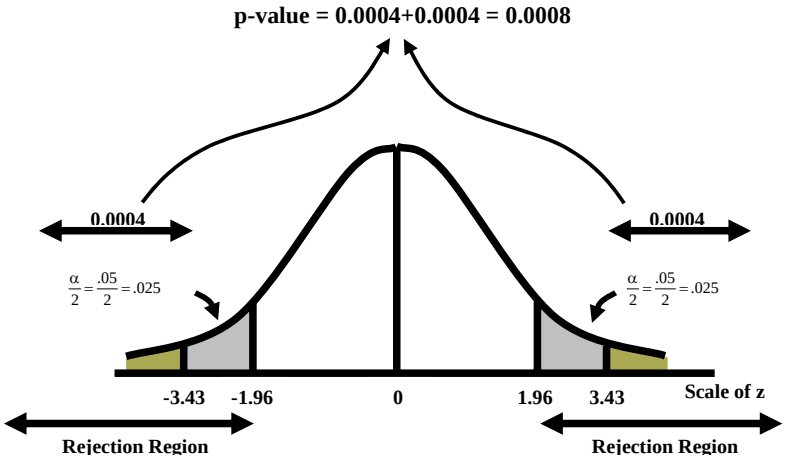
Conduct a hypothesis test at the 2% level of significance to determine whether the hourly wages of the computer analysts are still the same as those of registered nurses. In order to carry out this test, the Null and Alternative Hypotheses were set up as follows:
Null and Alternative Hypotheses:
H0 : $\mu 1 - \mu 2 = 0$
HA : $\mu 1 - \mu 2 \neq 0$
(Two-tailed test)
The computed value of our test statistic came out to be 3.43, whereas, at the 5% level of significance, the critical value was 2.33, hence, we rejected H0.



Hence, we concluded that there was a significant difference between the average hourly wage of a temporary computer analyst and the average hourly wage of a temporary registered nurse. This conclusion could also have been reached by using the

P-VALUE METHOD

I. Looking up the probability of Z > 3.43 in the area table of the standard normal distribution yields an area of .5000 – .4996 = .0004.
II. To compute the p-value, we need to be concerned with the region less than –3.43 as well as the region greater than 3.43 (because the rejection region is in both tails).



The p-value is 0.0004 + 0.0004 = 0.0008. Since this value is very small, it means that the result that we have obtained in this example is highly improbable if, in fact, the null hypothesis is true. Hence, with such a small p-value, we decide to reject the null hypothesis.
The above example shows that: The p-value is a property of the data, and it indicates "how improbable" the obtained result really is. A simple rule is that if our p-value is less than the level of significanceα, then we should reject H0, whereas if our p-value is greater than the level of significanceα, then we should accept H0. (In the above example, α = 0.02 whereas the p-value is equal to 0.0008, hence we reject H0.)

RELATIONSHIP BETWEEN CONFIDENCE INTERVAL AND TESTS OF HYPOTHESIS

Some of the students may already have an idea that there exists some kind of a relationship between the confidence interval for a population parameter θ and a test of hypothesis aboutθ. (After all: When deriving the confidence interval for μ, the area that was kept in the middle of the sampling distribution of $\bar{X}$ was equal to 1-α so that the area in each of the right and left tails was equal to $\alpha/2$. And, when testing the hypothesis H0 : $\mu = \mu 0$ versus HA : $\mu \neq \mu 0$ at level of significance α, the area in each of the right and left tails was again equal to $\alpha/2$.)Hence, consider the following proposition: Let [L, U] be a 100(1 - α)% confidence interval for the parameter θ.Then we will accept the null hypothesis H0 : $\theta = \theta 0$ against H1 : $\theta \neq \theta 0$ at a level of significance α if $\theta 0$ falls inside the confidence interval, but if $\theta 0$ falls outside the interval [L, U], we will reject H0.In the language of hypothesis testing, the (1 - α) 100% confidence interval is known as the acceptance region and the region outside the confidence interval is called the rejection or critical region. The critical values are the end points of the confidence interval. The students are encouraged to work on this point on their own. As we approach the end of this course, we present an Overview of the Science of Statistics in Today's World: Statistics is a vast discipline! In this course, we have discussed the very basic and fundamental concepts of statistics and probability. But, there are numerous other topics that could have been discussed if we had the time. We could have talked about the Latin Square Design, we could have considered Inference Regarding Regression and Correlation Coefficients, we could have discussed Non-Parametric Statistics, and so on, and so forth.
The students are encouraged to study some of these concepts on their own --- as and when time permits --- in order to develop a better understanding and appreciation of the importance of the science of Statistics. In this course, numerous examples were discussed and many numerical problems were presented.

The solutions of these problems were presented in detail, and the various steps were worked out. In doing so, the purpose was to develop in the students a better understanding of the core concepts of the various techniques that were applied. But, it is interesting and useful to note that, a lot many of these numerical problems can be solved within seconds by using the wide variety of statistical packages that are available. These include **SPSS, SAS, Statistica, Statgraph, Minitab, Stata, S-Plus, etc.** (The students are welcome to try out some of these packages on their own.)Towards the end of this course, we present one of the latest definitions of Statistics:

LATEST STATISTICAL DEFINITION

Statistics is a science of decision making for governing the state affairs. It collects analyzes, manages, monitors, interprets, evaluates and validates information. Statistics is Information Science and Information Science is Statistics. It is an applicable science as its tools are applied to all sciences including humanities and social sciences.

- THE END -