

VU LEARNING HUB

vulearninghub.xyz

HIGHLIGHTED HANDOUT

BIO401

Biostatistics

BIO401 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>



Virtual University of Pakistan
FEDERAL GOVERNMENT UNIVERSITY

COURSE HANDOUTS

Biostatistics

BIO401

Order Your VU Books Now

Get all Virtual University books in high-quality hard copy, delivered right to your doorstep. Visit vubookshoppk.com



Order Now

TABLE OF CONTENTS

Sr #	Topic Name
1	Introduction to Biostatistics
2	Basic Terminology - I
3	Basic Terminology - II
4	Data
5	Variables
6	Measurement Scales - I
7	Measurement Scales - II
8	Types of Statistics
9	Sampling and Statistical Inference
10	Basic Terminology for sampling
11	Types of Sampling Methods
12	Simple Random Sampling
13	Stratified Random Sampling
14	Cluster Sampling
15	Systematic Sampling
16	Non-Probability Sampling Methods
17	Convenience Sampling
18	Describing Data with Tables - I
19	Describing Data with Tables - II
20	Describing Data with Tables - III
21	Contingency Tables
22	Describing Data with Charts - I
23	Describing Data with Charts - II
24	Describing Data with Charts - III
25	Box and Whisker Plot
26	Microsoft Excel Demo
27	Describing Data: From its Shapes
28	Describing Data: From its peak
29	Describing Shape: From Box and Whisker Plot
30	The Central Tendency of a data-set
31	Measures of Dispersion
32	Mean Deviation, Standard Deviation and Variance & Coefficient of variation
33	Chebychev's Inequality, The Empirical Rule & The Five-Number Summary
34	Box and Whisker Plot, Pearson's Coefficient of Skewness
35	Set Theory, Counting Rules
36	Permutations, Combinations

37	Definitions of Probability
38	Application of Addition Theorem, Conditional Probability & Multiplication Theorem
39	Bayes' Theorem, Discrete Random Variable
40	Graphical Representation of the Distribution Function of a Discrete Random Variable
41	Properties of Expected Values, Covariance & Correlation & Discrete Probability Distributions
42	Binomial Distribution, Introduction to the Hypergeometric Distribution
43	Hypergeometric Distribution, Poisson Distribution & Continuous Uniform Distribution
44	Appropriate Use of Measures of Location
45	Quantiles (Cut Points)
46	Examples of Quantiles
47	Measures of Dispersion
48	Absolute Measures of Dispersion
49	Examples for Absolute Measures of Dispersion
50	Relative Measures of Dispersion
51	Examples of Measures of Dispersion
52	Appropriate usage of the Measures of Dispersion
53	Calculating Skewness and making decision using Numerical measures
54	Calculating Kurtosis and making decision using Numerical measures
55	Counting Rules
56	Permutations
57	Combinations
58	Bernoulli Distribution
59	Bernoulli Trials and Bernoulli Distribution
60	Standard Error
61	Sampling Distribution of the sample mean
62	Expected value and variance of the sampling distribution of the sample mean
63	Estimation
64	Point Estimation
65	Interval Estimation
66	Correlation
67	Assumptions used for Correlation
68	Types of correlation and their usage
69	Partial Correlation
70	Multiple Correlation
71	Calculating and interpreting Correlation

72	Testing Hypothesis about Correlation coefficient
73	Regression
74	Assumptions for Regression Model
75	Steps in Regression Analysis
76	A Tutorial to SPSS - I
77	A Tutorial to SPSS - II
78	A Tutorial to SPSS - III
79	A Tutorial to SPSS - IV
80	A Tutorial to SPSS - V
81	A Tutorial to SPSS - VI
82	A Tutorial to SPSS - VII
83	A Tutorial to SPSS - VIII
84	A Tutorial to SPSS - IX
85	A Tutorial to SPSS - X
86	A Tutorial to SPSS - XI
87	A Tutorial to SPSS - XII
88	A Tutorial to SPSS - XIII

Topic 1

Introduction to Biostatistics

We shall define biostatistics as the **application of statistical methods to the solution of biological problems**. The biological problems of this definition are those arising in the basic biological sciences as well as in such applied areas as the health-related sciences and the agricultural sciences. Biostatistics is also called **biological statistics or biometry**.

The definition of biostatistics leaves us somewhat up in the air-"statistics" has not been defined. Statistics is a science well known by name even to the layman. The number of definitions you can find for it is limited only by the number of books you wish to consult. We might define **statistics in its modern sense as the scientific study of numerical data based on natural phenomena**.

All parts of this definition are important and deserve emphasis: Scientific study: Statistics must meet the commonly accepted criteria of validity of scientific evidence. We must always be objective in presentation and evaluation of data and adhere to the general ethical code of scientific methodology, or we may find that the old saying **that "figures never lie, only statisticians do"** applies to us.

Biostatistics, in its present form, is the cumulative result of four centuries of contributions from many mathematicians and scientists. Some are well known, and some are obscure; some are famous people you never would've suspected of being statisticians, and some are downright eccentric and unsavory characters. This list gives you some highlights of the contributions of a few of the many people who made statistics (and therefore biostatistics) what it is today.

- Thomas Bayes (ca. 1701–1761). ...
- Pierre-Simon LaPlace (1749–1827). ...
- Carl Friedrich Gauss (1777–1855). ...
- John Snow (1813–1858). ...
- Florence Nightingale (1820–1910). ...
- Karl Pearson (1857–1936). ...
- William S. Gosset (1876–1937).
- Ronald A. Fisher (1890–1962)
- John W. Tukey (1915–2000)
- David R. Cox (1924–).

Applications of Biostatistics:

Various uses of Biostatistics and their interpretation in a practical manner over different areas are as below:

1. System Biology
2. Public Health Studies related to Nutrition,
3. Health Service Research,
4. Environmental Health
5. Epidemiology
6. To design and analyze the clinical operations
7. Assessment of effects of any disease over the
8. Patients and its outcomes
9. Ecology and its forecasting
10. Biological sequence analysis
11. Genomics data analysis and studies
12. Experiments on diseases and their effects
13. Gene network analysis and biological reasons.

Topic 2

Basic Terminology - I

Population

The average person thinks of a population as a collection of entities, usually people.

A population or collection of entities may, however, consist of animals, machines, places, or cells. For our purposes, we define a population of entities as the largest collection of entities for which we have an interest at a particular time.

If we take a measurement of some variable on each of the entities in a population, we generate a population of values of that variable. We may, therefore, define **a population of values as the largest collection of values of a random variable for which we have an interest at a particular time.**

If, for example, we are interested in the weights of all the children enrolled in a certain county elementary school system, our population consists of all these weights. If our interest lies only in the weights of first-grade students in the system, we have a different population—weights of first-grade students enrolled in the school system. Hence, populations are determined or defined by our sphere of interest.

Sample:

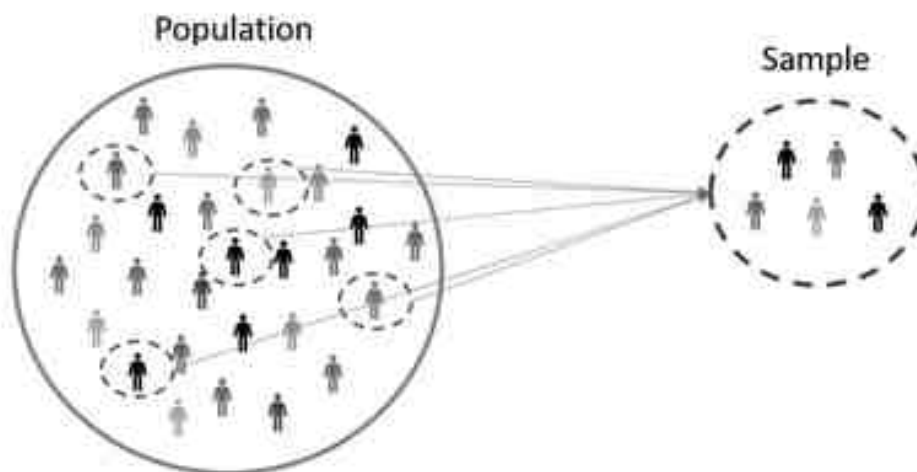
A sample may be defined simply as a part of a population. Suppose our population consists of the weights of all the elementary school children enrolled in a certain county school system. If we collect for analysis the weights of only a fraction of these children, we have only a part of our population of weights, that is, we have a sample

It is clearly impossible for most scientific studies to survey all individuals of a group about which conclusions are to be drawn. Cost and time considerations force the researcher to settle for studying a subset of a group to form conclusions about the entire group. For example, if one wished to know the systolic, diastolic, and mean blood pressures of the entire population of the United States, this testing could be done as part of the next census, in which all members of the population would be surveyed. However, the cost of training census takers for this time-consuming task and the need to hire additional census takers because of the increased time required to check everyone would make sampling the entire population an impractical task. Similarly, in biomedical research, time and financial costs preclude study of every member of the population. However, if a representative sample of an appropriate population can be obtained and studied, conclusions about the sample can be properly extrapolated to the defined population.

Difference between Population and Sample:

POPULATION	SAMPLE
▪ The measurable quality is called a parameter.	▪ The measurable quality is called a statistic.
▪ The population is a complete set.	▪ The sample is a subset of the population.
▪ Reports are a true representation of opinion.	▪ Reports have a margin of error and confidence interval.
▪ It contains all members of a specified group.	▪ It is a subset that represents the entire population.

Presentation of Population and Sample:



Type of population:

Populations may be

Finite:

The finite population is also known as a countable population in which the population can be counted. In other words, it is defined as the population of all the individuals or objects that are finite. For statistical analysis, the finite population is more advantageous than the infinite population. Examples of finite populations are employees of a company, potential consumer in a market.

Infinite

The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

Some other concepts which are important to know:

Target Population:

The general population that the study seeks to understand.

Source population:

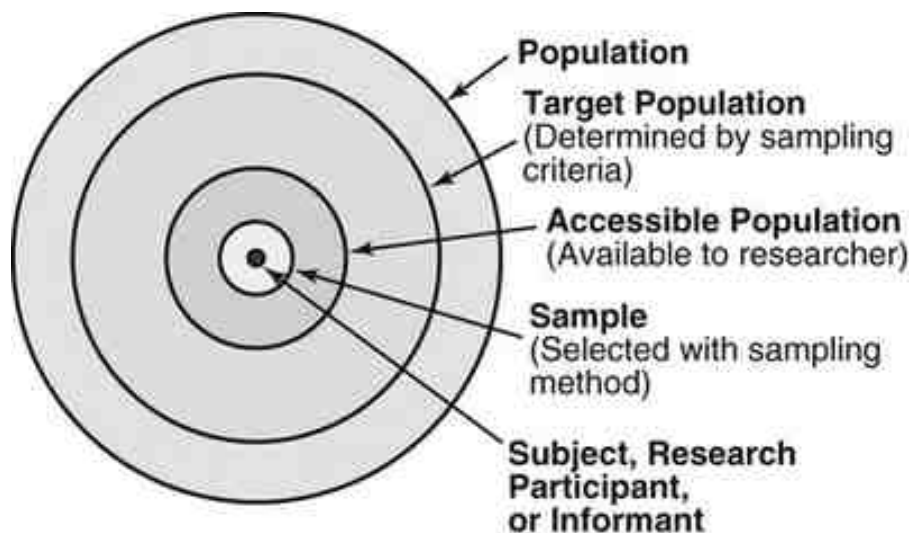
The specific individuals from which a representative sample will be drawn.

Sample Population:

Individuals ask to participate.

Study population:

Eligible participants.



Census:

A census is a survey conducted on the full set of observation objects belonging to a given population or universe.

Census, an enumeration of people, houses, firms, or other important items in a country or region at a particular time. Used alone, the term usually refers to a population census.

Population parameter:

You may have already heard of parameters in math, which are values that are *held constant* for a given mathematical function. In statistics, the definition of parameter is different. A parameter is data that refers to something about an **entire population**.

A **parameter** is any summary number, like an average or percentage, that describes the entire population.

The population mean μ (the greek letter "mu") and the population proportion p are two different population parameters. For example:

- We might be interested in learning about μ , the average weight of all middle-aged female Americans. The population consists of all middle-aged female Americans, and the parameter is μ .
- Or, we might be interested in learning about p , the proportion of likely American voters approving of the president's job performance. The population comprises all likely American voters, and the parameter is p .

Topic 3

Basic Terminology – II

A **statistic** refers to measures about the sample, while a **parameter** refers to measures about the population.

Examples of statistics vs parameters	
Sample statistic	Population parameter
Proportion of 2000 randomly sampled participants that support the death penalty.	Proportion of all US residents that support the death penalty.
<u>Median</u> income of 850 college students in Boston and Wellesley.	Median income of all college students in Massachusetts.
Standard deviation of weights of avocados from one farm.	Standard deviation of weights of all avocados in the region.
Mean screen time of 3000 high school students in India.	Mean screen time of all high school students in India.

Statistical Notations:

Different symbols are used for statistics versus parameters to show whether a sample or a population is being referred to. Greek letters and capital letters usually refer to populations, whereas Latin letters and lower-case letters refer to samples.

Symbols for statistics vs parameters		
	Sample statistic	Population parameter
Proportion	$\hat{p}$ (called “p-hat”)	P
Mean	$\bar{x}$ (called “x-bar”)	μ (Greek letter “mu”)
Standard deviation	s (Latin letter “s”)	σ (Greek letter “sigma”)
Variance	s^2	σ^2

Characteristic of a good Sample:

- **It should be random:** A random sample is one in which each item in the sample has an equal chance of being selected. Statistical inferences may not be valid unless the sample is random.
- **It should be representative:** the sample should be representative of the differing items in the whole population. For example it should contain a similar proportion of high and low value items of the population.
- **Protective:** Protective of the auditor. More intensive auditing should occur on high value items known to be high risk.
- **Unpredictable** - client should not be able to know which items will be examined.

Topic 4

Data

The raw material of statistics is data. For our purposes we may define data as numbers. The two kinds of numbers that we use in statistics are numbers that result from the taking—in the usual sense of the term—of a measurement, and those that result from the process of counting.

For example, when a nurse weighs a patient or takes a patient's temperature, a measurement, consisting of a number such as 150 pounds or 100 degrees Fahrenheit, is obtained. Quite a different type of number is obtained when a hospital administrator counts the number of patients—perhaps 20—discharged from the hospital on a given day. Each of the three numbers is a datum, and the three taken together are data.

Statistical data: When it means statistical data it refers to numerical descriptions of things. These descriptions may take the form of counts or measurements. Thus statistics of malaria cases in one of malaria detection and treatment posts of Ethiopia include fever cases, number of positives obtained, sex and age distribution of positive cases, etc.

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.

Types of Data:

There are two types of data

Primary Data:

Primary data is the kind of data that is collected directly from the data source without going through any existing sources. It is mostly collected specially for a research project and may be shared publicly to be used for another research.

Primary data is often reliable, authentic, and objective in as much as it was collected with the purpose of addressing a particular research problem. It is noteworthy that primary data is not commonly collected because of the high cost of implementation.

Secondary Data:

Secondary data is the data that has been collected in the past by someone else but made available for others to use. They are usually once primary data but become secondary when used by a third party.

Secondary data are usually easily accessible to researchers and individuals because they are mostly shared publicly. This, however, means that the data are usually general and not tailored specifically to meet the researcher's needs as primary data does.

BASIS FOR COMPARISON	PRIMARY DATA	SECONDARY DATA
Meaning	Primary data refers to the first hand data gathered by the researcher himself.	Secondary data means data collected by someone else earlier.
Data	Real time data	Past data
Process	Very involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview, etc.	Government publications, websites, books, journal articles, internal records etc.
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Always specific to the researcher's needs.	May or may not be specific to the researcher's need.
Available in	Crude form	Refined form

**BASIS FOR
COMPARISON****PRIMARY DATA****SECONDARY DATA**

Accuracy and
Reliability

More

Relatively less

Distinguish between primary sources of data and secondary sources of data:

Primary data refers to the firsthand data gathered by the researcher himself. Sources of primary data are surveys, observations, questionnaires, and interviews as explained below:

Survey: Survey method is one of the primary sources of data which is used to collect quantitative information about items in a population. Surveys are used in different areas for collecting the data even in public and private sectors. A survey may be conducted in the field by the researcher. The respondents are contacted by the research person personally, telephonically or through mail. This method takes a lot of time, efforts, and money but the data collected are of high accuracy, current and relevant to the topic. When the questions are administered by a researcher, the survey is called a structured interview or a researcher-administered survey.

Observations: Observation as one of the primary sources of data. Observation is a technique for obtaining information involves measuring variables or gathering of data necessary for measuring the variable under investigation. Observation is defined as accurate watching and noting of phenomena as they occur in nature with regards to cause-and-effect relation. Interview: Interviewing is a technique that is primarily used to gain an understanding of the underlying reasons and motivations for people's attitudes, preferences, or behavior. Interviews can be undertaken on a personal one-to-one basis or in a group.

Questionnaires: Questionnaire as one of the primary sources of data is an observational technique which comprises series of items presented to a respondent in a written form, in which the individual is expected to respond in writing. Here the respondents are given list of written items which he responds to by ticking the one he considers appropriate.

While secondary sources mean data collected by someone else earlier. Secondary data are the data collected by a party not related to the research study but collected these data for some other purpose and at different time in the past. If the researcher uses these data, then these become secondary data for the current users. Sources of secondary data are government publications websites, books, journal articles, internal records.

Topic 5

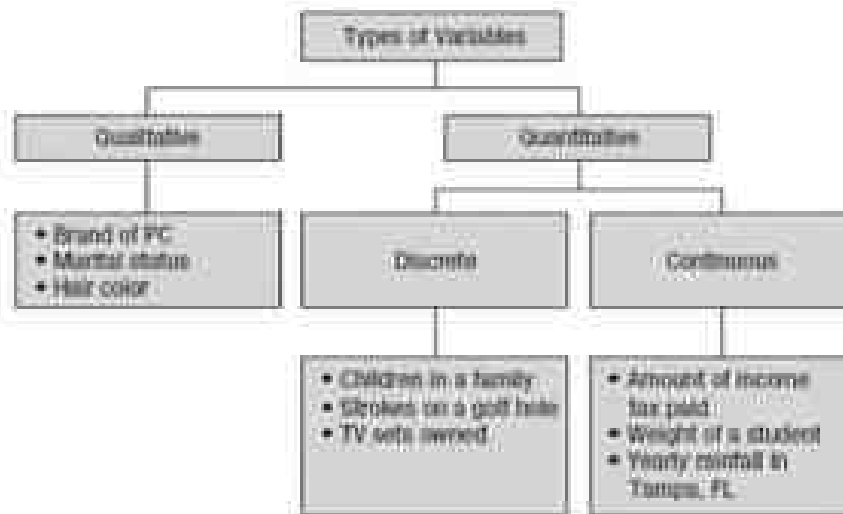
Variables

If, as we observe a characteristic, we find that it takes on different values in different persons, places, or things, we label the characteristic a variable. We do this for the simple reason that the characteristic is not the same when observed in different possessors of it. Some examples of variables include diastolic blood pressure, heart rate, the heights of adult males, the weights of preschool children, and the ages of patients seen in a dental clinic.

Qualitative variable: a variable or characteristic which cannot be measured in quantitative form but can only be identified by name or categories, for instance place of birth, ethnic group, type of drug, stages of breast cancer (I, II, III, or IV), degree of pain (minimal, moderate, severe or unbearable).

Quantitative variable: A quantitative variable is one that can be measured and expressed numerically, and they can be of two types (discrete or continuous). The values of a discrete variable are usually whole numbers, such as the number of episodes of diarrhea in the first five years of life. A continuous variable is a measurement on a continuous scale. Examples include weight, height, blood pressure, age, etc.

Summary of Types of Variables



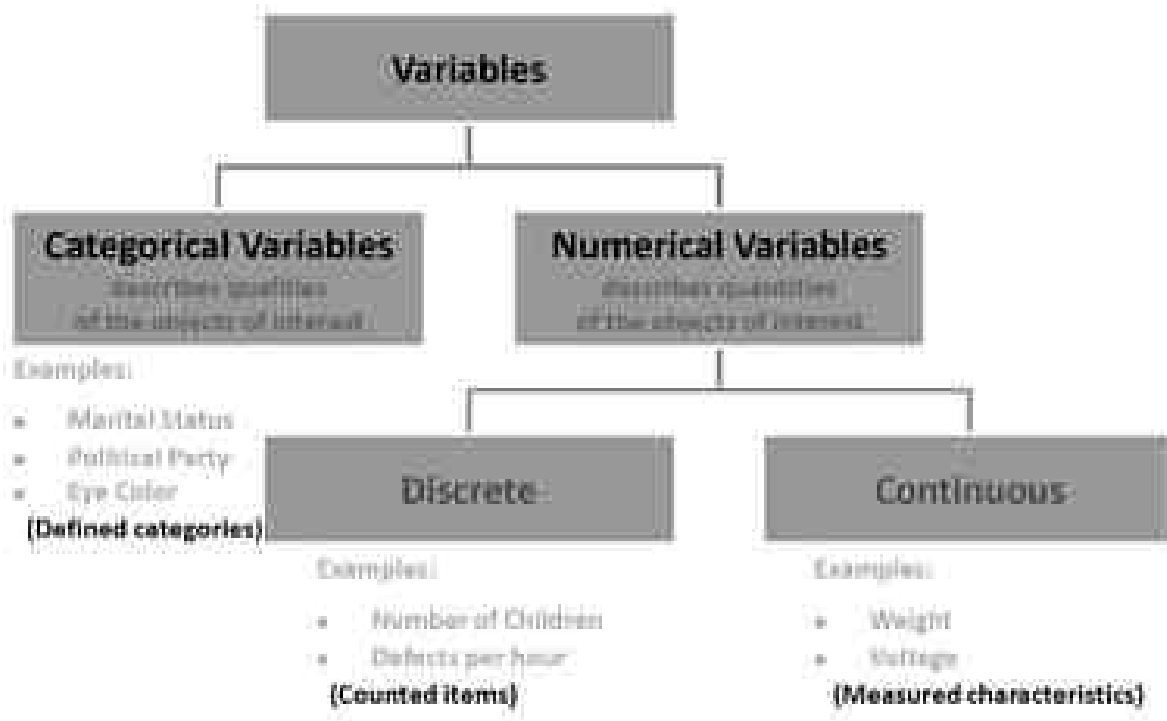
- Quantitative variable may be classified as discrete or continuous.

1. 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.

2. 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 i^{th} 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.



Topic 6

Measurement Scales – I

Normally, when one hears the term measurement, they may think in terms of measuring the length of something (ie. the length of a piece of wood) or measuring a quantity of something (ie. a cup of flour). This represents a limited use of the term measurement. In statistics, the term measurement is used more broadly and is more appropriately termed scales of measurement. Scales of measurement refer to ways in which variables/numbers are defined and categorized. Each scale of measurement has certain properties which in turn determines the appropriateness for use of certain statistical analyses.

Properties of measurement scales:

Each scale of measurement satisfies one or more of the following properties of measurement.

- **Identity.** Each value on the measurement scale has a unique meaning.
- **Magnitude.** Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
- **Equal intervals.** Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
- **A minimum value of zero.** The scale has a true zero point, below which no values exist.

Types of Measurement Scales:

There are four types of measurement scales

- Nominal
- Ordinal
- Interval
- Ratio

1. Nominal data:-

A nominal scale is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.

In other words, Data that represent categories or names. There is no implied order to the categories of nominal data. In these types of data, individuals are simply placed in the proper category or group, and the number in each category is counted. Each item must fit into exactly one category. The simplest data consist of unordered, dichotomous, or "either - or" types of observations, i.e., either the patient lives or the patient dies, either he has some particular attribute, or he does not. e.g. Nominal scale data: survival status of propranolol - treated and control patients with myocardial infarction/

e.g. Nominal scale data: survival status of propranolol - treated and control patients with myocardial infarction

Status 28 days after hospital admission	Propranolol -treated patient	Control Patients
Dead	7	17
Alive	38	29
Total	45	46
Survival rate	84%	63%

The above table presents data from a clinical trial of the drug propranolol in the treatment of myocardial infarction. There were two group of myocardial infarctions. There were two group of patients with MI. One group received propranolol; the other did not and was the control. For each patient the response was dichotomous; either he survived the first 28 days after hospital admission, or he succumbed (died) sometime within this time period. With nominal scale data the obvious and intuitive descriptive summary measure is the proportion or percentage of subjects who exhibit the attribute. Thus, we can see from the above table that 84 percent of the patients treated with propranolol survived, in contrast with only 63% of the control group.

Some other examples of nominal data:

- **Eye color** - brown, black, etc.
- **Religion** - Christianity, Islam, Hinduism, etc
- **Sex** - male, female

2. Ordinal Data: -

The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the “order.” Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.

Example:

- Ranking of school students – 1st, 2nd, 3rd, etc.
- Ratings in restaurants
- Evaluating the frequency of occurrences
 - Very often
 - Often
 - Not often
 - Not at all

- Assessing the degree of agreement
 - Totally agree
 - Agree
 - Neutral
 - Disagree
 - Totally disagree

Characteristics of the Ordinal Scale

- The ordinal scale shows the relative ranking of the variables
- It identifies and describes the magnitude of a variable
- Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables
- The interval properties are not known
- The surveyors can quickly analyse the degree of agreement concerning the identified order of variables

Topic 7

Measurement Scales – II

3. Interval Data:-

The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

Characteristics of Interval Scale:

- The interval scale is quantitative as it can quantify the difference between the values
- It allows calculating the mean and median of the variables
- To understand the difference between the variables, you can subtract the values between the variables
- The interval scale is the preferred scale in Statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.

4. Ratio Data:-

The highest level of measurement is the ratio scale. This scale is characterized by the fact that equality of ratios as well as equality of intervals may be determined. Fundamental to the ratio scale is a true zero point. The measurement of such familiar traits as height, weight, and length makes use of the ratio scale.

Characteristics of Ratio Scale:

- Ratio scale has a feature of absolute zero
- It doesn't have negative numbers, because of its zero-point feature
- It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.
- Ratio scale has unique and useful properties. One such feature is that it allows unit conversions like kilogram – calories, gram – calories, etc.

Levels of Measurement

Nominal	Ordinal	Interval	Ratio
"Eye color"	"Level of satisfaction"	"Temperature"	"Height"
Named	Named	Named	Named
	Natural order	Natural order	Natural order
		Equal interval between variables	Equal interval between variables
			Has a "true zero" value, thus ratio between values can be calculated

Topic 8

Types of Statistics

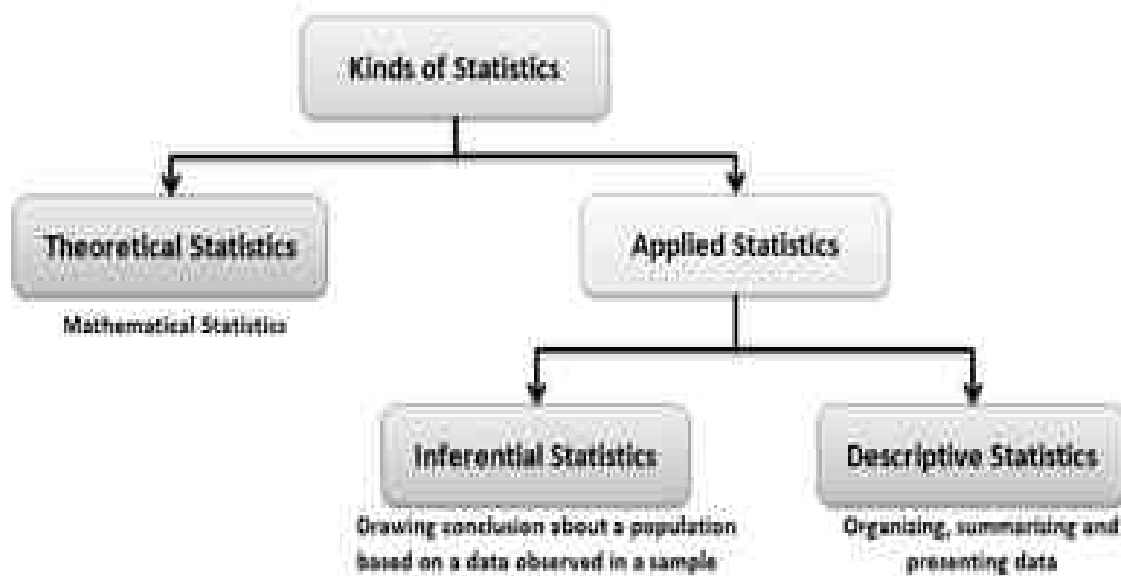
Statistics simply means numerical data and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data. It is a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life. It is an area of applied mathematics concern with data collection analysis, interpretation, and presentation. Statistics deals with how data can be used to solve complex problems. Some people consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Importance of Statistics:

- Statistics makes the work simple & provides a clear picture on the work we do on daily basis
- The statistical methods help us to research on different streams such as medicine, economics, business, social science and so on.
- Statistics provides us different types of organized data with the help of graphs, diagrams and charts
- Statistics comes handy while we do critical analysis.

➤ Statistics have majorly categorized into two types:

1. Descriptive statistics
2. Inferential statistics



Descriptive statistics

In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorized into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

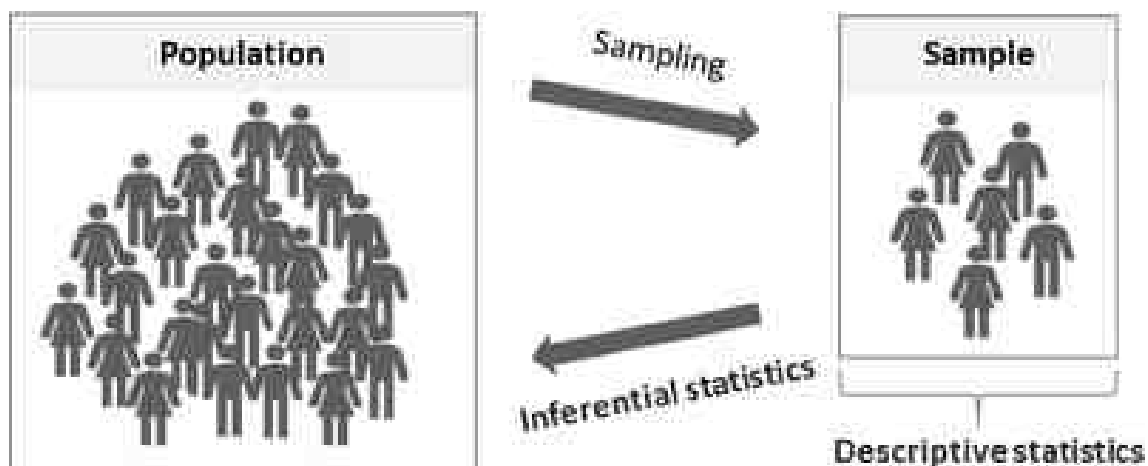
Topic 9

Sampling and Statistical Inference

Inferential statistics:

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed, and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to **draw conclusions from the** data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use **information collected from a sample to make decisions, predictions, or inferences from a population.** It grants us permission to give statements.



Importance of statistical inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data. The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

Statistical inference can be divided into two main branches

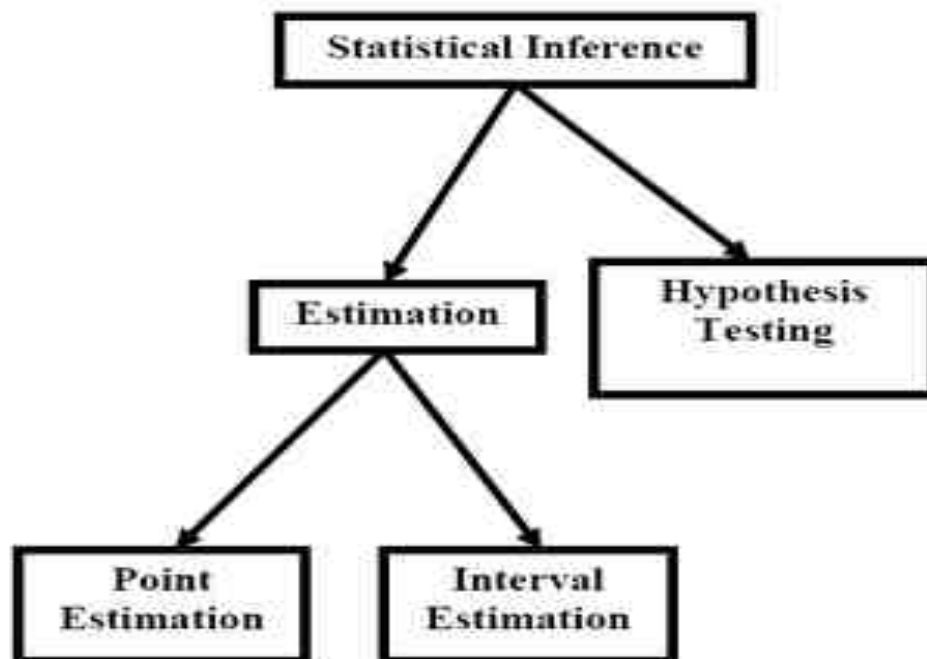
- Estimation
- Hypothesis-testing.

Presenting some basic concepts for statistical inference:

- **Estimand:** The parameter in the population which is unknown and that will be estimated in a statistical analysis.
- **Estimation:** the process of finding an **estimate**.
- **Estimate:** the estimated quantity of interest representing the estimand given some assumptions and data.
- **Estimator:** A *rule* for calculating an estimate of a given quantity based on observed data.

Estimation itself can be further divided into two branches

- Point estimation
- Interval 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.

The sample mean $\bar{x}$ is a point estimate of the population mean μ . Similarly, the sample proportion p is a point estimate of the population proportion P .

For example: Based on sample results, we estimate that p , the proportion of all U.S. adults who are in favor of stricter gun control, is 0.6.

More importantly, point estimates and parameters represent fundamentally different things.

- Point estimates are calculated from the data; parameters are not.
- Point estimates vary from study to study; parameters do not.
- Point estimates are random variables; parameters are constants

Interval estimation

we estimate an unknown parameter using an **interval of values** that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter).

For example: Based on sample results, we are 95% confident that p , the proportion of all U.S. adults who are in favor of stricter gun control, is between 0.57 and 0.63.

Sampling Distribution

The second important point is that the concept of sampling distributions forms the basis for both estimation and hypothesis-testing,

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 without replacement.

Sampling distribution of the mean:

Although point estimate $\bar{x}$ is a valuable reflection of parameter μ , it provides no information about the precision of the estimate. We ask: How precise is $\bar{x}$ as estimate of μ ? How much can we expect any given $\bar{x}$ to vary from μ ?

The variability of $\bar{x}$ as the point estimate of μ starts by considering a hypothetical distribution called the **sampling distribution of a mean** (SDM for short). Understanding the SDM is difficult because it is based on a thought experiment that doesn't occur in actuality, being a hypothetical distribution based on mathematical laws and probabilities. The SDM imagines what would happen if we took repeated samples of the same size from the same (or similar) populations done under

the identical conditions. From this hypothetical experiment we “build” a pmf or pdf that is used to determine probabilities for various hypothetical outcomes.

Without going into too much detail, the SDM reveals that:

- $\bar{x}$ is an unbiased estimate of μ ;
- the SDM tends to be normal (Gaussian) when the population is normal or when the sample is adequately large;

- the standard deviation of the SDM is equal to $\sigma/\sqrt{n}$. This statistic—which is called the **standard**

error of the mean (SEM)—predicts how closely the $\bar{x}$ s in the SDM are likely to cluster around the value of μ and is a reflection of the precision of $\bar{x}$ as an estimate of μ :

$$\text{SEM} = \sigma / \sqrt{n}$$

Note that this formula is based on σ and not on sample standard deviation s . Recall that σ is NOT calculated from the data and is derived from an external source. Also note that the SEM is inversely proportion to the square root of n . 

Confidence Intervals:

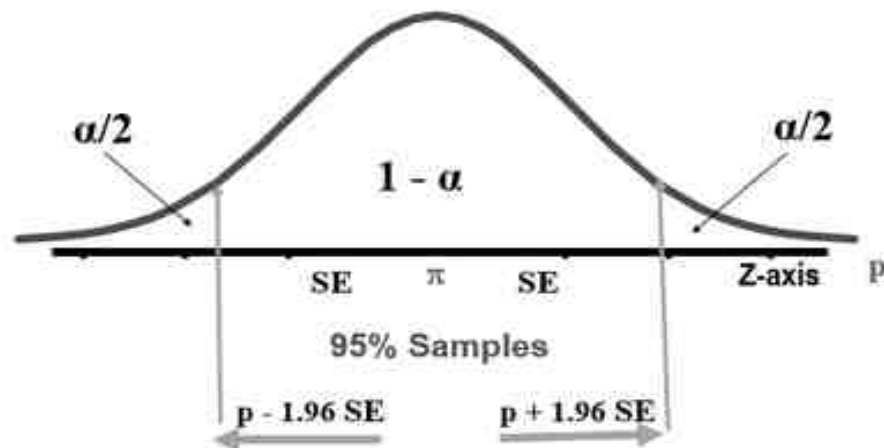
Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic $\pm$ margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic $\pm$ margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.



Hypothesis testing:

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 Hypothesis.

Null Hypothesis

A null hypothesis, generally denoted by the symbol H_0 , 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 inches. This statement is taken as a hypothesis and is written symbolically as $H_0 : \mu = 62''$. 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 H_0 is rejected. It is customarily denoted by H_1 or H_A . A null hypothesis H_0 is thus tested against an alternative hypothesis H_1 . For example, if our null hypothesis is $H_0 : \mu = 62''$, then our alternative hypothesis may be $H_1 : \mu \neq 62''$ or $H_1 : \mu < 62''$.

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, H_0 is either actually true, or it is false. Also, as far as our decision regarding H_0 is concerned, there are two

possibilities --- either we will accept H_0 , or we will reject H_0 . 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)

TYPE-I AND TYPE-II ERRORS

On the basis of sample information, we may reject a null hypothesis H_0 , when it is, in fact, true or we may accept a null

hypothesis H_0 , 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

$$\alpha = P(\text{Type I error})$$

$$= P(\text{reject } H_0 | H_0 \text{ is true}),$$

$$\beta = P(\text{Type II error})$$

$$= P(\text{accept } H_0 | H_0 \text{ is false}).$$

Factors Increasing type II Error:

True Value of Population Parameter

- Increases When Difference Between Hypothesized Parameter & True Value Decreases

Significance Level α

- Increases When α Decreases

Population Standard Deviation σ

- Increases When σ Increases

Sample Size n

- Increases When n Decreases

P value:

When you perform a statistical test a p -value helps you determine the significance of your results in relation to the null hypothesis.

The null hypothesis states that there is no relationship between the two variables being studied (one variable does not affect the other). It states the results are due to chance and are not significant in terms of supporting the idea being investigated. Thus, the null hypothesis assumes that whatever you are trying to prove did not happen.

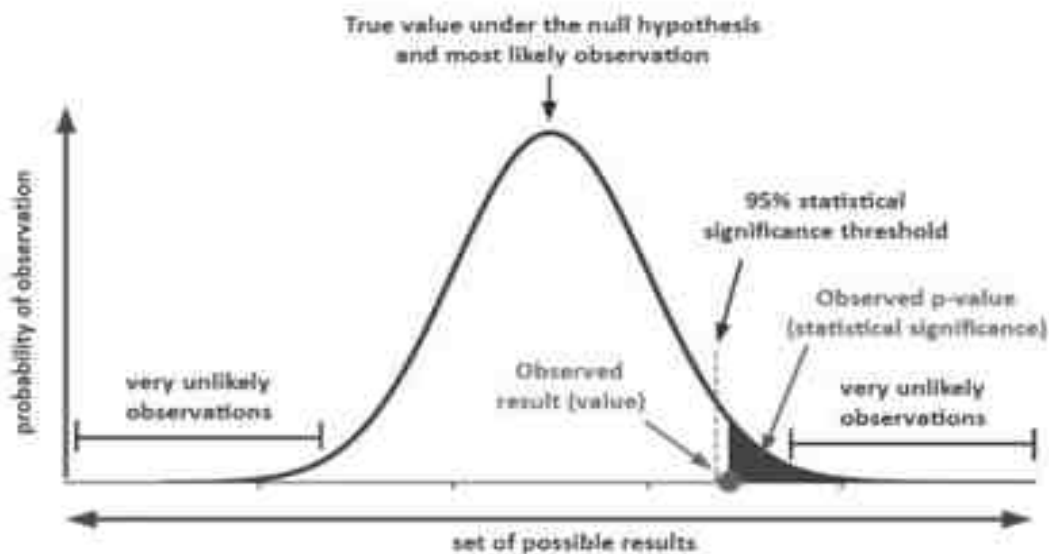
The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The alternative hypothesis states that the independent variable did affect the dependent variable, and the results are significant in terms of supporting the theory being investigated (i.e. not due to chance).

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p -value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That's pretty tiny. On the other hand, a large p -value of .9(90%) means your results have a 90% probability of being completely random and *not* due to anything in your experiment. Therefore, the smaller the p -value, the more important ("significant") your results.

When you run a hypothesis test, you compare the p value from your test to the alpha level you selected when you ran the test. Alpha levels can also be written as percentages.

graphically, the p value is the area in the tail of a probability distribution. It's calculated when you run hypothesis test and is the area to the right of the test statistic (if you're running a two-tailed test, it's the area to the left *and* to the right)



P value vs Alpha level:

Alpha levels are controlled by the researcher and are related to confidence levels. You get an alpha level by subtracting your confidence level from 100%. For example, if you want to be 98 percent confident in your research, the alpha level would be 2% (100% – 98%). When you run the hypothesis test, the test will give you a value for p. Compare that value to your chosen alpha level. For example, let's say you chose an alpha level of 5% (0.05). If the results from the test give you:

- A small p (≤ 0.05), reject the null hypothesis. This is strong evidence that the null hypothesis is invalid.
- A large p (> 0.05) means the alternate hypothesis is weak, so you do not reject the null.

Hypothesis Testing: Steps:

Test the Assumption that the true mean SBP of participants is 120 mmHg.

State H_0 $H_0: m = 120$

State H_1 $H_1: m \neq 120$

Choose α $\alpha = 0.05$

Choose n $n = 100$

Choose Test: Z, t, X^2 Test (or p Value)

Compute Test Statistic (or compute P value)

Search for Critical Value

Make Statistical Decision rule

Express Decision

One sample-mean Test

Assumptions

- ❖ Population is normally distributed

t test statistic

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Example Normal Body Temperature:

What is **normal body temperature**? Is it actually 37.6°C (on average)?

State the null and alternative hypotheses

$H_0: m = 37.6^\circ\text{C}$

$H_a: m \neq 37.6^\circ\text{C}$

Data: random sample of $n = 18$ normal body temps

37.2	36.8	38	37.6	37.2	36.8	37.4	38.7	37.2
36.4	36.6	37.4	37	38.2	37.6	36.1	36.2	37.5

Summarize data with a test statistic

Variable	n	Mean	SD	SE	t	P
Temperature	18	37.22	0.68	0.161	2.38	0.029

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Students's t Distribution Table:

Degrees of freedom	Probability (p value)		
	0.10	0.05	0.01
1	6.314	12.706	63.657
5	2.015	2.571	4.032
10	1.813	2.228	3.169
17	1.740	2.110	2.898
20	1.725	2.086	2.845
24	1.711	2.064	2.797
25	1.708	2.060	2.787
∞	1.645	1.960	2.576

Find the p -value

$$Df = n - 1 = 18 - 1 = 17$$

From SPSS: p -value = 0.029

From t Table: p -value is between 0.05 and 0.01.

Area to left of $t = -2.11$ equals area to right of $t = +2.11$.

The value $t = 2.38$ is between column headings 2.110 & 2.898 in table, and for $df=17$, the p -values are 0.05 and 0.01.

Decide whether or not the result is statistically significant based on the p -value

Using $\alpha = 0.05$ as the level of significance criterion, the results are **statistically significant** because 0.029 is less than 0.05. In other words, we can reject the null hypothesis.

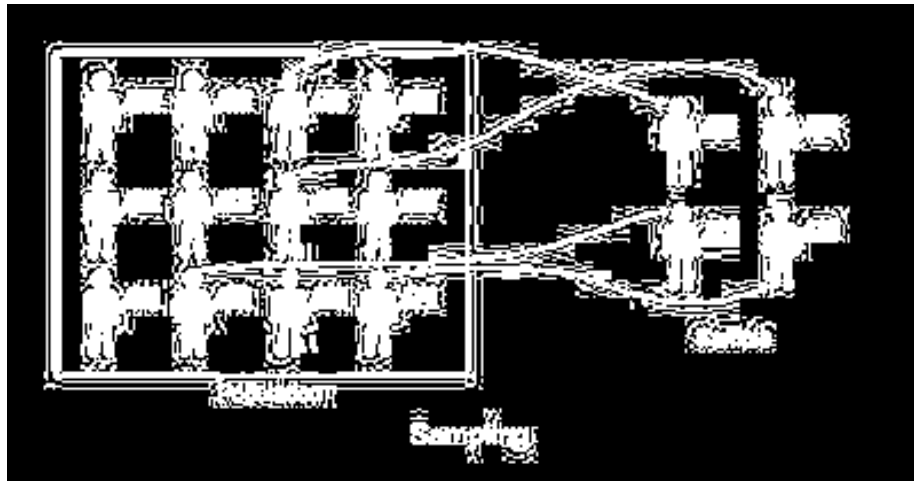
Report the Conclusion

We can conclude, based on these data, that the mean temperature in the human population does not equal 37.6.

Topic 10

Basic Terminology for sampling

Statistics is the science of data. Data are the numerical values containing some information. Statistical tools can be used on a data set to draw statistical inferences. These statistical inferences are in turn used for various purposes. For example, the government uses such data for policy formulation for the welfare of the people, marketing companies use the data from consumer surveys to improve the company and to provide better services to the customer, etc. Such data is obtained through sample surveys. Sample surveys are conducted throughout the world by governmental as well as nongovernmental agencies. For example, “National Sample Survey Organization (NSSO)” conducts surveys in India, “Statistics Canada” conducts surveys in Canada, agencies of United Nations like “World Health Organization (WHO), “Food and Agricultural Organization (FAO)” etc. conduct surveys in different countries.



Sampling theory provides the tools and techniques for data collection, keeping in mind the objectives to be fulfilled and the nature of the population. There are two ways of obtaining the information

1. Sample surveys
2. Complete enumeration or census

Sample surveys collect information on a fraction of the total population, whereas census collects information on the whole population. Some surveys, e.g., economic surveys, agricultural surveys etc. are conducted regularly. Some surveys are need-based and are conducted when some need arises, e.g., consumer satisfaction surveys at a newly opened shopping mall to see the satisfaction level with the amenities provided in the mall .

Sampling unit:

An element or a group of elements on which the observations can be taken is called a sampling unit. The objective of the survey helps in determining the definition of the sampling unit.

For example, if the objective is to determine the total income of all the persons in the household, then the sampling unit is a household. If the objective is to determine the income of any particular person in the household, then the sampling unit is the income of the particular person in the household. So the definition of sampling unit depends and varies as per the objective of the survey. Similarly, in another example, if the objective is to study the blood sugar level, then the sampling unit is the value of the blood sugar level of a person. On the other hand, if the objective is to study the health conditions, then the sampling unit is the person on whom the readings on the blood sugar level, blood pressure and other factors will be obtained. These values will together classify the person as healthy or unhealthy.

Population:

Collection of all the sampling units in a given region at a particular point of time or a particular period is called the population. For example, if the medical facilities in a hospital are to be surveyed through the patients, then the total number of patients registered in the hospital during the time period of the survey will be the population. Similarly, if the production of wheat in a district is to be studied, then all the fields cultivating wheat in that district will constitute the population. The total number of sampling units in the population is the population size, generally denoted by N . The population size can be finite or infinite (N is large).

Census:

The complete count of the population is called a census. The observations on all the sampling units in the population are collected in the census. For example, in India, the census is conducted at every tenth year in which observations on all the persons staying in India is collected.

Sample:

One or more sampling units are selected from the population according to some specified procedure. A sample consists only of a portion of the population units. Such a collection of units is called the sample. In the context of sample surveys, a collection of units like households, people, cities, countries etc. is called a finite population. A census is a 100% sample, and it is a complete count of the population.

Representative sample: When all the salient features of the population are present in the sample, then it is called a representative sample.

It goes without saying that every sample is considered as a representative sample.

For example, if a population has 30% males and 70% females, then we also expect the sample to have nearly 30% males and 70% females.

In another example, if we take out a handful of wheat from 100 Kg. bag of wheat, we expect the same quality of wheat in hand as inside the bag. Similarly, it is expected that a drop of blood will give the same information as all the blood in the body.

Sampling frame:

The list of all the units of the population to be surveyed constitutes the sampling frame. All the sampling units in the sampling frame have identification particulars. For example, all the students in a particular university listed along with their roll numbers constitute the sampling frame. Similarly, the list of households with the name of the head of family or house address constitutes the sampling frame. In another example, the residents of a city area may be listed in more than one frame - as per automobile registration as well as the listing in the telephone directory.

Ways to ensure representativeness: There are two possible ways to ensure that the selected sample is representative.

1. Random sample or probability sample:

The selection of units in the sample from a population is governed by the laws of chance or probability. The probability of selection of a unit can be equal as well as unequal.

2. Non-random sample or purposive sample:

The selection of units in the sample from the population is not governed by the probability laws. For example, the units are selected on the basis of the personal judgment of the surveyor. The persons volunteering to take some medical test or to drink a new type of coffee also constitute the sample on non-random laws.

Another type of sampling is Quota Sampling. The survey, in this case, is continued until a predetermined number of units with the characteristic under study are picked up.

For example, in order to conduct an experiment for rare type of disease, the survey is continued till the required number of patients with the disease are collected.

Advantages of sampling:

Sampling has various benefits to us. Some of the advantages are listed below:

- Sampling saves time to a great extent by reducing the volume of data. You do not go through each of the individual items.
- Sampling Avoids monotony in works. You do not have to repeat the query again and again to all the individual data.
- When you have limited time, survey without using sampling becomes impossible. It allows us to get near-accurate results in much lesser time
- When you use proper methods, you are likely to achieve higher level of accuracy by using sampling than without using sampling in some cases due to reduction in monotony, data handling issues etc.
- By using sampling, you can get detailed information on the data even by employing small amount of resources

Sampling Errors:

A sampling error is a deviation in the sample statistics value versus the true population parameter value. Sampling errors occur because the sample is not representative of the population or is biased in some way. Even randomized samples will have some degree of sampling error because a sample is only an approximation of the population from which it is drawn.

If, for example , $\bar{x}$ is the mean obtained from a sample of size n and μ is the corresponding population parameter, then the difference between $\bar{x}$ and μ is sampling error, that is

$$\text{sampling error} = \bar{x} - \mu$$

Topic 11

Types of Sampling Methods

Variability control in sample surveys:

Variability control is an important issue in any statistical analysis. A general objective is to draw statistical inferences with minimum variability. There are various types of sampling techniques which are adopted in different conditions. These techniques help in controlling the variability at different stages.

Probability Sampling

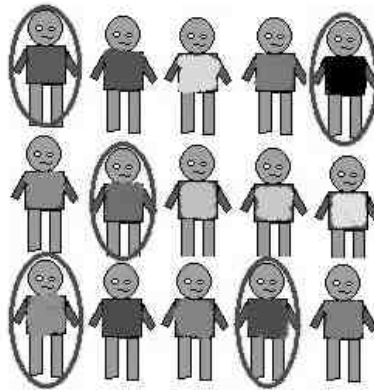
This Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample. It's alternatively known as random sampling. Such sampling techniques can be classified in the following way.

- Simple random sampling
- Stratified sampling
- Systematic sampling
- Cluster (area) sampling
- Multistage sampling



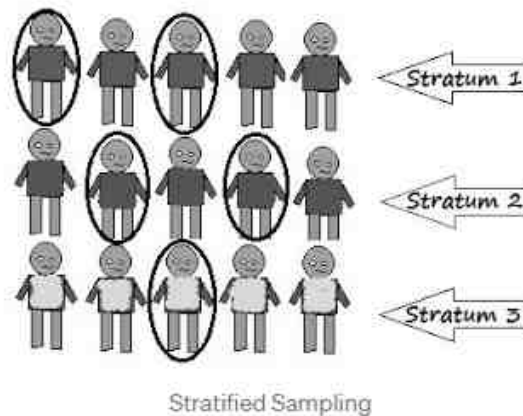
Simple Random Sampling: Every element has an equal chance of getting selected to be the part sample. It is used when we don't have any kind of prior information about the target population.

For example: Random selection of 20 students from class of 50 student. Each student has equal chance of getting selected. Here probability of selection is $1/50$



Stratified Sampling:

This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous and heterogeneous among the other subgroups formed. And then the elements are randomly selected from each of these strata. We need to have prior information about the population to create subgroups.



Stratified Sampling

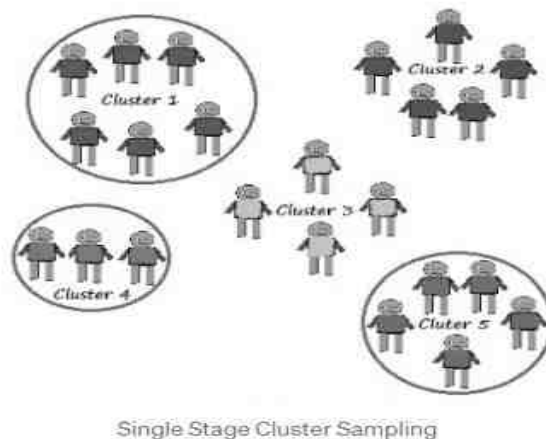
Cluster Sampling

Our entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling. Clusters are identified using details such as age, sex, location etc.

Cluster sampling can be done in following ways:

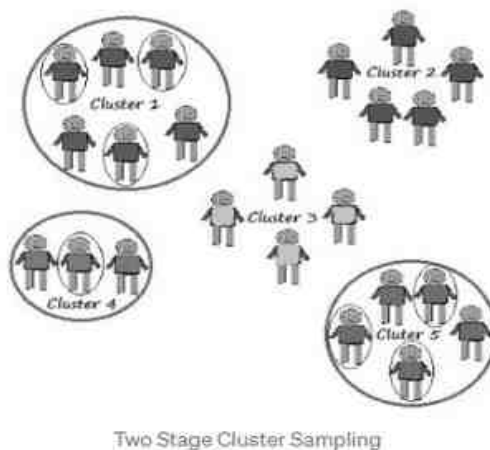
Single Stage Cluster Sampling

Entire cluster is selected randomly for sampling.



Two Stage Cluster Sampling

Here first we randomly select clusters and then from those selected clusters we randomly select elements for sampling



Systematic Clustering

Here the selection of elements is systematic and not random except the first element. Elements of a sample are chosen at regular intervals of population. All the elements are put together in a sequence first where each element has the equal chance of being selected.

For a sample of size n , we divide our population of size N into subgroups of k elements. We select our first element randomly from the first subgroup of k elements.

To select other elements of sample, perform following:
We know number of elements in each group is k i.e N/n

So if our first element is n_1 then

Second element is n_1+k i.e n_2

Third element n_2+k i.e n_3 and so on..

Taking an example of $N=20$, $n=5$

No of elements in each of the subgroups is N/n i.e $20/5 = 4 = k$

Now, randomly select first element from the first subgroup.

If we select $n_1 = 3$

$$n_2 = n_1 + k = 3 + 4 = 7$$

$$n_3 = n_2 + k = 7 + 4 = 11$$



Multi-Stage Sampling

It is the combination of one or more methods described above. Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity. One or more clusters can be randomly selected from each stratum. This process continues until the cluster can't be divided anymore. For example country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.

Topic 12

Simple Random Sampling

A sample is defined to be a simple random sample (SRS) if it is selected in such a manner that (i) each unit in the population has equal probability of being included in the sample and (ii) each possible sample of the same size has an equal probability of being the sample selected.

Suppose a finite population contains N units and a sample of size n is to be selected. If we sample with replacement, the number of all the possible samples of size n that could be selected is N^n , as the first unit of the sample can be selected in N different ways, the second unit can also be selected in N ways and so on. When the sample without replacement, then the number of all possible samples when the order of units is considered, is the number of permutations of n units from N , i.e. ${}^N P_n = N(N-1)\dots(N-n+1)$

But in the practical problem, we ignore the order in which units are drawn. Then the number of different samples of n units that can be selected when order is disregarded, is the number of combinations of n units from a finite population of N units, i.e., $\binom{N}{n} = \frac{N!}{n!(N-n)!}$. Thus there are

$\binom{N}{n}$ samples that could be selected and these samples occur with equal probabilities.

As an illustration, suppose we wish to select a random sample of size 2 from a population of say, 5 students, identified as A, B, C, D and E. If we sample with replacement, then there are $(5)^2 = 25$ possible samples, which are listed below

AA	BA	CA	DA	EA
AB	BB	CB	DB	EB
AC	BC	CC	DC	EC
AD	BD	CD	DD	ED
AE	BE	CE	DE	EE

Where the 1st letter corresponds to the student selected on 1st draw and the second letter corresponds to the student chosen on 2nd draw.

We see that A appears in 9 of 25 samples, so $P(A) = \frac{9}{25}$.

Similarly, $P(B) = P(C) = P(D) = P(E) = \frac{9}{25}$, showing that each unit of the population has an equal probability of being in the sample. Furthermore each of the 25 samples has an equal probability of $\frac{1}{25}$ of being selected.

When the sampling is done without replacement and order is disregarded there are $\binom{5}{2} = 10$ possible distinct samples which are listed below

AB, AC, AD, AE, BC, BD, BE, CD, CE, DE

Now A appears in 4 of the 10 samples, so $P(A) = \frac{4}{10}$

Similarly, $P(B) = P(C) = P(D) = P(E) = \frac{4}{10}$ showing that each unit in the population has an equal probability $\left(\frac{4}{10}\right)$ and each of the 10 distinct samples has the same probability of being $\left(\frac{1}{10}\right)$ of being selected.

a simple random sample is also known as unrestricted random sample an important advantage of simple random sampling is that it provides unbiased estimates of population mean population total and all sampling variants of the estimates

selection of simple random sample:

A simple random sample can be selected by the following methods

Goldfish Bowl Procedure:

Allot to each unit in the population at different serial numbers from one to N and record each number on a card or a slip of paper. Place these numbered cards or the folded slip of paper in a bowl or a basket and mix them thoroughly. Then draw out blindly the desired number of the card or the holding slip one by one for the sample, mixing thoroughly after each drawing. The population units corresponding to the numbers appearing on the selected cards or slips of papers are then included in the sample and the desired information obtained. This method of random selection works for populations but becomes particularly impossible when the population size is quite large or infinite. This difficulty is a welcome by a procedure, similar to drawing slips from a basket, that uses a random number table.

Using a random number table.

Assign a number from 1 to N to each of the N units in the population. consult a table of random numbers and select randomly a starting point in the table. Read digits in groups of two three or more according to the largest number assigned to a unit in the population, from the table vertically, horizontally and diagonally. record the number, discarding a number that is greater than N and that appears a second time if something is without replacement. Continue this process of selection until the desired sample size is reached.

Using a computer.

computer programs are available that provide random numbers. The sampling unit corresponding to the random numbers are included in the sample.

Example.

Assume that a population consist of 5 students and marks obtained by them in a certain statistics class are 20, 15, 12, 16 and 18. draw all possible random samples of two students when sampling is performed (i) with replacement, (ii) without replacement. calculate the mean marks for each sample

Solution:

Let the five students be identified as A, B, C, D, and E. Then (i) the number of possible random samples of two students which can be selected with replacement from this population is $(5)^2 = 25$

Let X_1 denotes the marks of students selected first and X_2 the marks of students selected on the second draw then the possible random samples off size $n = 2$ with values of $\bar{X}$ are given below.

Sample no.	Sample Students	Sample Marks	Sample Mean Marks
1	AA	20, 20	20
2	AB	20, 15	17.5
3	AC	20, 12	16
4	AD	20, 16	18
5	AE	20, 18	19
6	BA	15, 20	17.5
7	BB	15, 15	15
8	BC	15, 12	13.5
9	BD	15, 16	15.5
10	BE	15, 18	16.5
11	CA	12, 20	16
12	CB	12, 15	13.5
13	CC	12, 12	12
14	CD	12, 16	14
15	CE	12, 18	15
16	DA	16, 20	18
17	DB	16, 15	15.5
18	DC	16, 12	14
19	DD	16, 16	16
20	DE	16, 18	17
21	EA	18, 20	19
22	EB	18, 15	16.5
23	EC	18, 12	15
24	ED	18, 16	17
25	EE	18, 18	18

The number of random samples of 2 students that can be drawn without replacement is $\binom{5}{2} = 10$.

these samples with values of mean marks are given below:

Sample no.	Sample Students	Sample Marks	Sample Mean Marks
1	AB	20, 15	17.5
2	AC	20, 12	16
3	AD	20, 16	18
4	AE	20, 18	19
5	BC	15, 12	13.5
6	BD	15, 16	15.5
7	BE	15, 18	16.5
8	CD	12, 16	14
9	CE	12, 18	15
10	DE	16, 18	17

Topic 13

Stratified Random Sampling

A sample of size n is defined to be a stratified random sample if it is selected from a population which has been divided into a number of non-overlapping groups or sub populations called strata, such that part of the sample is drawn at random from each stratum. Stated differently, let a heterogeneous population containing N units to be divided into k subpopulations or strata of sizes $N_1, N_2, \dots, N_k$ in such a manner that all units in each stratum are believed to be very similar with respect to the elements of interest. Then a stratified random sample of size n will be composed of the simple random samples of predetermined sizes $n_1, n_2, \dots, n_k$ ($\sum n_i = n$) drawn independently from strata $1, 2, \dots, k$ respectively.

It is to be emphasized that stratification requires that each of all these strata should be internally homogeneous but externally the strata should differ from one another.

A population may be stratified according to size such as large, medium and small;

to the administrative grouping such as provinces and districts; to the geographic area such as urban and rural; Characteristics such as age group, sex, family size, occupation, education, taste of consumers etc. the advantages of stratified random samplings are

The advantages of stratified random sampling are low cost, greater accuracy and a better coverage. Stratified random sampling is used when (i) the variations among strata are greater than the variations within strata, (ii) information about some parts of the population is desired.

The purpose of stratified random sampling is three-fold. Firstly, the strata obtained by subdividing the heterogeneous population into homogeneous groups adequately represent the population so the information concerning individual stratum is gathered. Secondly, it provides improved estimates of the population characteristic. Thirdly, it reduces the variance of the estimator.

Allocation

of

sample

sizes:

By allocation of sample means the way the total sample size n is distributed among the various strata into which the population has been divided. Four methods of allocating the sample numbers are available. They are

- 1) Equal allocation:** The allocation is called equal when from each stratum an equal number of sampling units is selected. That is, the total sample size n is distributed equally among all the k strata. Thus sample size n_i for equal allocation is

$$n_i = \frac{n}{k} \quad \text{for } i = 1, 2, \dots, k$$

- 2) Proportional allocation:** The allocation is said to be proportional when the total sample size n is distributed among the different strata in proportion to the sizes of strata. In other words, the allocation is proportional if

$$n_i = n \cdot \frac{N_i}{N} \quad \text{for } i = 1, 2, \dots, k$$

Where N_i is the population size of the i th stratum, n_i is the i th stratum sample size and N is the total size of population. A sample size of n from stratum i is drawn by random numbers and

investigated. this way of allocation is the next simplest method and it is the most frequently used method. the advantage of proportional allocation is that it does not require information either on the stratum variance all on the cost of sampling units in different strata.

3) Optimum allocation: The allocation is called optimum when the total sample size n is allocated among the different strata in such a way that for a given cost of selecting the sample, the variance of the estimated mean $\bar{X}_{st}$, i.e. $Var(\bar{X}_{st})$ is minimized. the stratum sample size n_i for this method of allocation is

$$n_i = n \cdot \frac{N_i \sigma_i / \sqrt{c_i}}{\sum N_i \sigma_i / \sqrt{c_i}}$$

N_i is the population of i th stratum

σ_i is the stratum standard deviation, and

c_i is the cost of serving one unit in the i th stratum.

This method of allocation tell us that we should select a larger size sample from a given stratum if

- (i) the stratum comprises a larger number of units, i.e. N_i is larger,
- (ii) the variation within the stratum is greater, i.e. σ_i is bigger,
- (iii) the sampling units in the stratum are less costly to measure, i.e. c_i it's smaller

when the information on the stratum standard deviation, σ_i is not available we either estimate it by s_i or we may use the proportional location.

4) Neyman allocation: This method of allocation was proposed by J Neyman in 1934 and it consist of finding n_i which minimizes the variance of the stratified sample mean for a fixed total sample size n , assuming the cost of surveying the units, i.e. c_i is to be the same in all strata. the stratum sample size n_i is given by the

$$n_i = n \cdot \frac{N_i \sigma_i}{\sum N_i \sigma_i} \text{ for } i = 1, 2, \dots, k$$

Neyman allocation becomes the same as the proportion allocation when their stratum standard deviations are equal.

Example:

Suppose a population of $N = 9$ is stratified into 3 strata with the following measurements:

Stratum 1	$X_{11} = 1, X_{12} = 2, X_{13} = 4$
Stratum 2	$X_{21} = 6, X_{22} = 8$
Stratum 3	$X_{31} = 11, X_{32} = 15, X_{33} = 16, X_{34} = 19$

If two measurements are drawn from strata for the sample, state how many samples of size 6 could be chosen from this population?

List these samples and compute the mean for each sample.

Here the population consists of 3 strata and from each stratum 2 units are to be selected to make up a sample $n = 6$. Assuming the sampling without replacement, we can choose $\binom{3}{2} = 3$ possible

sub-samples from the first stratum, $\binom{2}{2} = 1$ possible subsample from the second stratum and

$\binom{4}{2} = 6$ possible subsamples from the third stratum.

Each of the possible subsamples from stratum 1 is to be associated with the subsample from stratum 2 and then these combinations are further associated with each of the possible subsamples from stratum 3.

Hence, in all there are $3 \times 1 \times 6$, i.e. 18 possible samples of size 6 with 2 measurements from each stratum.

The first sample consists of (1, 2, 6, 8, 11, 15) of which (1, 2) is from stratum 1, (6, 8) is from stratum 2 and (11, 15) is from stratum 3. The 18 possible samples are listed below and the sample means appear in the last column.

Sample No.	Sample data from stratum			Total	Sample Mean
	1	2	3		
1	1, 2	6, 8	11, 15	43	7.17
2	1, 2	6, 8	11, 16	44	7.33
3	1, 2	6, 8	11, 19	47	7.83
4	1, 2	6, 8	15, 16	48	8
5	1, 2	6, 8	15, 19	51	8.5
6	1, 2	6, 8	16, 19	52	8.67
7	1, 4	6, 8	11, 15	45	7.5
8			11, 16	46	7.67
9			11, 19	49	8.17
10			15, 16	50	8.33
11			15, 19	53	8.83
12			16, 19	54	9
13	2, 4	6, 8	11, 15	46	7.67
14			11, 16	47	7.83
15			11, 19	50	8.33
16			15, 16	51	8.5
17			15, 19	54	9
18			15, 19	55	9.17

Topic 14

Cluster Sampling

A random sample is said to be cluster sample if it consists of first selecting at random groups open individual units, called **clusters** (created as sample units) into which a population can be divided and then including in the sample either all the units from each of the chosen clusters, or selecting a random sample of the units which the cluster comprises.

In other words, suppose that a population is first divided into M similar groups (equal or unequal in size), called clusters, such as the blocks of the cities, households, classes, etc. A random sample consisting of a number of clusters (say, m) is then selected from all these clusters where each chosen cluster is either subsampled or all the subunits in the sampled clusters are included in the sample. Such a sample is called a cluster sample.

Cluster Sampling requires that the clusters should be as internally dissimilar as possible and different clusters should be very similar.

The procedure is called **One stage cluster sampling** when all the units which each of the sampled cluster comprises are included. If each of the sampled cluster is subsampled, then the sampling plan is called **Two stage cluster sampling** or **Subsampling**. The plan is called **Multistage cluster sampling** when more than two stages are involved in making the sample. When the clusters related to geographical areas, the sampling is known as area sampling.

As each cluster is treated as a single sampling unit in the selection process, the cluster are therefore called the **Primary sampling units** (psu) while the subunits comprising a cluster, are called the **Secondary sampling units** (ssu).

The advantage of cluster sampling is saving in cost and time, i.e., the cost of sample selection and travel expenses of interviewers are considerably low. The master sampling is used when (i) sampling frames of adequate coverage are not available (ii) the variation among clusters is smaller than the variations within clusters. It is to be noted that cluster sampling is mostly used in statistical quality control.

Topic 15

Systematic Sampling

A sample of size n is defined to be systematic random sample it is obtained by choosing one unit at random from the first k units and thereafter selecting every k th unit after the N units in the population have been serially numbered from 1 to N or arranged in a systematic fashion. The letter

k , call to sampling interval, stands for some integer nearest to $\frac{N}{n} = \frac{\text{Population size}}{\text{Sample size}}$ and sample is

generally expressed by saying “a 1 in k samples”. For example, when the number of units selected at random from the first k units is i ($i = 1, 2, \dots, k$), the systematic sample of size n will contain the units with the numbers $i + k, i + 2k, \dots, i + (n-1)k$ suppose that $i = 7$ and $k = 20$, then a systematic sample consisting of every 20th unit, will be composed of the units numbered 7, 27, 47, 67 and so on. Thus in systematic sampling, sampling units are selected at uniform interval after a random start.

When the sampling interval corresponds to some periodic or cyclic characteristics in the population, the systematic sampling will result in a non-representative sample.

For example, suppose every 20th shop in a big bazaar is a corner shop and the sampling interval is also 20. if the random start coincides with a corner shop, then the sample will include all the corner shops, and the sample will be highly non representative as different characteristics (say, dealing in beverages) are associated with the corner shops.

when we are providing of this sort of periodicity is to take a fractional sampling interval and then to round off the numbers obtained, e.g. If sampling interval is 20.7 and we start at 7, then the subsequent numbers are 27.7, 48.4, 69.1, 89.8, etc. and the rounding off results in 28, 48, 69, 90, etc. The sample will then consist of the units with serial numbers 7, 28, 48, 69, 90, etc.

The systematic sampling is that it saves much time and effort, it is economical, it is easily selected and conveniently worked out. Furthermore, a systematic sample is more representative of the population sampled as the sample is more evenly distributed across the population.

Topic 16

Non-Probability Sampling Methods

It does not rely on randomization. This technique is more reliant on the researcher's ability to select elements for a sample. Outcome of sampling might be biased and makes difficult for all the elements of population to be part of the sample equally. This type of sampling is also known as non-random sampling.

- Convenience Sampling
- Purposive Sampling
- Quota Sampling
- Referral /Snowball Sampling

Convenience Sampling: Here the samples are selected based on the availability. This method is used when the availability of sample is rare and also costly. So based on the convenience samples are selected.

For example: Researchers prefer this during the initial stages of survey research, as it's quick and easy to deliver results.

Purposive Sampling

This is based on the intention or the purpose of study. Only those elements will be selected from the population which suits the best for the purpose of our study.

For Example: If we want to understand the thought process of the people who are interested in pursuing master's degree then the selection criteria would be "Are you interested for Masters in..?" All the people who respond with a "No" will be excluded from our sample.

Quota Sampling

This type of sampling depends of some pre-set standard. It selects the representative sample from the population. Proportion of characteristics/ trait in sample should be same as population. Elements are selected until exact proportions of certain types of data is obtained or sufficient data in different categories is collected.

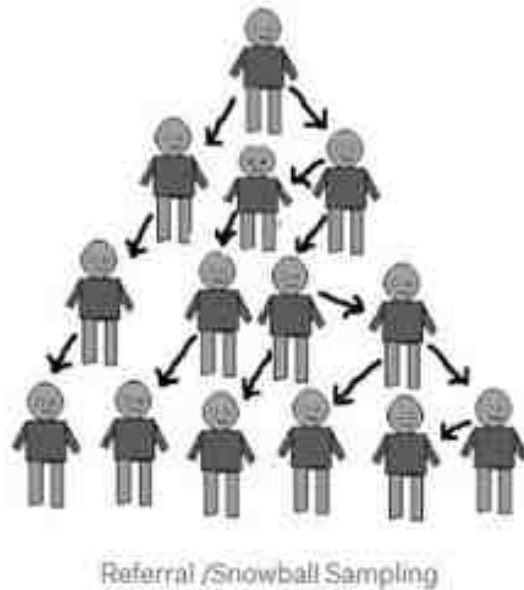
For example: If our population has 45% females and 55% males then our sample should reflect the same percentage of males and females.

Referral /Snowball Sampling

This technique is used in the situations where the population is completely unknown and rare.

Therefore we will take the help from the first element which we select for the population and ask him to recommend other elements who will fit the description of the sample needed.

So this referral technique goes on, increasing the size of population like a snowball.



For example: It's used in situations of highly sensitive topics like HIV Aids where people will not openly discuss and participate in surveys to share information about HIV Aids.

Not all the victims will respond to the questions asked so researchers can contact people they know or volunteers to get in touch with the victims and collect information

Topic 17

Convenience Sampling

Convenience sampling is defined as a method adopted by researchers where they collect market research data from a conveniently available pool of respondents. It is the most used sampling technique as it's incredibly prompt, uncomplicated, and economical. In many cases, members are readily approachable to be a part of the sample.

Researchers use various sampling techniques in situations where there are large populations. In most cases, testing the entire community is practically impossible because they are not easy to reach. Researchers use convenience sampling in situations where additional inputs are not necessary for the principal research. There are no criteria required to be a part of this sample. Thus, it becomes incredibly simplified to include elements in this sample. All components of the population are eligible and dependent on the researcher's proximity to get involved in the sample.

The researcher chooses members merely based on proximity and doesn't consider whether they represent the entire population or not. Using this technique, they can observe habits, opinions, and viewpoints in the easiest possible manner.

A good example of convenience sampling is: A new NGO wants to establish itself in 20 cities. It selects the top 20 cities to serve based on the proximity to where they're based.

Top six advantages of using convenience sampling:

Here are the advantages of adopting a convenience sampling approach:

- I. **Collect data quickly:** In situations where time is a constraint, many researchers choose this method for quick data collection. The rules to gather elements for the sample are least complicated in comparison to techniques such as simple random sampling, stratified sampling, and systematic sampling. Due to this simplicity, data collection takes minimal time.

- II. **Inexpensive to create samples:** The money and time invested in other probability sampling methods are quite large compared to convenience sampling. It allows researchers to generate more samples with less or no investment and in a brief period.
- III. **Easy to do research:** The name of this surveying technique clarifies how samples are formed. Elements are easily accessible by the researchers and so, collecting members for the sample becomes easy.
- IV. **Low cost:** Low cost is one of the main reasons why researchers adopt this technique. When on a small budget, researchers – especially students, can use the budget in other areas of the project.
- V. **Readily available sample:** Data collection is easy and accessible. Most convenience sampling considers the population at hand. Samples are readily available to the researcher. They do not have to move around too much for data collection. Quotas are met quickly, and the data collection can commence even within a few hours.
- VI. **Fewer rules to follow:** It doesn't require going through a checklist to filter members of an audience. Here, gathering critical information and data becomes uncomplicated. For instance, if an NGO wants to survey women's empowerment, they can go to schools, colleges, offices, etc. in their proximity and gather quick responses.



Disadvantages of Convenience Sampling:

1. Highly vulnerable to selection bias and influences beyond the control of the researcher
2. High level of sampling error
3. Studies that use convenience sampling have little credibility due to reasons above

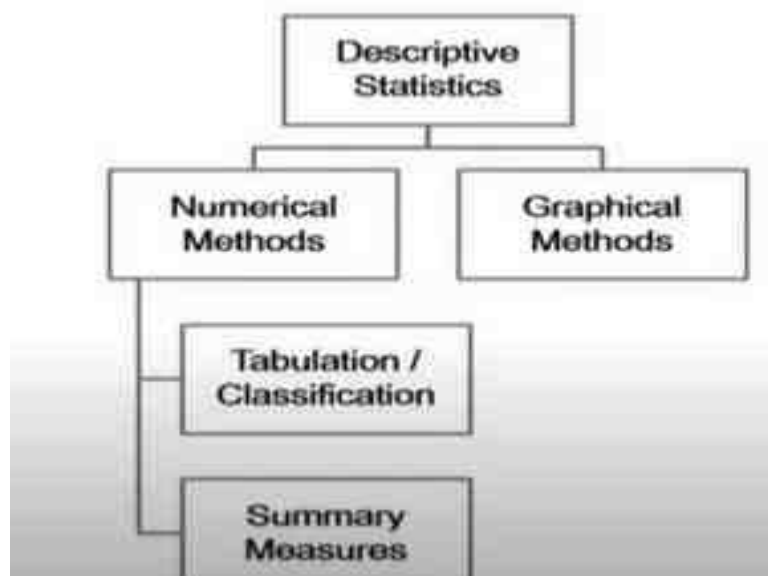
Topic 18

Describing Data with Tables - I

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

Descriptive Statistics is performed using **Tabular**, **Graphical**, and **Numerical** methods.



Tabular Methods are used to summarize the data in table form. It is a systematic organization of information in grid row and columnar structure. The most frequently used tabular format for data summarization is Frequency table and Cross-tabulation.

Graphical Methods are a visual way of presenting data using charts and graphs. The visuals make the data intuitive and self-understandable. The most frequently used visual representation of data are Bar Plot, Histogram, Pareto Chart, Box Plot, Pie Chart, Line Plot, and Scatter Plot.

Arrange data into an array.

The first step in organizing the data is to arrange them in an array so that we can observe the data in a more meaningful and systematic manner. Notice that data can be arranged from lowest to highest values (ascending order) or from highest to lowest values (descending order).

Examples:

Input : array = {5, 9, 1, 4, 6, 8};

Output: 9, 8, 6, 5, 4, 1 or

Output: 1, 4, 5, 6, 8, 9

Frequency Distribution:

The basic function of statistics is to organize and summarize data. Data collected in any research project are in raw and unorganized form. Not much meaning can be conveyed by raw data unless the data are arranged or grouped in a certain manner to give more insight into the data.

For example, I want to conduct a study to identify the level of customer loyalty in my company. In other words, I want to find out how many customers love my company, and they will never choose another company. In this way, I can learn more about my customers. For collecting data, I can create an online questionnaire and distribute randomly among my customers. After collecting data, I will have a big mass of data that is meaningless. To describe customer loyalty, draw conclusions, or make inferences about customer loyalty, I must organize the data in some meaningful way. The most convenient method of organizing data is to construct a frequency distribution. According to Bluman (2013), a frequency distribution is the organization of raw data in table form, using classes and frequencies.

Frequency distribution is based on two components

- Class: is a Qualitative and Quantitative category.
- Class frequency: is the number of data values contained in a specified class.

The reason for construction frequency distribution are as follows:

To organize the data in a meaningful, intelligible way.

To enable the reader to determine the nature or shape of the distribution.

To facilitate computational procedures for measures of average and spread

To enable the researcher to draw charts and graphs for the presentation of data.

To enable the reader to make comparisons among different data sets.

Topic 19

Describing Data with Tables – II

Types of frequency distributions that are most often used are the categorical frequency distribution and the grouped and ungrouped frequency distribution.

Categorical frequency distribution: is used for qualitative data {Nominal & Ordinal}

Grouped frequency distribution: is used for quantitative data { Interval & Ratio}

Ungrouped frequency distribution: is used for quantitative data. It shows the frequency of an item in each separate data value rather than groups of data values.

A frequency distribution table is a chart that shows the frequency of each of the items in a data set. Let's consider an example to understand how to make a frequency distribution table using tally marks.

A jar containing beads of different colors- **red, green, blue, black, red, green, blue, yellow, red, red, green, green, green, yellow, red, green, yellow.** To know the exact number of beads of each particular color, we need to classify the beads into categories. An easy way to find the number of beads of each color is to use tally marks. Pick the beads one by one and enter the tally marks in the respective row and column. Then, indicate the frequency for each item in the table.

Categories	Tally Marks	Frequency
Red		5
Green		6
Blue		2
Black		1
Yellow		3

Thus, the table so obtained is called a frequency distribution table.

Relative Frequency:

It shows the proportion of data values that fall into a given class.

In some situation, Relative frequency is more important than the actual number of data values that fall into that class. For example, if we want to compare the age distribution of students in statistics class with the age distribution of students in accounting class, you would use relative frequency distributions. The reason is that since the population of these two classes are different. To convert a frequency into a proportion or relative frequency, we should divide the frequency for each class by the total of the frequencies. The sum of the relative frequencies will always be 1.

A relative frequency is the **ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes.**

To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

$$\text{R.F} = \frac{\text{frequency of the class}}{\text{total}}$$

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
15	9	$\frac{9}{20}$	$\frac{9}{20}$
16	8	$\frac{8}{20}$	$\frac{17}{20}$
17	5	$\frac{5}{20}$	$\frac{22}{20}$
18	3	$\frac{3}{20}$	$\frac{25}{20}$
19	2	$\frac{2}{20}$	$\frac{27}{20}$
20	3	$\frac{3}{20}$	$\frac{30}{20}$

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the table below.

Cumulative relative frequency = sum of previous relative frequencies + current class frequency

Ungrouped frequency distribution:

The ungrouped frequency distribution is a type of frequency distribution that displays the frequency of each individual data value instead of groups of data values. In this type of frequency distribution, we can directly see how often different values occurred in the table.

Steps for construction of Ungrouped frequency distribution:

Step 1: Construct an ordered Array

Step 2: Construct Classes

Step 3: Construct Tally Marks

Step 4: Count the tally marks to construct frequency

Example: There are 20 students in a class. The teacher, Ms. Jolly, asked the students to tell their favorite subject. The results are as follows –

Mathematics, English, Science, Science, Mathematics, Science, English, Art, Mathematics, Mathematics, Science, Art, Art, Science, Mathematics, Art, Mathematics, English, English, Mathematics.

Solution: 20 students have indicated their choices of preferred subjects. Let us represent this data using tally marks. The tally marks are showing the frequency of each subject.

Subject	Tally Marks	Number of Students
Art		4
Mathematics	/	7
Science	/	5
English		4

According to the above frequency distribution, mathematics is the most liked subject.

Topic 20

Grouped frequency distribution

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 (X_0) = 30.1,

Largest Value (X_m) = 44.9,

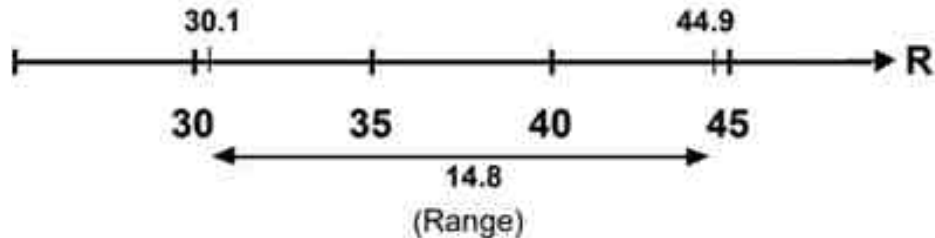
Step-2

Find the range which is defined as the difference between the largest value and the smallest value

In our example:

$$\begin{aligned}\text{Range} &= X_m - X_0 \\ &= 44.9 - 30.1 \\ &= 14.8\end{aligned}$$

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

$$\text{Class interval} = h = 14.8 / 5 = 2.96$$

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.

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	 	14
39.0 – 41.9	 	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	$14/30 = 0.467$	46.7
39.0 – 41.9	8	$8/30 = 0.267$	26.7
42.0 – 44.9	2	$2/30 = 0.067$	6.7
	30		

Topic 21

Contingency tables

Contingency tables (also called crosstabs or two-way tables) are used in statistics to summarize the relationship between several categorical variables. A contingency table is a special type of frequency distribution table, where two variables are shown simultaneously. The entries in the cells of a two-way table can be frequency counts or relative frequencies (just like a one-way table).

Example:

	Dance	Sports	TV	Total
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

The two-way table above shows the favorite leisure activities for 50 adults - 20 men and 30 women. Because entries in the table are frequency counts, the table is a frequency table .

Topic 22

Describing Data with Charts – I

Before we describe data with charts, we will discuss the basic concept of data.

Basic Concepts in Data Analysis

These are the basic concepts in data analysis.

- Data, Information, and Knowledge
- Populations and Samples
- Variables and Observations
- Types of Data: Categorical and Numerical
- Types of Data: Cross Sectional and Time Ordered

Data, Information, and Knowledge

We live in the information age. In the same way that the development of industry created the industrial age, the development of information technology systems, and especially the internet, has created the information age. It has been a long-held belief by many philosophers that knowledge is power and that knowledge stems from understanding of information; information, in turn, is the assigning of meaning to data. To develop learners' understanding of information technology, we start by defining these three related concepts. The topics are hierarchical in that:



What is data?

The concept of data as it is used in the syllabus is commonly referred to as ‘raw’ data – a collection of text, numbers and symbols with no meaning. Data therefore has to be processed, or provided with a context, before it can have meaning.

Example

- 3, 6, 9, 12
- cat, dog, gerbil, rabbit, cockatoo
- 161.2, 175.3, 166.4, 164.7, 169.3

These are meaningless sets of data. They could be the first four answers in the 3 x table, a list of household pets and the heights of 15-year-old students but without a context we don't know.

What is information?

It is important that students learn the concept of what ‘information’ is as used in information technology. Information is the result of processing data, usually by computer. This results in facts, which enables the processed data to be used in context and have meaning. Information is data that has meaning.

Data on its own has no meaning. It only takes on meaning and becomes information when it is interpreted. Data consists of raw facts and figures. When that data is processed into sets according to context, it provides information. Data refers to raw input that when processed or arranged makes meaningful output. Information is usually the processed outcome of data. When data is processed into information, it becomes interpretable and gains significance. In IT, symbols, characters, images, or numbers are data. These are the inputs an IT system needs to process in order to produce a meaningful interpretation. In other words, data in a meaningful form becomes information. Information can be about facts, things, concepts, or anything relevant to the topic concerned. It may provide answers to questions like who, which, when, why, what, and how. If we put Information into an equation it would look like this: $\text{Data} + \text{Meaning} = \text{Information}$

Example

Looking at the examples given for data:

- 3, 6, 9, 12

- cat, dog, gerbil, rabbit, cockatoo

- 161.2, 175.3, 166.4, 164.7, 169.3

Only when we assign a context or meaning does the data become information. It all becomes meaningful when we are told:

- 3, 6, 9 and 12 are the first four answers in the 3 x table

- cat, dog, gerbil, rabbit, cockatoo is a list of household pets

- 161.2, 175.3, 166.4, 164.7, 169.3 are the heights of 15-year-old students

What is knowledge?

When someone memorizes information, this is often referred to as ‘rote-learning’ or ‘learning by heart’. We can then say that they have acquired some knowledge. Another form of knowledge is produced as a result of understanding information that has been given to us, and using that information to gain knowledge of how to solve problems.

How are data, information and knowledge linked?

If we put Knowledge into an equation it would look like this: Information + application or use = Knowledge

Example

Looking at the examples given for data:

- 3, 6, 9, 12

- cat, dog, gerbil, rabbit, cockatoo

- 161.2, 175.3, 166.4, 164.7, 169.3

Only when we assign a context or meaning does the data become information. It all becomes meaningful when we are told:

- 3, 6, 9 and 12 are the first four answers in the 3 x table

- cat, dog, gerbil, rabbit, cockatoo is a list of household pets

- 161.2, 175.3, 166.4, 164.7, 169.3 are the heights of the five tallest 15-year-old students in a class.

If we now apply this information to gain further knowledge we could say that:

- 4, 8, 12 and 16 are the first four answers in the 4 x table (because the 3 x table starts at three and goes up in threes the 4 x table must start at four and go up in fours)

- The tallest student is 175.3cm.

- A lion is not a household pet as it is not in the list and it lives in the wild.

Topic 23

Describing Data with Charts – II

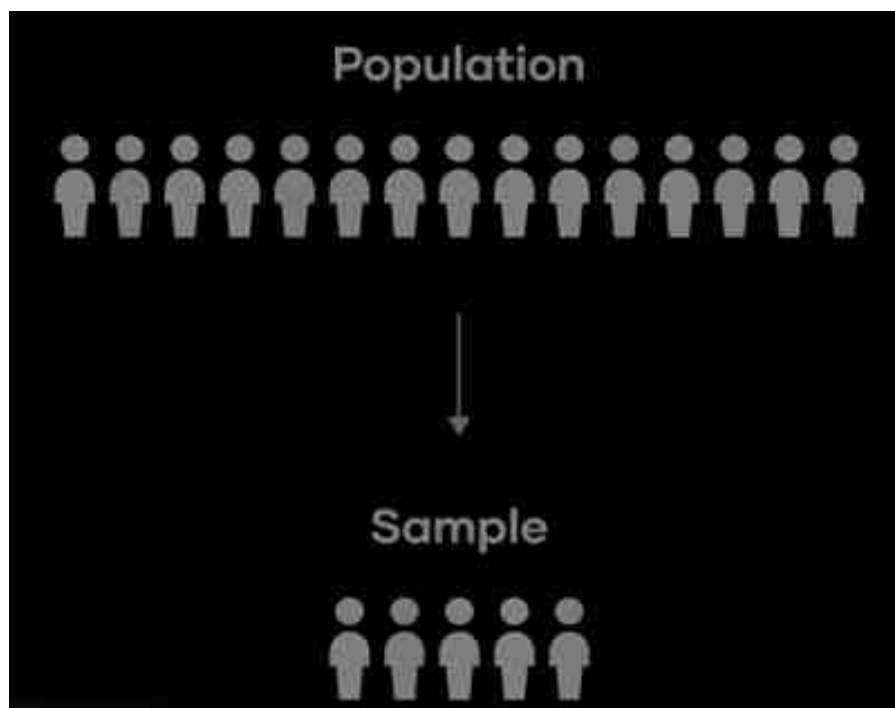
Population and Samples

Population

A population is the entire group that you want to draw conclusions about.

Sample

A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.



Variable and Observations

We call a set of data derived from an object (experimental unit) an observation. Each object is measured according to various aspects, such as temperature, concentration of some constituents, frequency of occurrence of some phenomenon, etc. Each of these aspects is denoted as a variable. By assembling all available data on all objects, we can build a table where the columns represent the variables and the rows represent the measured observations.

As an example, we can take the list of physical characteristics of 10 persons. The objects of the matrix are the persons, the variables are the measured properties, such as the weight or the colour of the eyes.

	Age [years]	Sex	Weight [kg]	Eye Color	Body Temperature [°C]
individuuum 1	42	female	52.9	brown	36.9
individuuum 2	37	male	87.0	green	36.3
individuuum 3	29	male	82.1	blue	36.4
individuuum 4	61	female	62.5	blue	36.7
individuuum 5	77	female	55.5	gray	36.6
individuuum 6	33	male	95.2	green	36.5
individuuum 7	32	female	81.8	brown	37.0
individuuum 8	45	male	78.9	brown	36.3
individuuum 9	18	male	83.4	green	36.6
individuuum 10	19	male	84.7	gray	36.1

Depending on the nature of the measured item and the measurement process the variables may be further classified into discrete (nominal and ordinal scales) and continuous variables (interval and

ratio scales). See the topic [Scale of Measurement](#) for more. The eye color is a discrete variable, the weight or the body temperature is a continuous variable.

Types of Data: Categorical and Numerical

The below picture explains the categorical and numerical data.

	A	B	C	D	E	F
1	Age	Gender	State	Children	Salary	Opinion
2	Middle Aged	1	Georgia	1	65,000	Strongly Agree
3	Elderly	2	Illinois	2	32,000	Neutral
4	Young	1	Florida	0	46,000	Agree
5	Middle Aged	1	Texas	3	52,000	Agree
6	Young	2	Virginia	0	38,000	Strongly Disagree

We can do arithmetic on numerical data (age and salary). These data are actual measurements. Categorical data is qualitative. Sometimes qualitative data is coded. For example, opinion can be coded 1-5 and arithmetic (calculations) can be performed. Such data is ordinal (has implied order). State is a categorical variable and cannot be used for calculations. Such data are nominal.

Types of Data: Cross Sectional and Time Ordered

Cross-sectional Data

Cross-section data is collected in a single time period and is characterized by individual units - people, companies, countries, etc. Some examples include:

- Student grades at the end of the current semester;
- Household data of the previous year - expenditure on food, unemployment, income, etc.
- Car data - average speed, horsepower, colour, etc.

With cross-sectional data the ordering of the data does not matter.

Time Series Data

Data collected at a number of specific points in time is called time series data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc. Time series data can be observed at many different *frequencies* (*hourly, daily, weekly, monthly, quarterly, annually*, etc.).

Unlike cross-sectional data, the ordering of the data is important in time-series data.

Topic 24

Describing Data with Charts – III

Frequency Histograms

A graphical display of distribution of frequencies is known as frequency histogram.

A **Frequency Table** showing a classification of the AGE of attendees at an event.

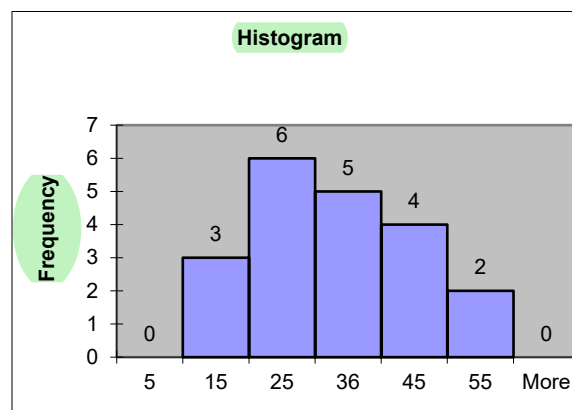
Class	Frequency	Relative Frequency	Percentage
10 but under 20	3	.15	15
20 but under 30	6	.30	30
30 but under 40	5	.25	25
40 but under 50	4	.20	20
50 but under 60	2	.10	10
Total	20	1	100

Class is a range for the values of a variable.

Frequency is the number of observations associated with a class.

Relative Frequency is the proportion of observations (frequency) associated with a class.

This is a graphical display of distribution of frequency table.



Bar and Pie Charts

Frequency tables, pie charts, and bar charts can be used to display the distribution of a single categorical variable. These displays show all possible values of the variable along with either the **frequency** (count) or **relative frequency** (percentage).

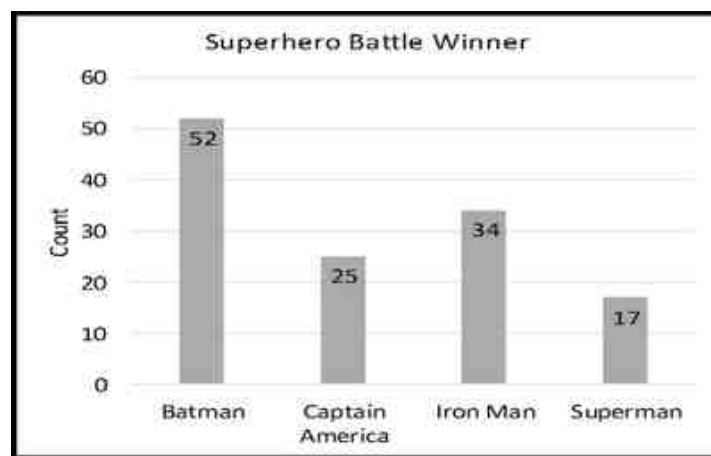
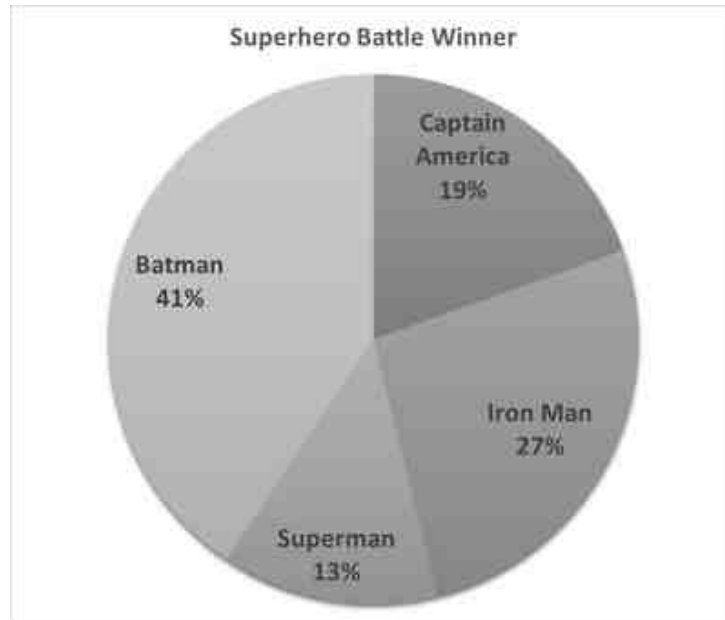
Relative frequencies are more commonly used because they allow you to compare how often values occur relative to the overall sample size. They are calculated by dividing the number of responses for a specific category by the total number of responses. Pie charts represent relative frequencies by displaying how much of the whole pie each category represents. Frequency tables and bar charts can display either the raw frequencies or relative frequencies.

A researcher asked her class to pick who would win in a battle of superheroes. Below is a frequency table and charts of the results:

	Captain America	Iron Man	Superman
Batman	52	34	17

Out of a total of 128 responses, 41% (or 52/128) of students reported that Batman would win the battle, followed by Iron Man with 27%, Captain America with 19%, and Superman with 13%.

A pie chart and bar chart of these results are shown below:

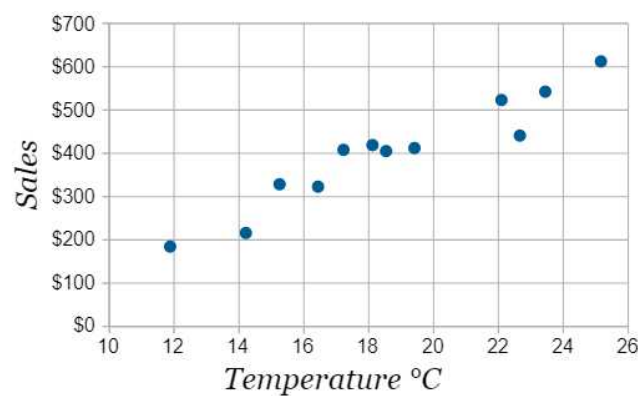


Scatter Plot

Scatter plots (also called *scatter graphs*) are similar to line graphs. A line graph uses a line on an X-Y axis to plot a continuous function, while a scatter plot uses dots to represent individual pieces of data. In statistics, these plots are useful to see if two variables are related to each other. For example, a scatter chart can suggest a linear relationship (i.e. a straight line). The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:

<i>Ice Cream Sales vs Temperature</i>	
Temperature °C	Ice Cream Sales
14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408

And here is the same data as a Scatter Plot:



It is now easy to see that warmer weather leads to more sales, but the relationship is not perfect.

Scatter plots are also called scatter graphs, scatter charts, scatter diagrams and scatter grams.

Time series plot

A time series graph is a line graph of repeated measurements taken over regular time intervals.

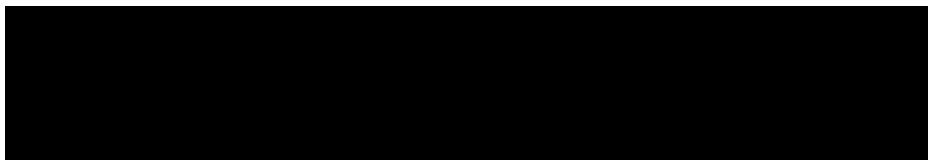
Time is always shown on the horizontal axis.

On time series graphs data points are drawn at regular intervals and the points joined, usually with straight lines.

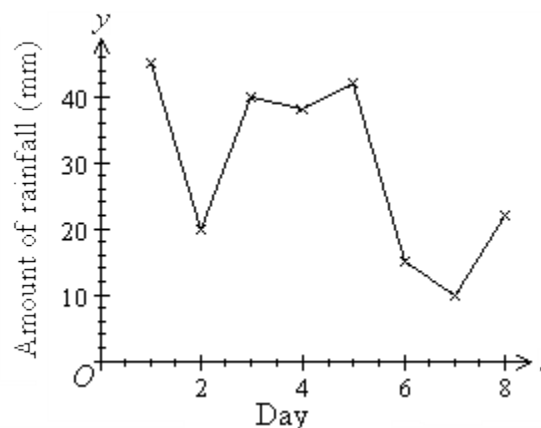
Time series graphs help to show trends or patterns.

Example

Draw a time series graph of the daily amount of rainfall (in millimetres) based on the following recorded data.



The time series graph is obtained by plotting y against t , as shown below.

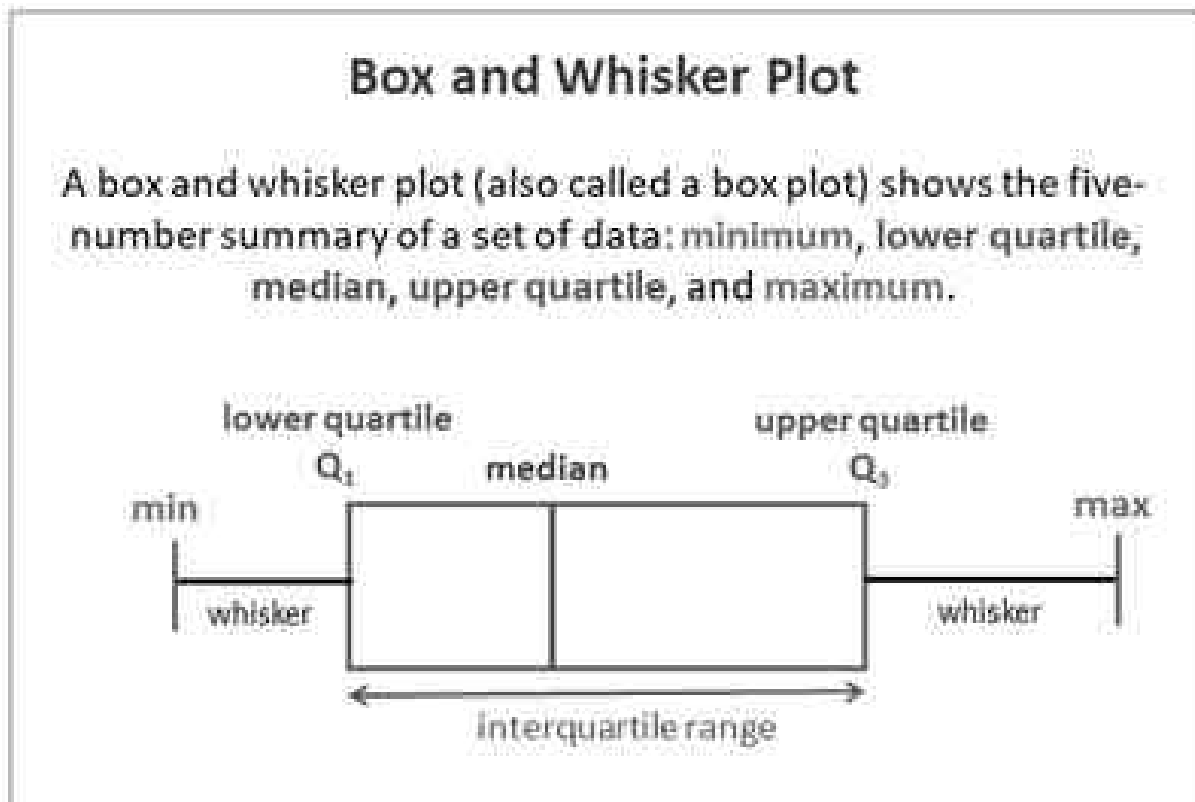


Topic 25

Box and whisker plot

A box plot (also called a box and whisker plot) shows data using the middle value of the data and the quartiles, or 25% divisions of the data.

The following diagram shows a box plot or box and whisker plot.



Drawing A Box And Whisker Plot

Example:

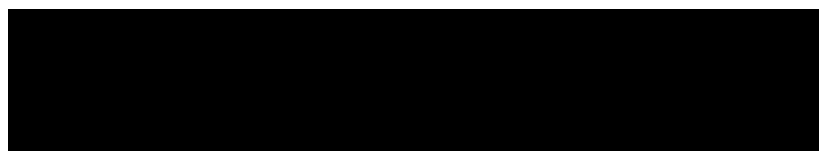
Construct a box plot for the following data:

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution:

Step 1: Arrange the data in ascending order.

Step 2: Find the median, lower quartile and upper quartile.



Median (middle value) = 22

Lower quartile (middle value of the lower half) = 12

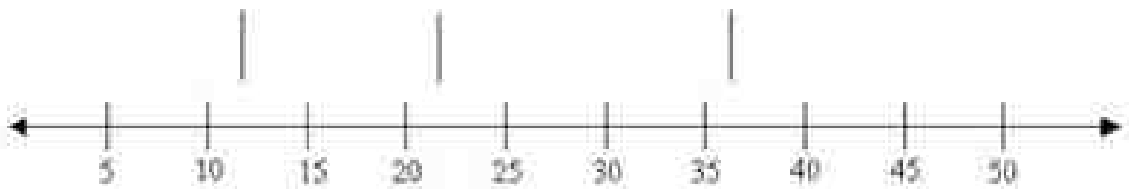
Upper quartile (middle value of the upper half) = 36

(If there is an even number of data items, then we need to get the average of the middle numbers.)

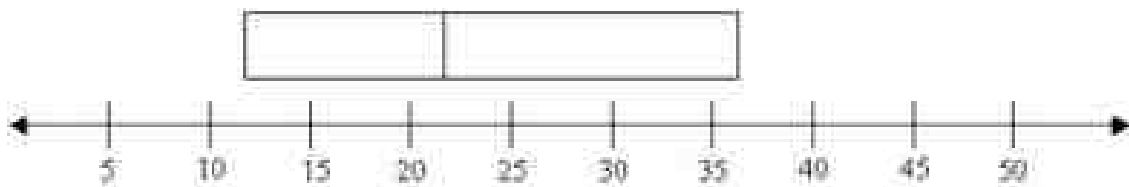
Step 3: Draw a number line that will include the smallest and the largest data.



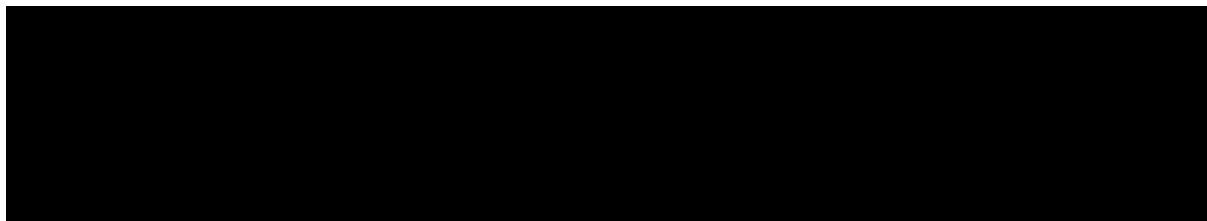
Step 4: Draw three vertical lines at the lower quartile (12), median (22) and the upper quartile (36), just above the number line.



Step 5: Join the lines for the lower quartile and the upper quartile to form a box.



Step 6: Draw a line from the smallest value (5) to the left side of the box and draw a line from the right side of the box to the biggest value (53).



Interquartile range

The interquartile range is the difference between the upper quartile and the lower quartile.

$$53 - 5 = 48$$

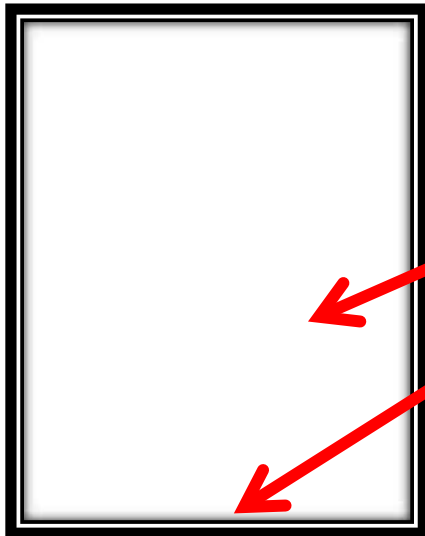
Topic 26

INTRODUCTION TO MS-EXCEL

MS-Excel

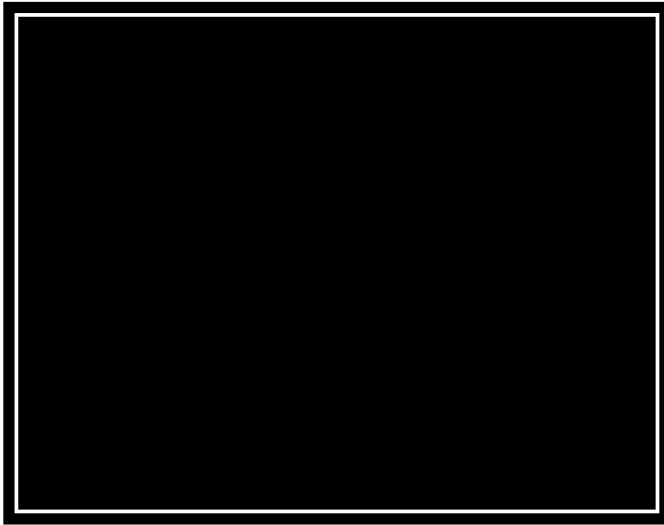
- Excel is a computer program used to create electronic spreadsheets.
- Within excel user can organize data, create chart and perform calculations.
- Excel is a convenient program because it allows user to create large spreadsheets, reference information, and it allows for better storage of information.
- Excel operates like other Microsoft (MS) office programs and has many of the same functions and shortcuts of other MS programs.

Overview of Excel



- Microsoft excel consists of workbooks. Within each workbook, there is an infinite number of worksheets.
- Each worksheet contains Columns and Rows.
- Where a column and a row intersect is called a cell. For e.g. cell D5 is located where column D and row 5 meet.
- The tabs at the bottom of the screen represent different worksheets within a workbook. You can use the scrolling buttons on the left to bring other worksheets into view.

Office Button



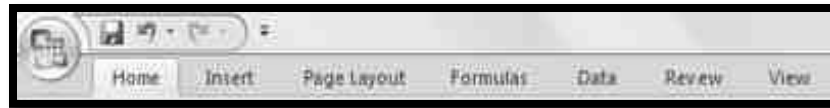
Office Button contains

-  New-to open new workbook. (CTRL+N)
-  Open-to open existing document (CTRL+O)
-  Save-to save a document. (CTRL+S)
-  Save as-to save copy document. (F12)
-  Print-to print a document. (CTRL+P)
-  Prepare-to prepare document for distribution.
-  Send-to send a copy of document to other people.
-  Publish-to distribute document to other people.
-  Close-to close a document (CTRL+W).

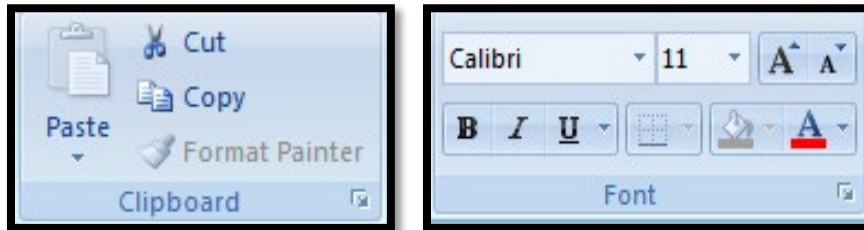
Ribbons

The three parts of the ribbon are

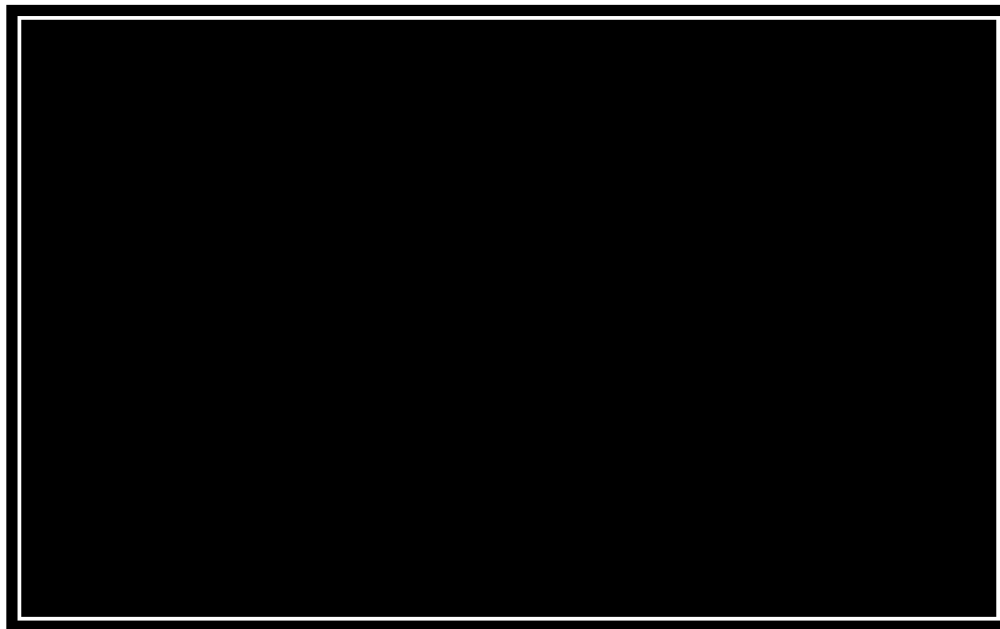
Tabs: There are seven tabs across the top of the excel window.



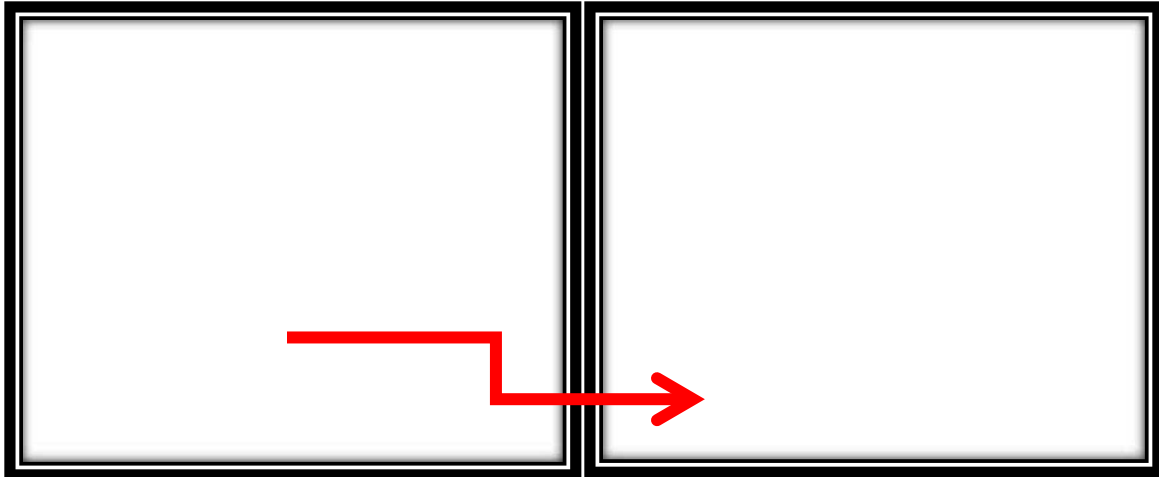
Groups: Groups are sets of related commands, displayed on tabs.



Commands: A command is a button, a menu or a box where you enter information.



Working with cells



To Cut and Paste Cell Contents

Select the cell or cells you wish to cut.

Click the Cut command in the Clipboard group on the Home tab.

Select the cell or cells where you want to paste the information.

Click the Paste command.

The cut information will be removed and now appear in the new cells.

Formatting Text



To format text in bold, italics or underline

Left-click a **cell** to select it or drag your cursor over the text in the formula bar to select it.

Click the **Bold, Italics or underline** command.



To change the font style

Select the cell or cells you want to format.

Left-click the **drop-down arrow** next to the **Font Style** box on the Home tab.

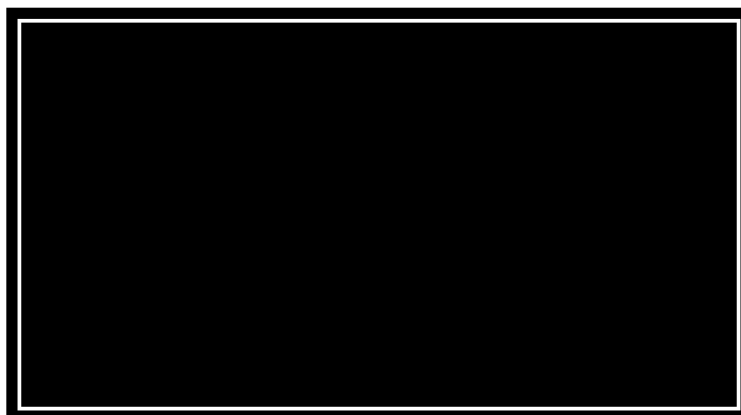
Select a **font style** from the list.

To change the font size

Select the cell or cells you want to format.

Left-click the **drop-down arrow** next to the **Font Size** box on the Home tab.

Select a **font size** from the list.



To add a border

Select the cell or cells you want to format.

Click the **drop-down arrow** next to the **Borders** command on the Home tab. A menu will appear with border options.



To change the text color

Select the cell or cells you want to format.

Left-click the **drop-down arrow** next to the **Text Color** command. A color palette will appear.

Select a color from the palette.



To add a fill colour:

Select the cell or cells you want to format.

Click the **Fill command**. A color palette will appear.

Select a color from the palette.



Conditional formatting

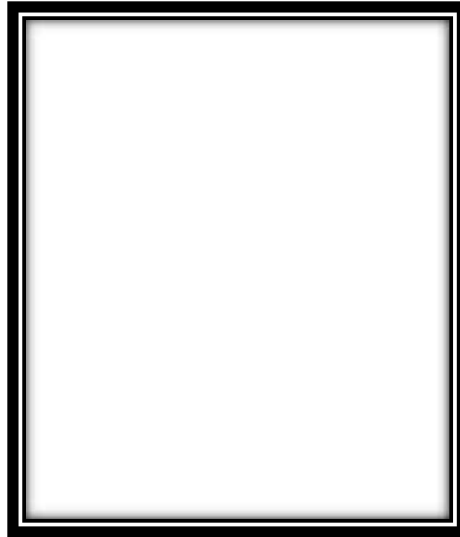
To apply conditional formatting

Select the cells you would like to format.

Select the Home tab.

Locate the Styles group.

Click the Conditional Formatting command. A menu will appear with your formatting options.

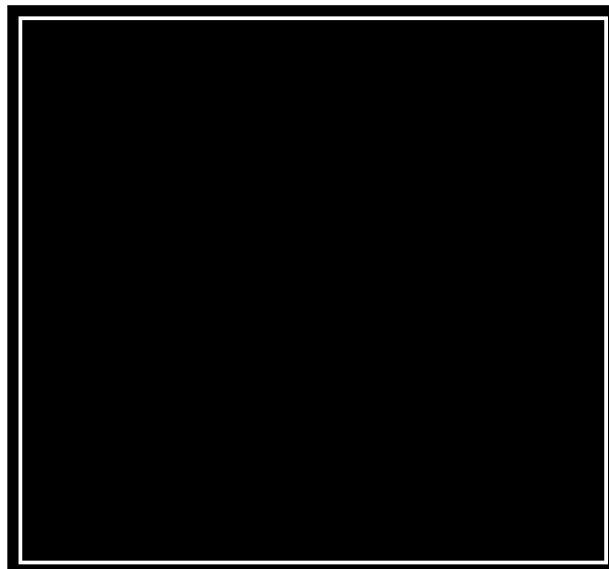


To remove conditional formatting

Click the Conditional Formatting command.

Select Clear Rules.

Choose to clear rules from the entire worksheet or the selected cells.

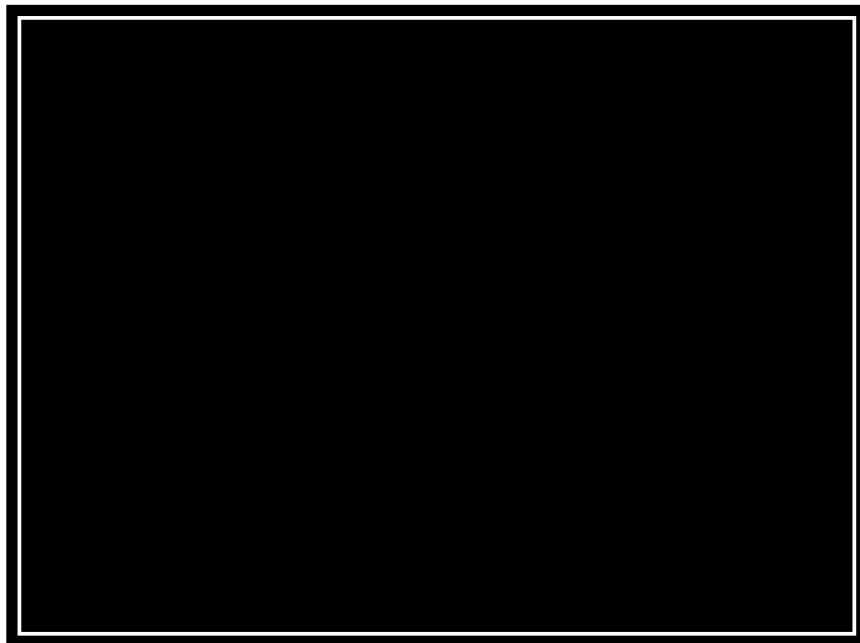


To apply new formatting



To apply new formatting:

Click the Conditional Formatting command. Select New Rules from the menu. There are different rules, you can apply these rules to differentiate particular cell.

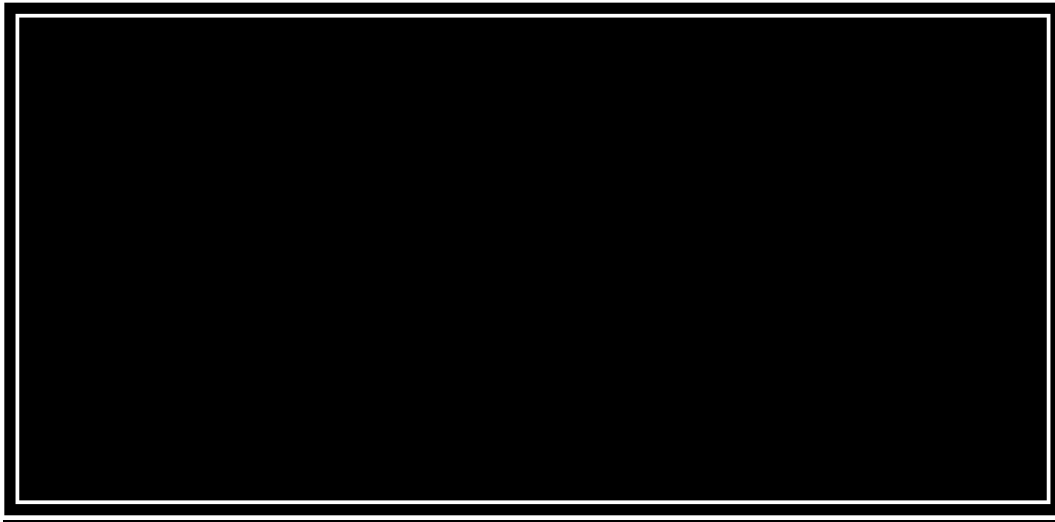


To manage conditional formatting

Click the **Conditional Formatting** command.

Select **Manage Rules** from the menu. The Conditional Formatting Rules Manager dialog box will appear. From here you can edit a rule, delete a rule, or change the order of rules.

To insert rows & columns



NOTE:

1. The new row always appears above the selected row.
2. The new column always appears to the left of the selected column.

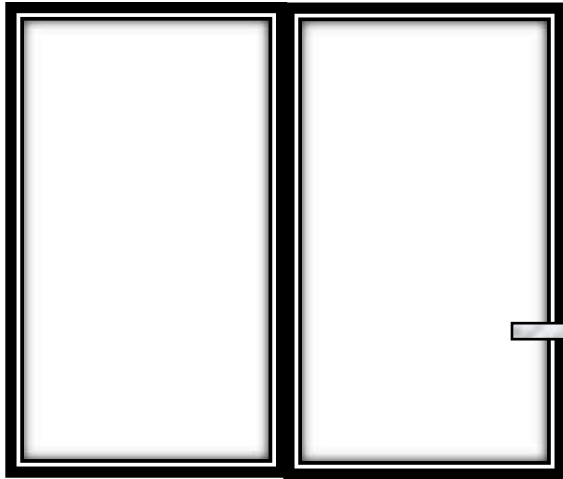
To insert rows:

1. Select the row below where you want the new row to appear.
2. Click the Insert command in the Cells group on the Home tab. The row will appear.

To Insert Columns:

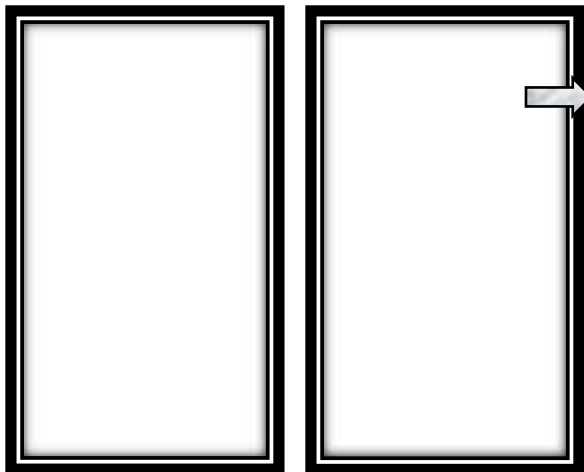
1. Select the column to the right of where you want the column to appear.
2. Click the Insert command in the Cells group on the Home tab. The column will appear.

Editing- fill



- ❑ In the lower right hand corner of the active cell is excel's "fill handle".when you hold your mouse over the top of it, your cursor will turn to a crosshair.

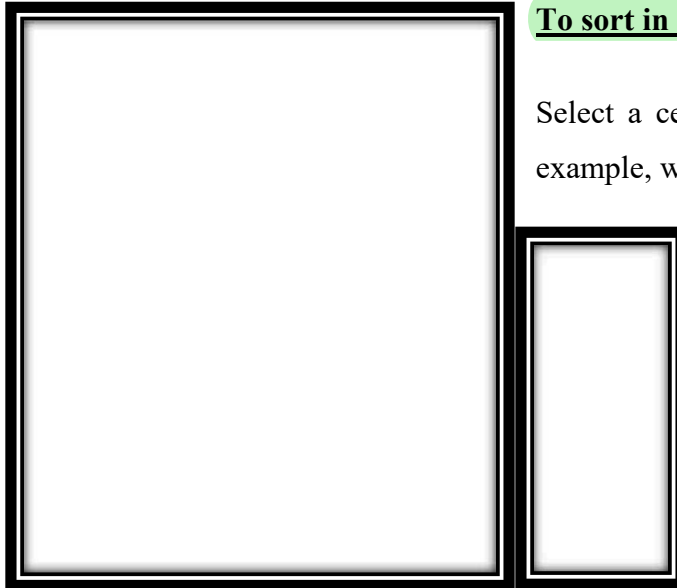
- ❑ If you have just one cell selected, if you click and drag to fill down a column or across a row, it will copy that number or text to each of the other cells.



- ❑ If you have two cells selected, excel will fill in a series. It will complete the pattern. For example, if you put 4 and 8 in two cells select them, click and drag the fill handle, excel will continue the pattern with 12,16,20.etc.

- ❑ Excel can also auto- fill series of dates, times, days of the week, months.

Sorting

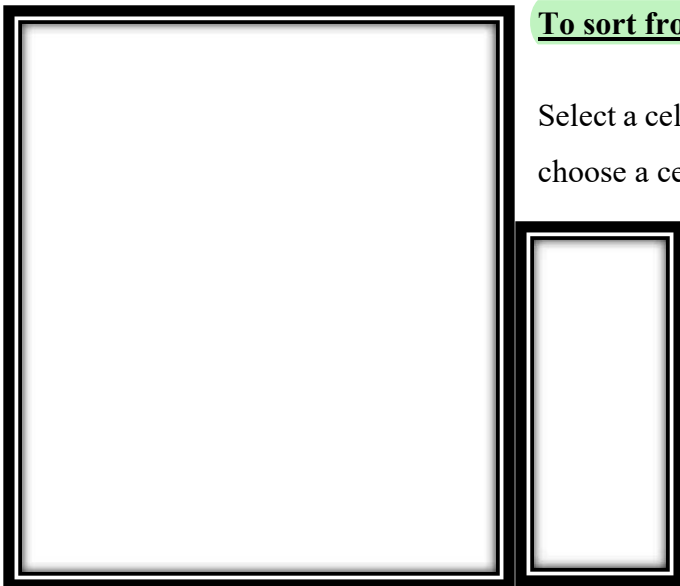


To sort in alphabetical order

Select a cell in the column you want to sort (In this example, we choose a cell in column Q).

Click the **Sort & Filter** command in the **Editing** group on the Home tab.

Select **Sort A to Z**. Now the information in the Category column is organized in alphabetical order.



To sort from smallest to largest

Select a cell in the column you want to sort (In this example, we choose a cell in column Q).

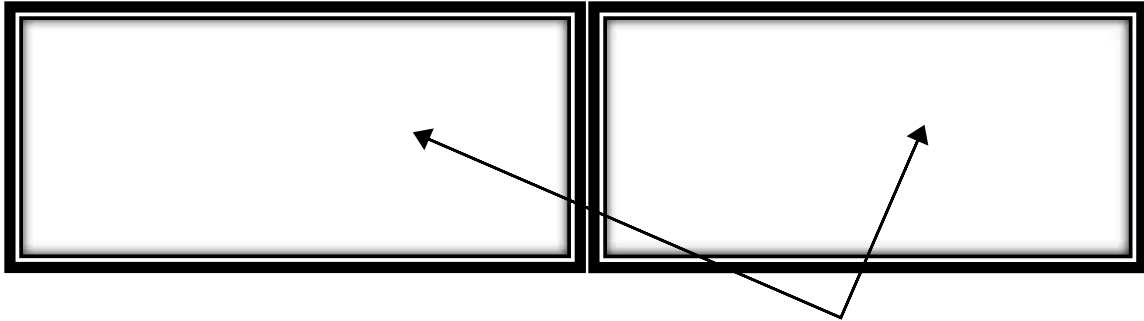
Click the **Sort & Filter** command in the **Editing** group on the Home tab.

Select **From Smallest to Largest**. Now the information is organized from the smallest to largest amount.

Cell referencing

Relative reference

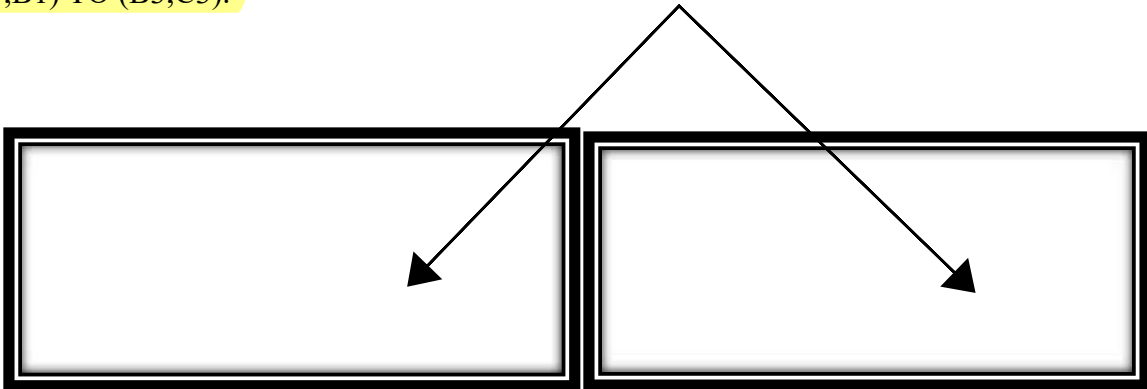
A relative cell reference as (a1) is based on the relative position of the cell. If the position of the cell that contains the reference changes, the reference itself is changed.



In cell (C1) sum function is used.

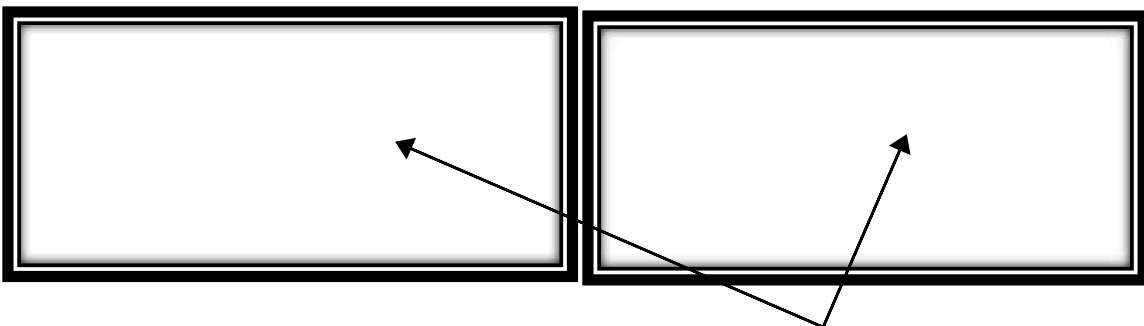
Then function from cell (C1) is copy to cell (D3).

When the position of the cell is changed from (C1) to (D3), then the reference is also changed from (A1,B1) TO (B3,C3).



Absolute reference

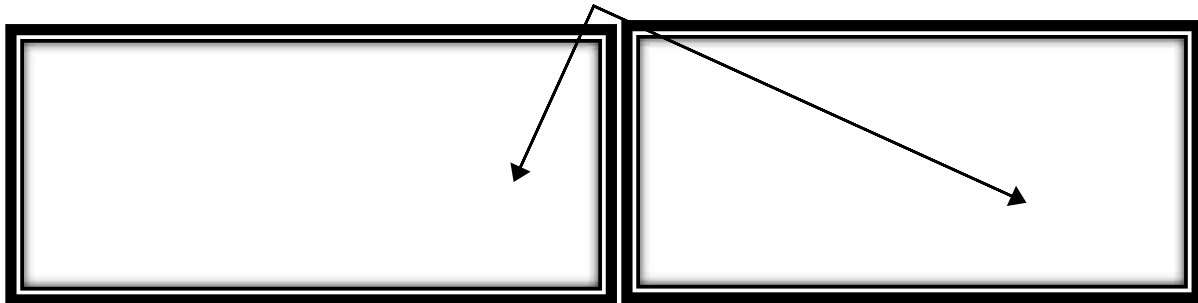
An **absolute cell reference** as (\$a\$1) always refers to a cell in a specific location. If the position of the cell that contains the formula changes, the absolute reference remains the same.



In cell (C1) sum function is used.

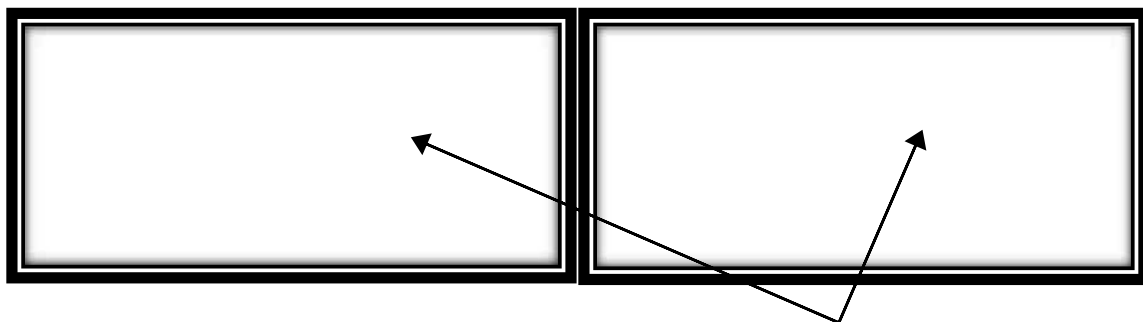
Then function from cell (C1) is copy to cell (D3).

When the position of the cell is changed from (C1) to (D3), then the absolute reference remains the same (A1,B1). **\$ is used for constant row or column.**



Mixed reference

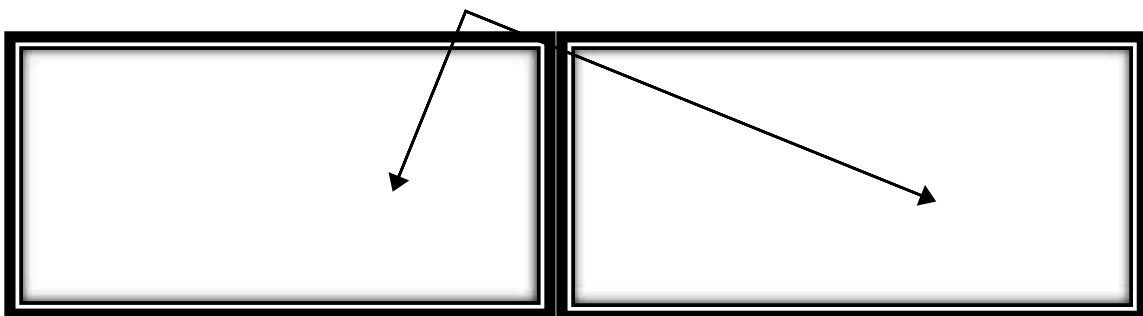
A **mixed reference** has either an absolute column and relative row or absolute row and relative column. An absolute column reference takes the form \$A1, \$B1. An absolute row reference takes the form A\$1, B\$1.



In cell (C1) sum function is used.

Then function from cell (C1) is copy to cell (D3).

When the position of the cell is changed from (C1) to (D3), then row reference is changed (from 1 to 3) but column reference remains same (A, B).



Functions

DATEDIF FUNCTION

SYNTAX OF DATEDIF

=DATEDIF(START_DATE,END_DATE,"INTERVAL")

START DATE-

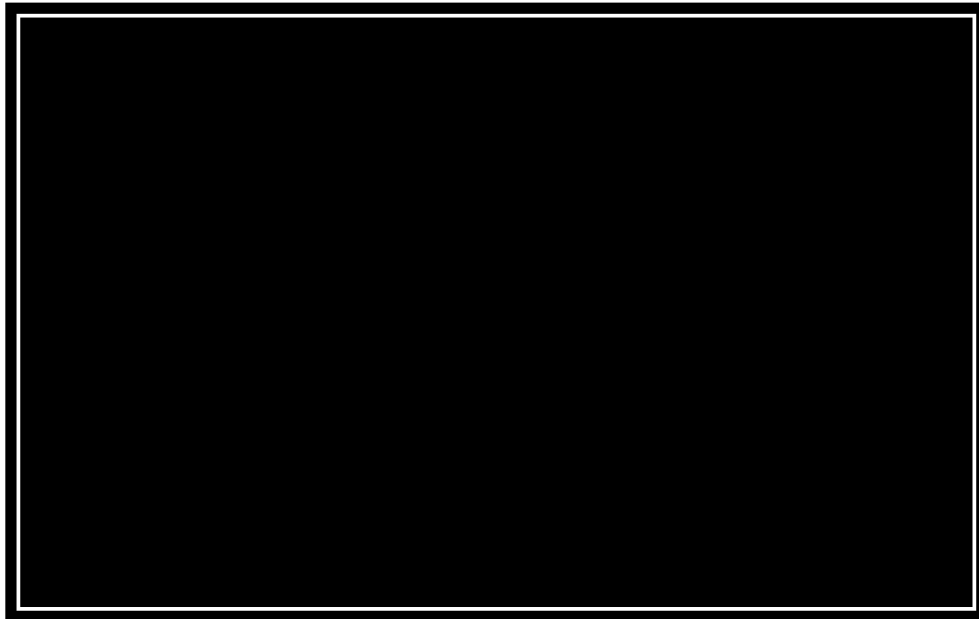
Date from which u want to calculate difference.

END DATE-

Date up to which u want to calculate difference.

INTERVAL-

Form in which u want to calculate difference.



“D”- DAYS

“M”- MONTHS

“Y”- YEARS

“YM”- MONTHS OVER YEAR

This says
that I am
19 years 6
months &

“MD”- DAYS OVER MONTH

SUMIF FUNCTION

SYNTAX OF SUMIF

=SUMIF(RANGE,CRITERIA,SUM_RANGE)

RANGE-

Range of cells on which conditions are applied.

CRITERIA-

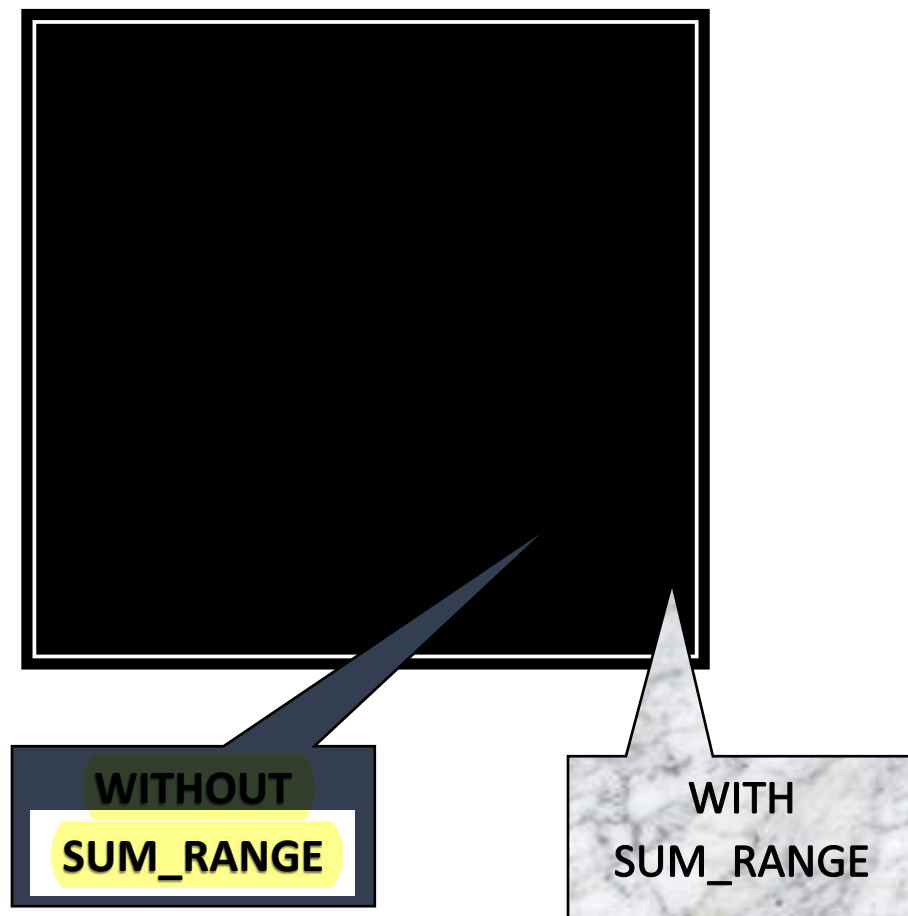
Condition that defines which cell or cells will be added.

SUM_RANGE-

Actual cells to sum.

NOTE:-

If sum range is not used then range is used for sum.



IF FUNCTION

SYNTAX OF IF

=IF(LOGICAL TEXT, VALUE IF TRUE, VALUE IF FALSE)

LOGICAL TEXT-

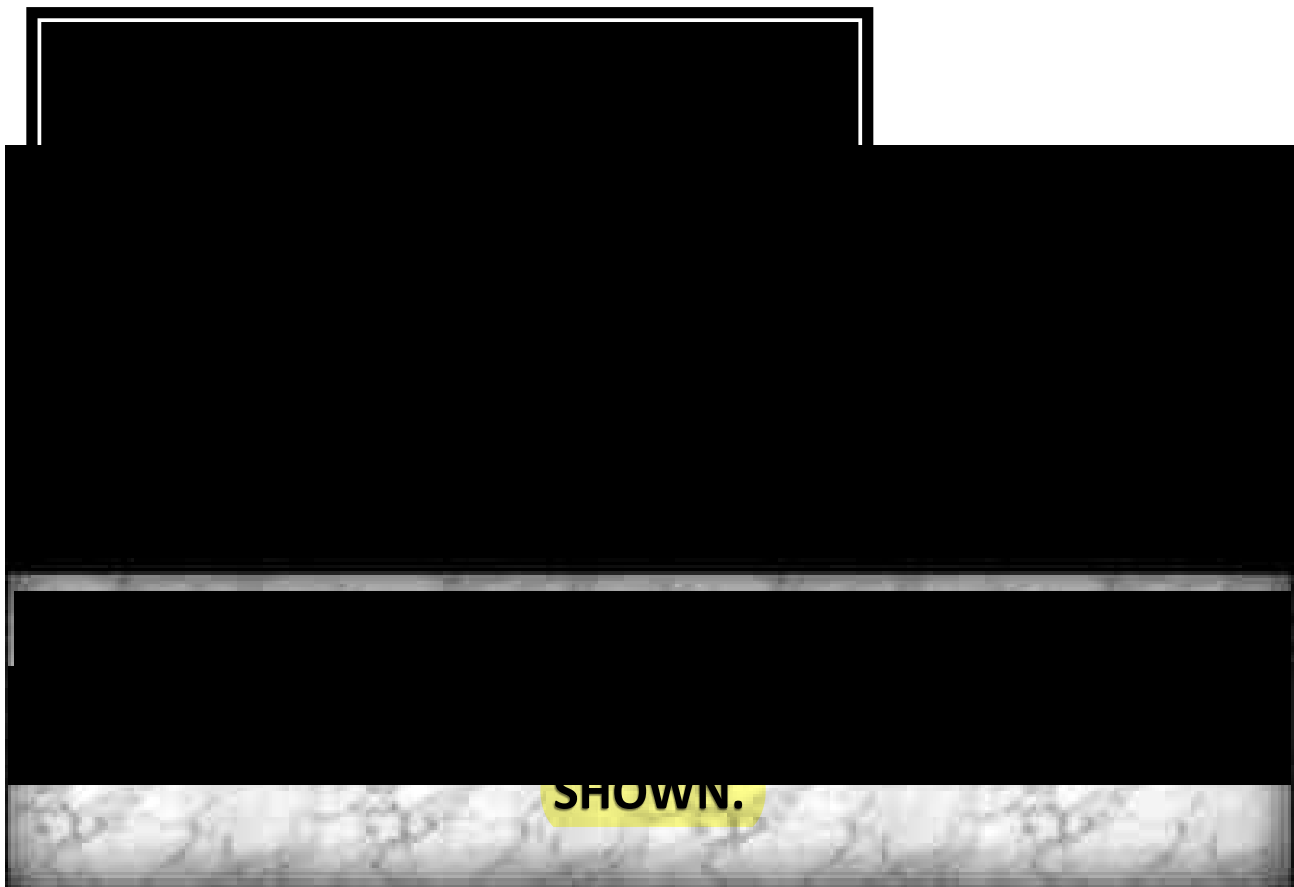
Any value or expression that can be evaluated to TRUE or FALSE.

VALUE IF TRUE-

Value that is returned if logical text is TRUE.

VALUE IF FALSE-

Value that is returned if logical text is FALSE.



Count functions

Syntax of functions

1. COUNT

=COUNT(VALUE1,VALUE2,...)

2. COUNTA

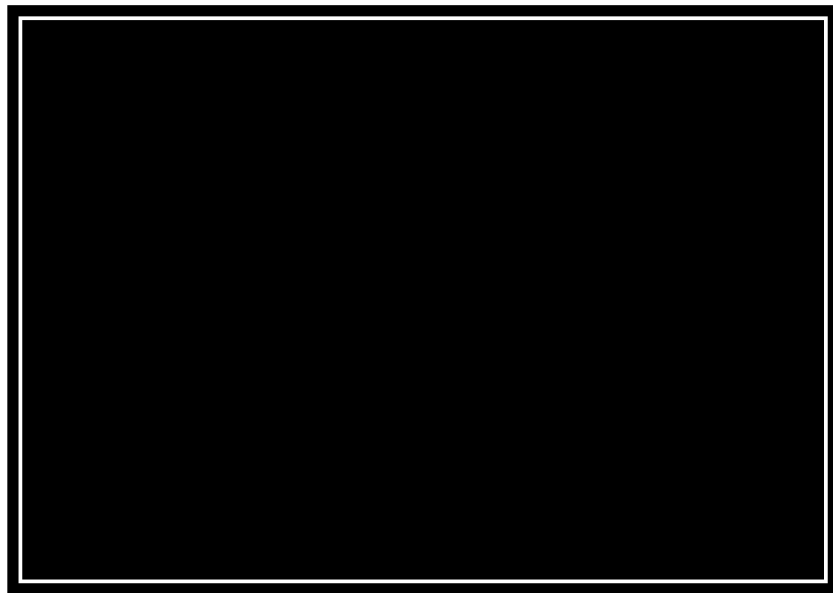
=COUNTA(VALUE1,VALUE2,...)

3. COUNTBLANK

=COUNTBLANK(RANGE)

4. COUNTIF

=COUNTIF(RANGE,CRITERIA)



1.

2.

3.

4.

COUNT ONLY
CELLS THAT
CONTAINS
NUMBER

COUNT CELLS
THAT ARE NOT
EMPTY.

COUNT
CELLS THAT
ARE BLANK.

COUNT NO. OF
CELLS THAT
MEET GIVEN
CONDITION.

Text functions

SYNTAX OF FUNCTIONS

1. LOWER FUNCTION

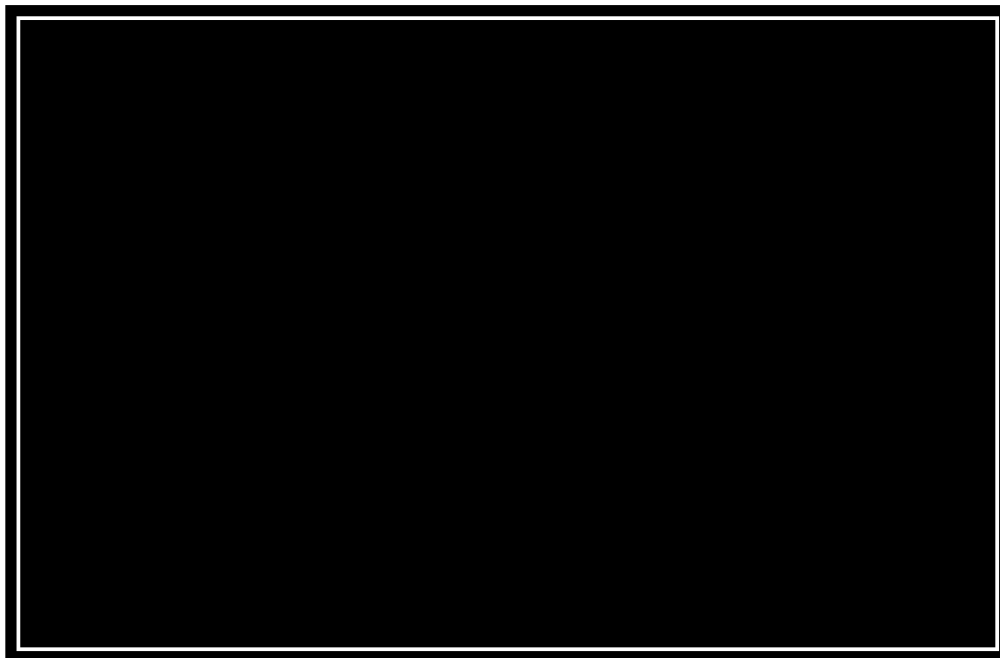
=LOWER(TEXT)

2. UPPER FUNCTION

=UPPER(TEXT)

3. PROPER FUNCTION

=PROPER(TEXT)



1.

TO
CONVER
T TEXT
FROM
CAPITAL
TO
SMALL.

2.

TO
CONVERT
TEXT
FROM
SMALL TO
CAPITAL.

3.

TO
CAPITALIS
ED EACH
WORD OF
TEXT.

SYNTAX OF FUNCTIONS

1. **LEFT FUNCTION** =LEFT(TEXT,NUM_CHARS)

2. **RIGHT FUNCTION** =RIGHT(TEXT,NUM_CHARS)

3. **MID FUNCTION**

=MID(TEXT,STARTNUM,NUM_CHAR)



1.

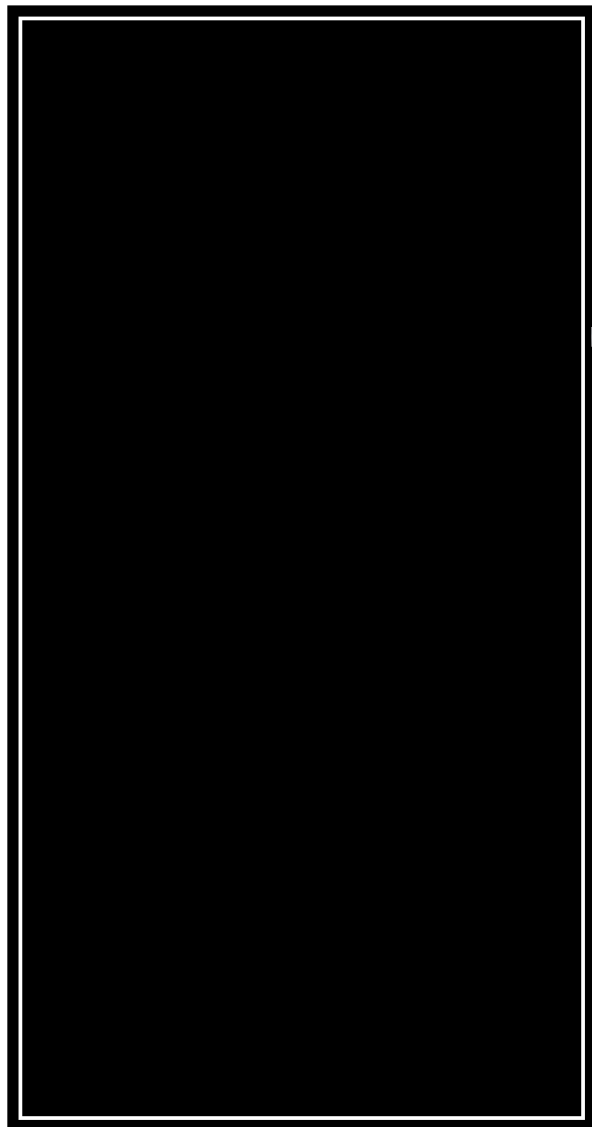
2.

3.

RETURN
SPECIFIED NO.
OF CHARACTER
FROM END OF
TEXT.

RETURN
SPECIFIED NO.
OF
CHARACTER
FROM START
OF TEXT.

RETURN
CHARACTER
FROM
MIDDLE OF
TEXT, GIVEN
A STARTING
POSITION.



USES OF FUNCTIONS

→ **NOW** RETURNS CURRENT DATE
AND TIME.

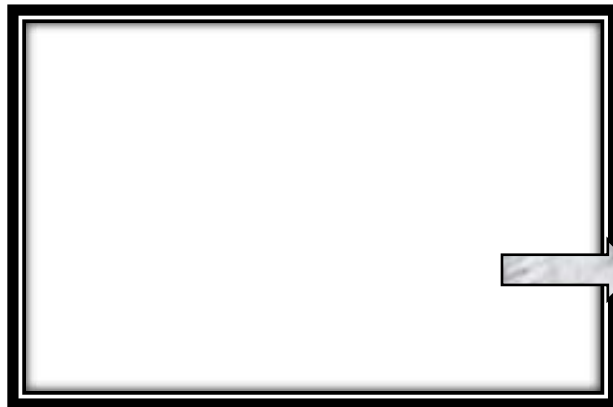
→ **TODAY** RETURNS CURRENT DATE
ONLY.

→ **MOD** RETURNS THE
REMAINDER AFTER A NO.
IS DIVIDED BY A DIVISOR.

→ **LEN** RETURNS THE NO. OF
CHARACTERS IN A TEXT STRING.

SUM ADD ALL THE NUMBERS. →

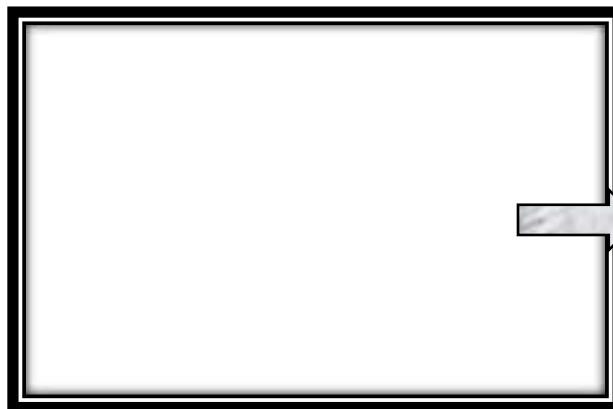
TRACE PRECEDENTS



SHOW ARROW THAT INDICATE WHAT CELLS AFFECT THE VALUE OF THE CURRENTLY SELECTED CELL.

IN THIS EXAMPLE CELLS **A1 & A3** AFFECT THE VALUE OF CELL **C2** & CELLS **A1 & A4** AFFECT THE VALUE OF CELL **C6**.

TRACE DEPENDENTS



SHOW ARROW THAT INDICATE WHAT CELLS ARE AFFECTED BY THE VALUE OF THE CURRENTLY SELECTED CELL.

IN THIS EXAMPLE CELL **C2 & C6** ARE AFFECTED BY THE VALUE OF CELL **A2** & CELL **C6** IS ALSO AFFECTED BY THE

CELL **A4**.

Topic 27

Describing Data from its shapes

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 (X_0) = 30.1,

Largest Value (X_m) = 44.9,

Step-2

Find the range which is defined as the difference between the largest value and the smallest value

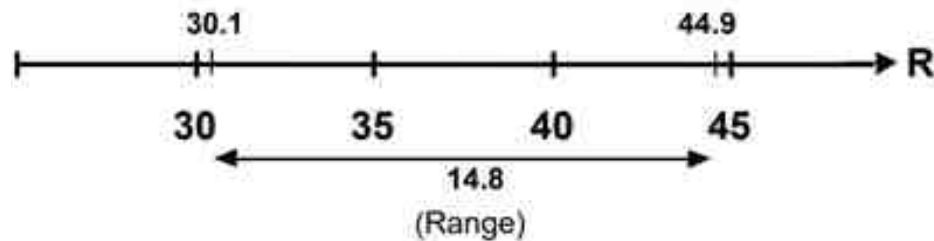
In our example:

$$\text{Range} = X_m - X_0$$

$$= 44.9 - 30.1$$

$$= 14.8$$

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

$$\text{Class interval} = h = 14.8 / 5 = 2.96$$

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.

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	 	14
39.0 – 41.9	 	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	$14/30 = 0.467$	46.7
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	$2/30 \times 100 = 6.7$	$7/50 \times 100 = 14$
33.0-35.9	$4/30 \times 100 = 13.3$	$10/50 \times 100 = 20$
36.0-38.9	$14/30 \times 100 = 46.7$	$16/50 \times 100 = 32$
39.0-41.9	$8/30 \times 100 = 26.7$	$9/50 \times 100 = 18$
42.0-44.9	$2/30 \times 100 = 6.7$	$8/50 \times 100 = 16$

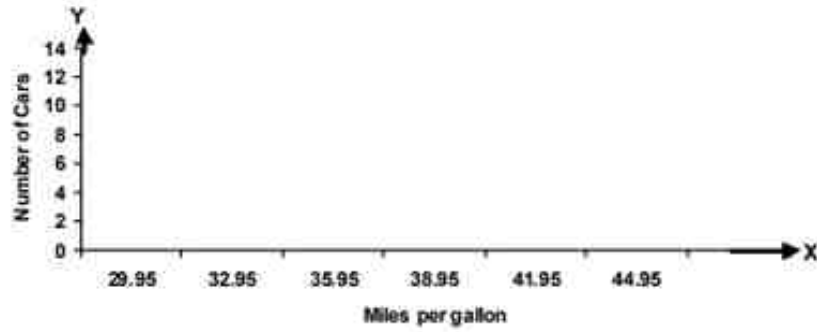
From the table it is clear that whereas 6.7% of the cars of model A fall in the mileage group 42.0 to 44.9, as many as 16% of the cars of model B fall in this group. Other comparisons can similarly be made. Let us now turn to the visual representation of a continuous frequency distribution. In this context,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

HISTOGRAM:

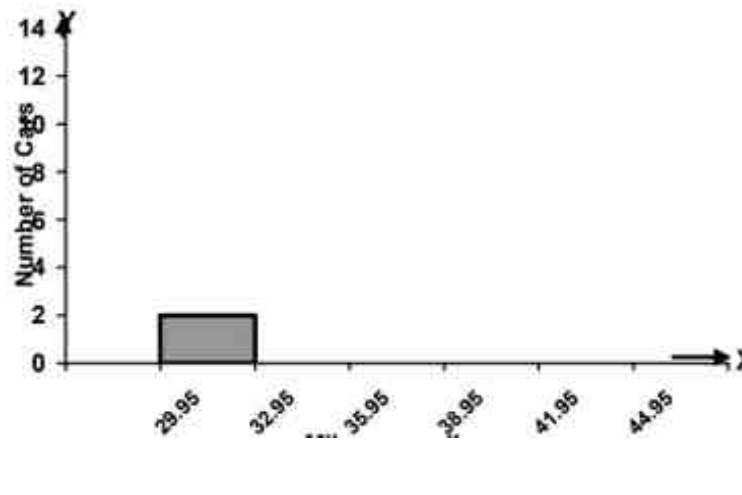
A histogram consists of a set of adjacent rectangles whose bases are marked off by class boundaries along the X-axis, and whose heights are proportional to the frequencies associated with the respective classes. It will be recalled that, in the last lecture, we were considering the mileage ratings of 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

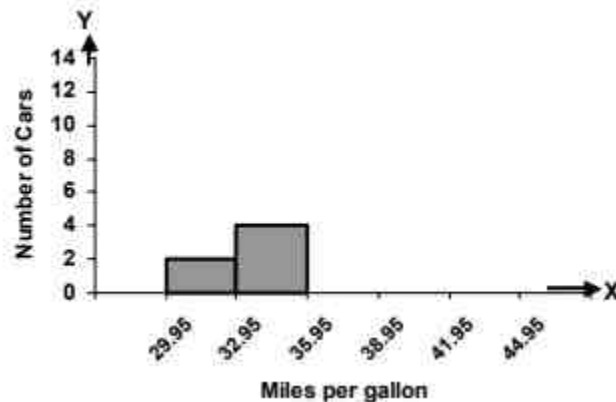


Now, as seen in the frequency table, the frequency of the first class is 2. As such, we will draw a rectangle of height equal to 2 units and obtain the following figure:

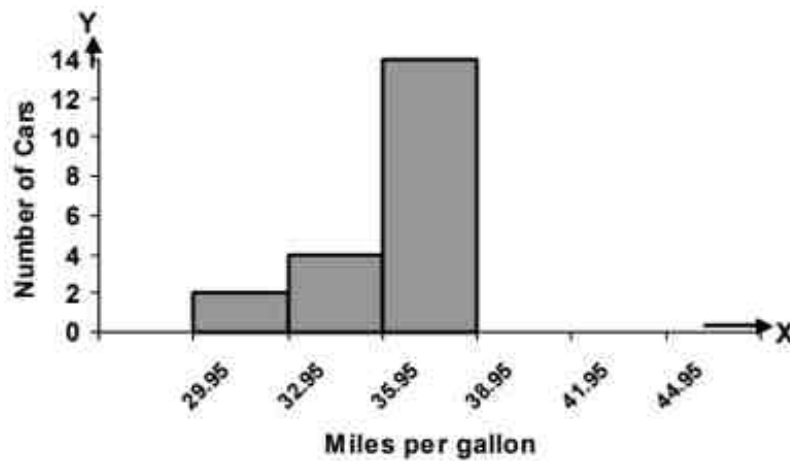


Miles per gallon

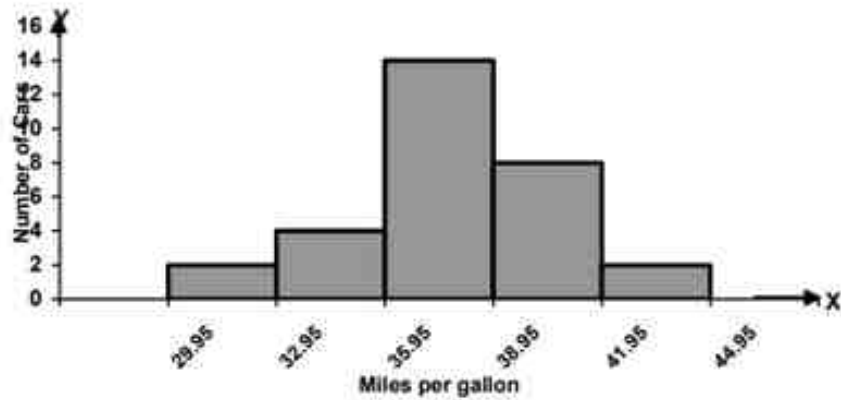
The frequency of the second class is 4. Hence we draw a rectangle of height equal to 4 units against the second class, and thus obtain the following situation:



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

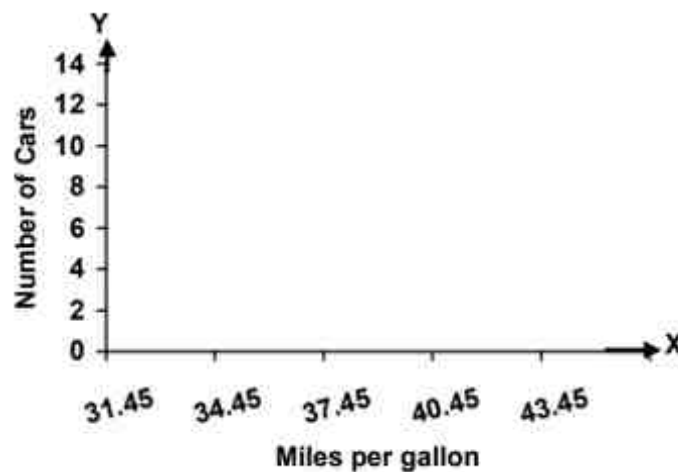
These mid-points are denoted by X. Now let us add two classes to my frequency table, one class in the very beginning, and one class at the very end.

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	

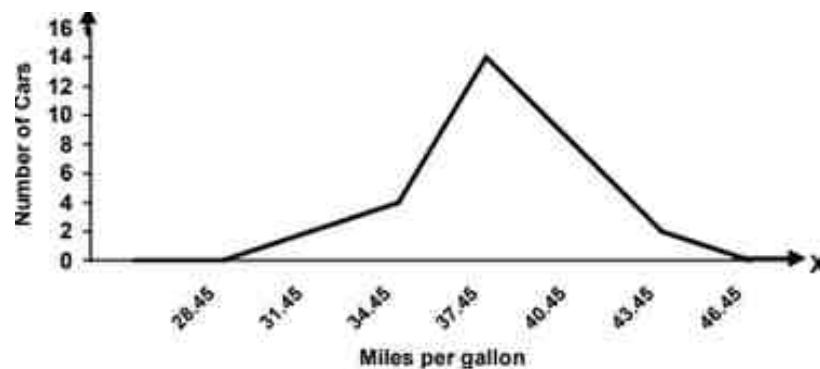
The frequency of each of these two classes is 0, as in our data set, no value falls in these classes.

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:

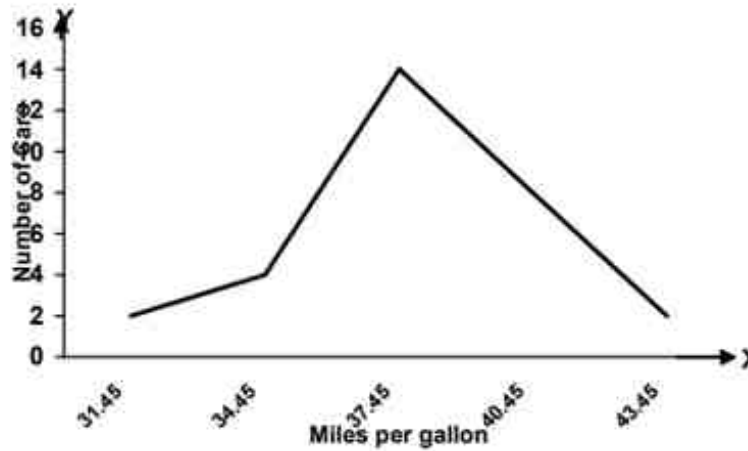


Next, we plot points on our graph paper according to the frequencies of the various classes, and join the points so obtained by straight line segments. In this way, we obtain the following frequency polygon:



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:

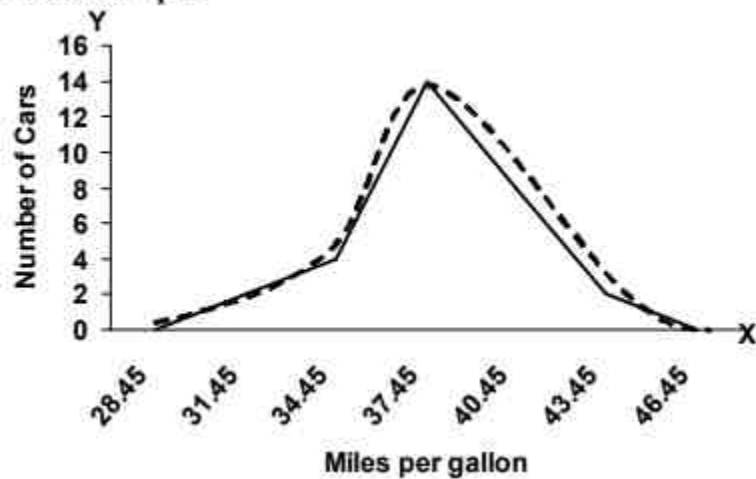


And since this graph is touching the X-axis, hence it cannot be called a frequency polygon (because it is not a closed figure)!

FREQUENCY CURVE:

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

In our example:



Topic 28

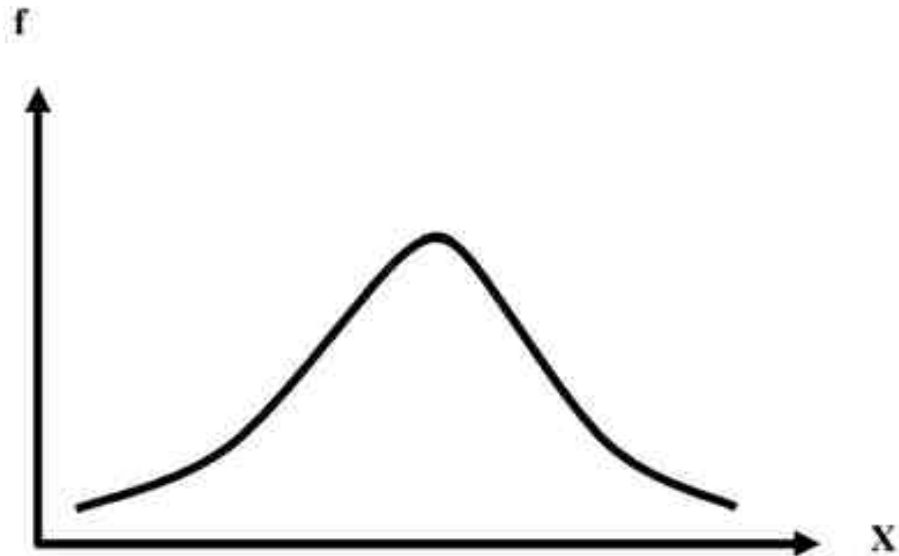
Describing Data: From its peak

VARIOUS TYPES OF FREQUENCY CURVES

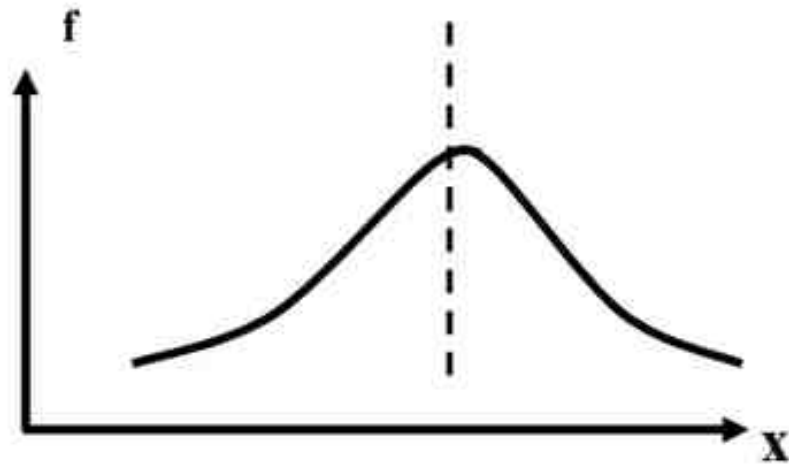
- the symmetrical frequency curve
- the moderately skewed frequency curve
- the extremely skewed frequency curve
- the U-shaped frequency curve

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

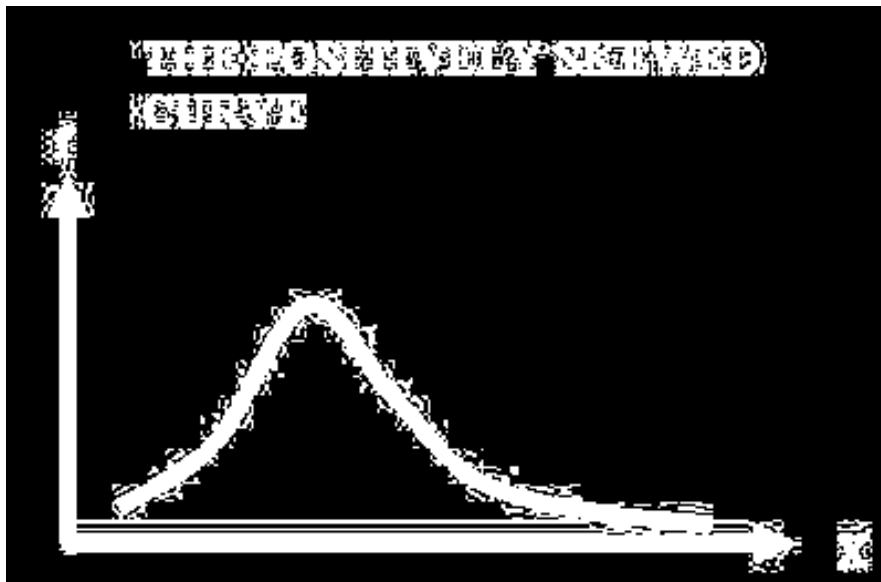
THE SYMMETRIC CURVE



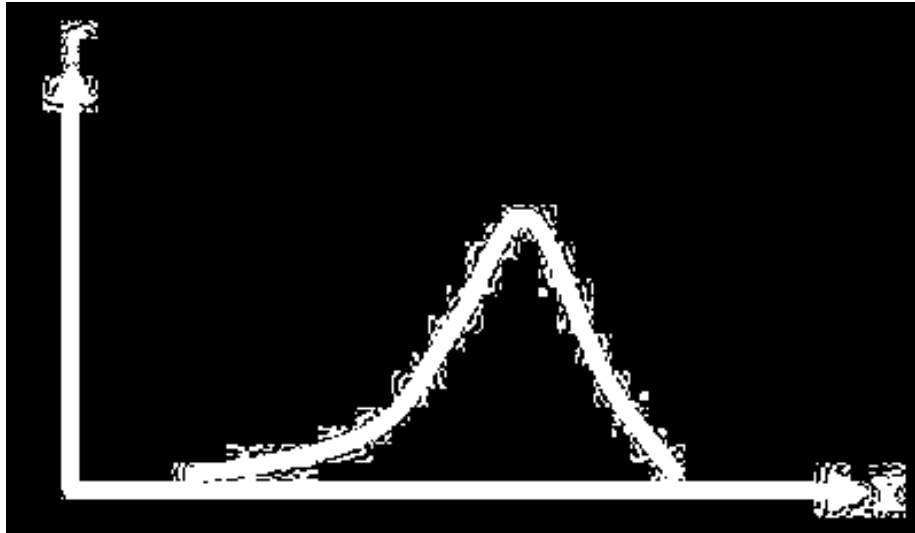
If we place a vertical mirror in the centre of this graph, the left hand side will be the mirror image of the right hand side.



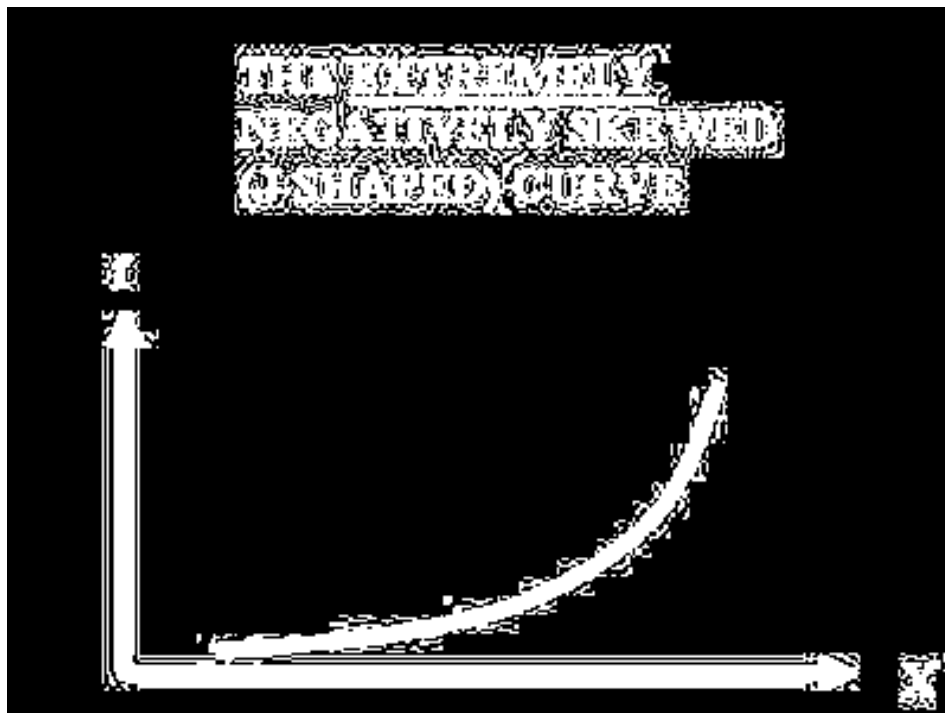
Next, we consider the moderately skewed frequency curve. We have the positively skewed curve and the negatively skewed curve. The positively skewed curve is that one whose right tail is longer than its left tail, as shown below



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:

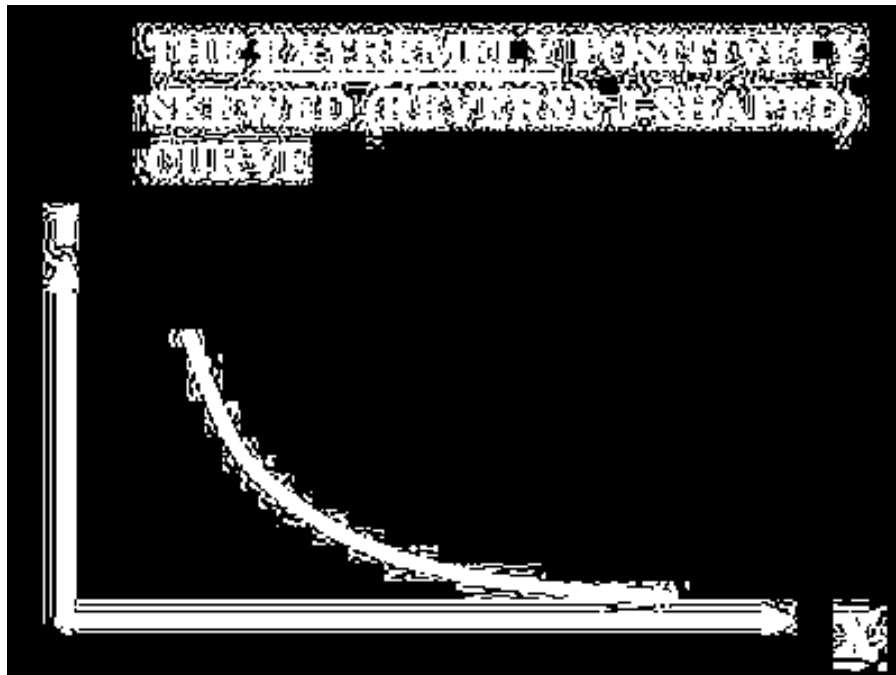


This is the case when the maximum frequency occurs at the end of the frequency table. For example, if we think of the death rates of adult males of various age groups starting from age 20 and going up to age 79 years, we might obtain something like this:

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.



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



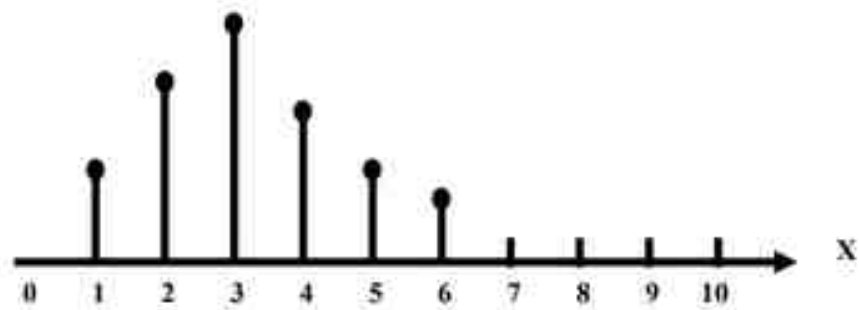
If we consider the example of the death rates not for only the adult population but for the population of ALL the age groups, we will obtain the **U-shaped distribution**.

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.

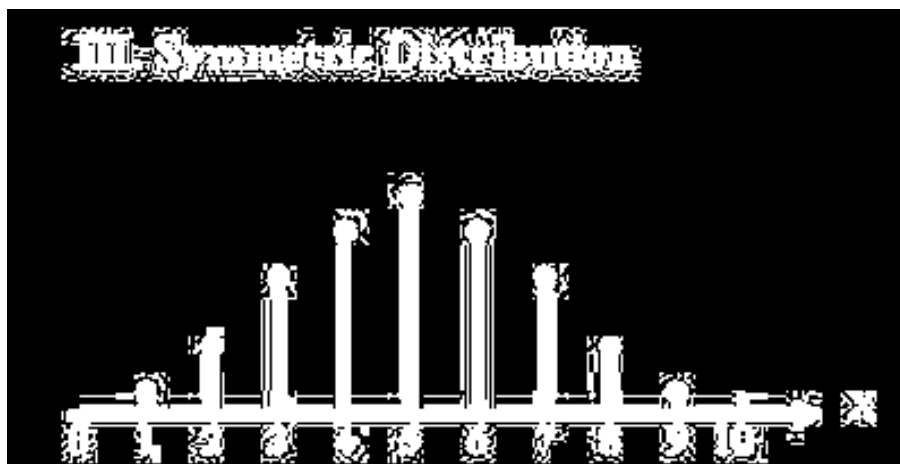
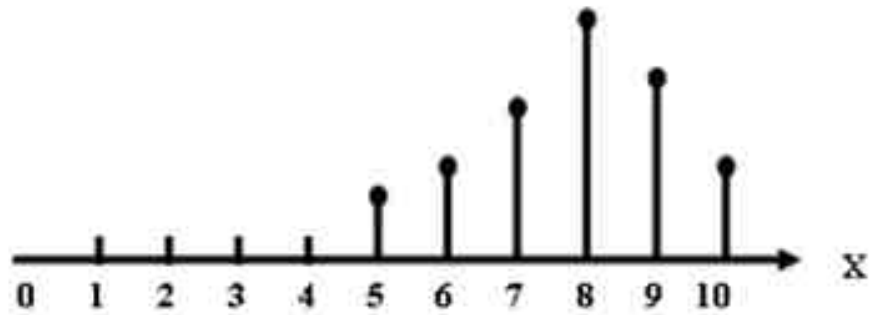
If we construct a frequency distribution of this data, and draw its histogram and its frequency curve, we will find that our data will generate a moderately skewed distribution. Until now, we have discussed the various possible shapes of the frequency distribution of a continuous variable. Similar shapes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

VARIOUS TYPES OF DISCRETE FREQUENCY DISTRIBUTION

I. Positively Skewed Distribution



II. Negatively Skewed Distribution



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.

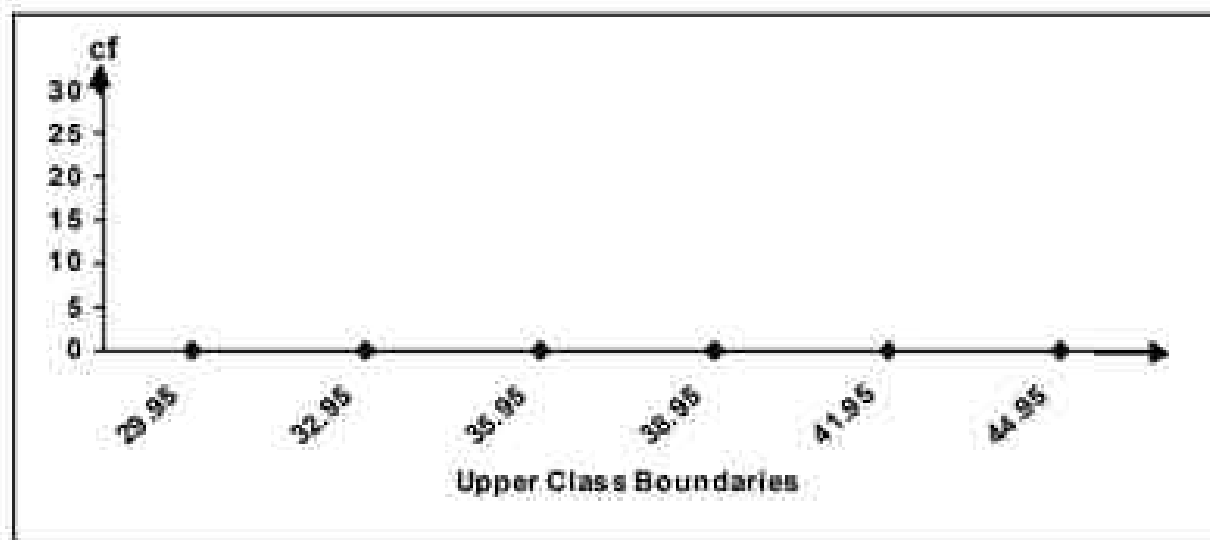
CUMULATIVE FREQUENCY DISTRIBUTION

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.

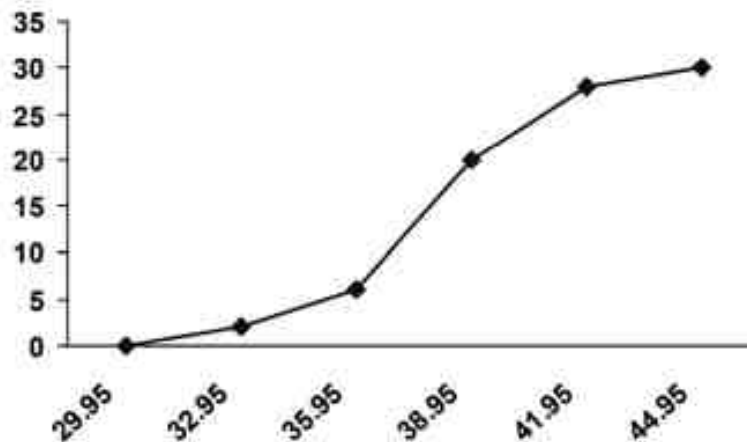
CUMULATIVE FREQUENCY POLYGON or OGIVE

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:

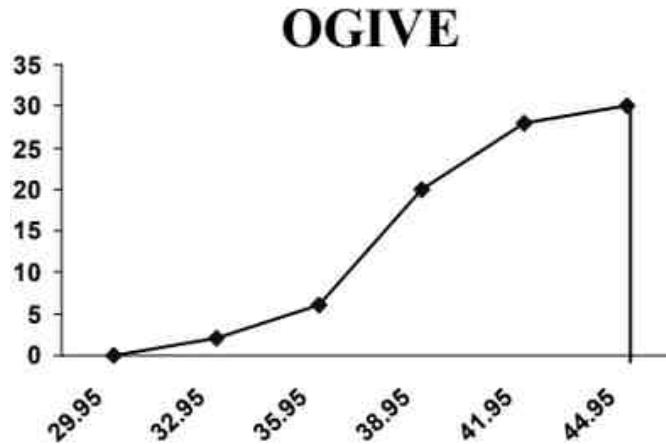
Cumulative Frequency Polygon or OGIVE



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.

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

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

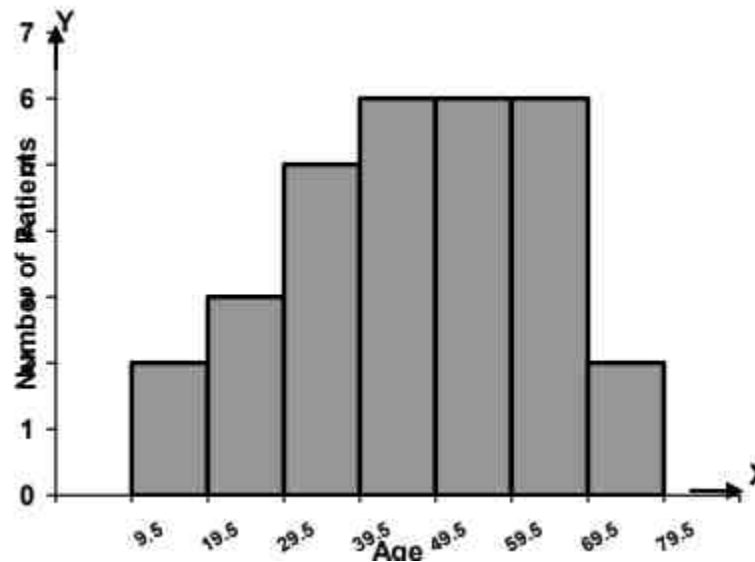
Stem (Leading Digit)	Leaf (Trailing Digit)
1	2 8
2	6 7 9
3	0 1 5 7 9
4	0 2 3 8 8 9
5	1 2 3 4 7 8
6	1 2 4 5 7 8
7	1 4

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

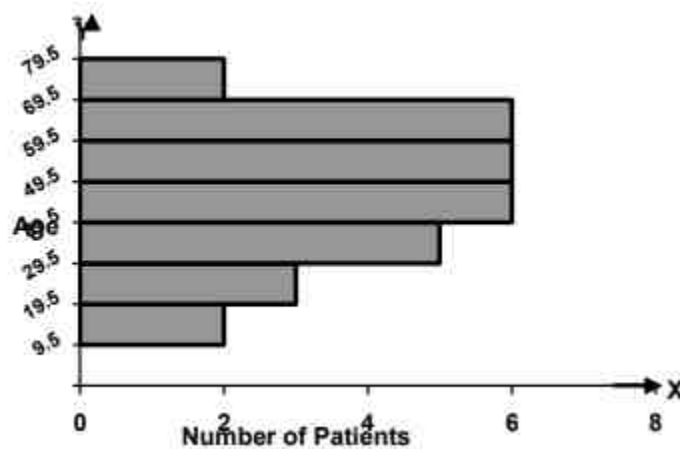
FREQUENCY DISTRIBUTION

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

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

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.

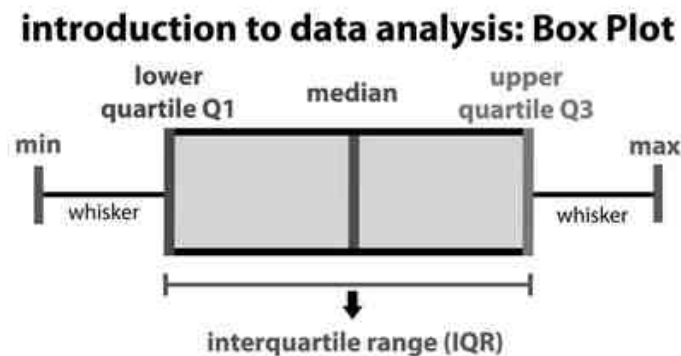
DESCRIPTION OF VARIABLE DATA:

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.

Topic 29

Describing Shape: From Box and Whisker Plot

In descriptive statistics, a box plot or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and Skewness through displaying the data quartiles (or percentiles) and averages. Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.



Definitions

Minimum Score

The lowest score, excluding outliers (shown at the end of the left whisker).

Lower Quartile

Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile).

Median

The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value and half are less.

Upper Quartile

Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value.

Maximum Score

The highest score, excluding outliers (shown at the end of the right whisker).

Whiskers

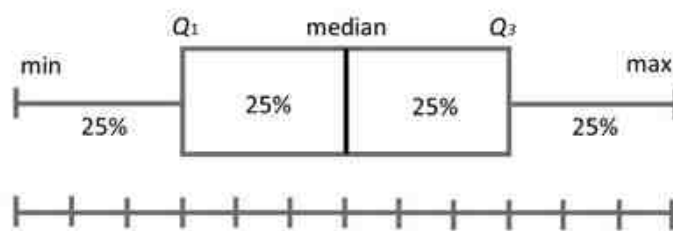
The upper and lower whiskers represent scores outside the middle 50% (i.e. the lower 25% of scores and the upper 25% of scores).

The Interquartile Range (or IQR)

This is the box plot showing the middle 50% of scores (i.e., the range between the 25th and 75th percentile).

Why are box plots useful?

Box plots divide the data into sections that each contain approximately 25% of the data in that set.



Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.

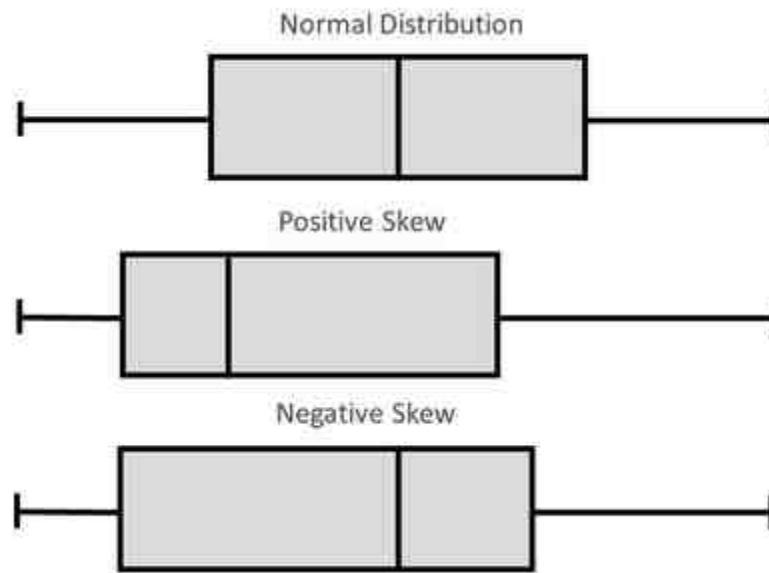
Note the image above represents data which is a perfect normal distribution and most box plots will not conform to this symmetry (where each quartile is the same length).

Box plots are useful as they show the average score of a data set.

The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

Box plots are useful as they show the skewness of a data set

The box plot shape will show if a statistical data set is normally distributed or skewed.

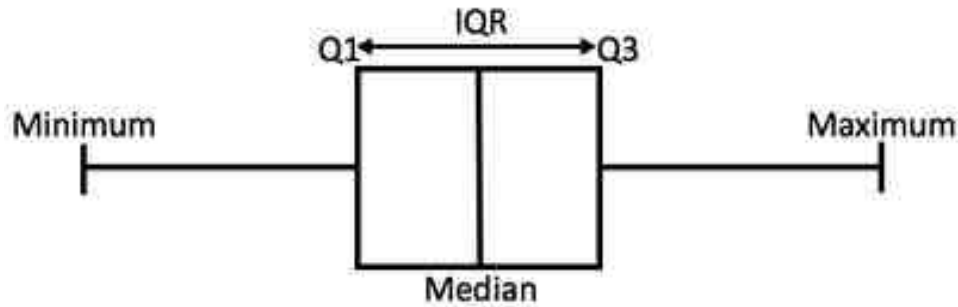


- When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.
- When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).
- When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).

Box plots are useful as they show the dispersion of a data set.

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.

The smallest value and largest value are found at the end of the ‘whiskers’ and are useful for providing a visual indicator regarding the spread of scores (e.g. the range).

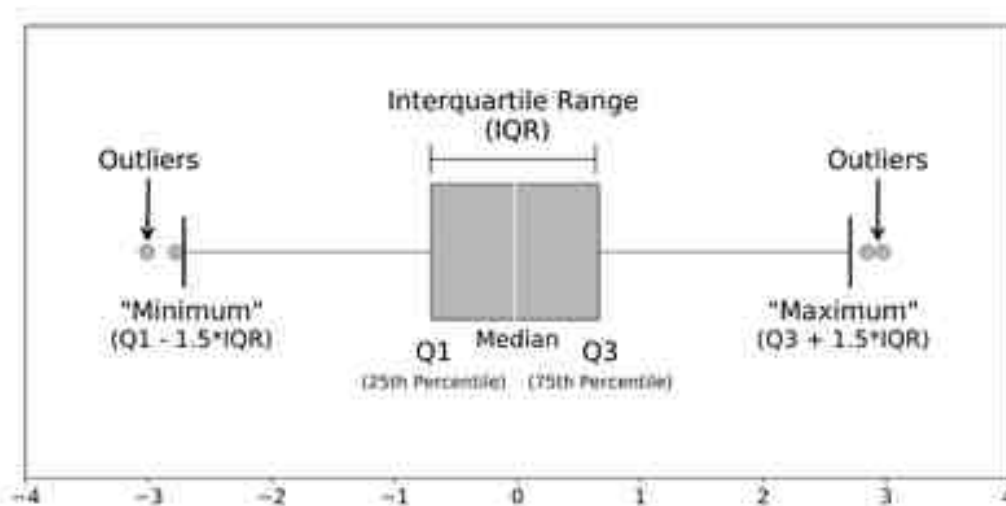


The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. $Q3 - Q1$).

Box plots are useful as they show outliers within a data set.

An outlier is an observation that is numerically distant from the rest of the data.

When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.



For example, outside 1.5 times the interquartile range above the upper quartile and below the lower quartile ($Q1 - 1.5 * IQR$ or $Q3 + 1.5 * IQR$).

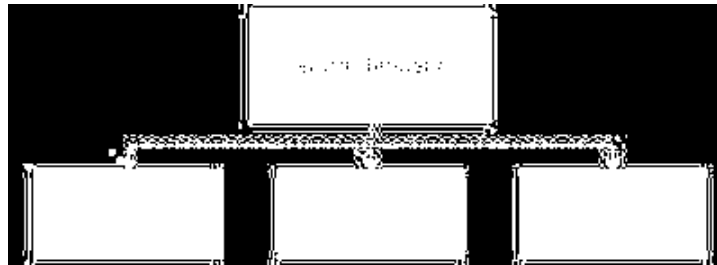
Topic:30

The Central Tendency of a data-set

What is Central Tendency?

Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.

The central tendency is one of the most quintessential concepts in statistics. Although it does not provide information regarding the individual values in the dataset, it delivers a comprehensive summary of the whole dataset.



- Mode: the most frequent value.
- Median: the middle number in an ordered dataset.
- Mean: the sum of all values divided by the total number of values.

AVERAGES (I.E. MEASURES OF CENTRAL TENDENCY)

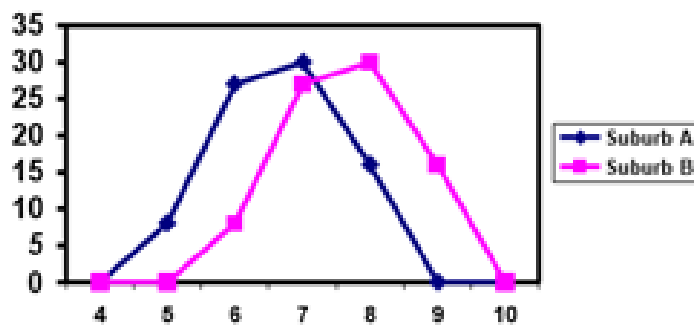
An average is a single value which is intended to represent a set of data or a distribution as a whole. It is more or less CENTRAL value ROUND which the observations in the set of data or distribution usually tend to cluster.

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		

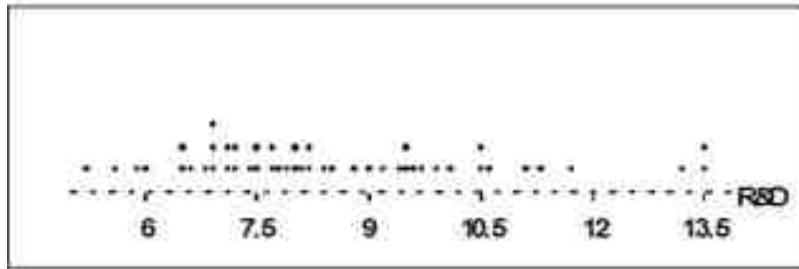
We can represent this data by means of a plot that is called dot plot.

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

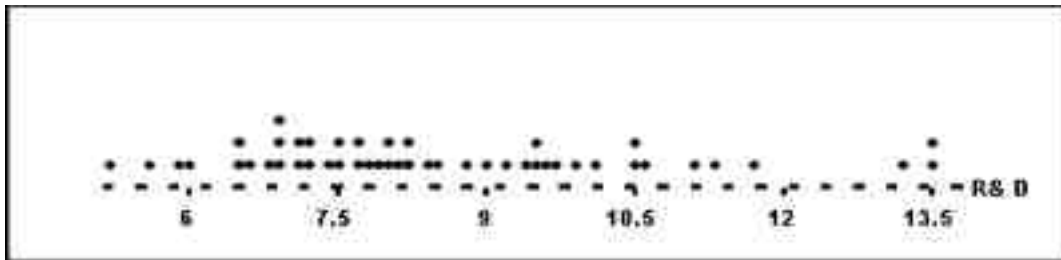
Dot Plot



*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-Hall 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.



$$X = 6.9$$

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 $f_m = 13$ 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

h = length of class interval of the modal class

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

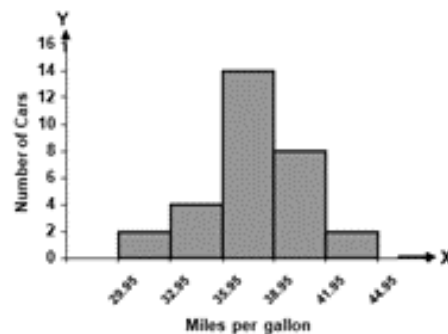
EPA MILEAGE RATINGS

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

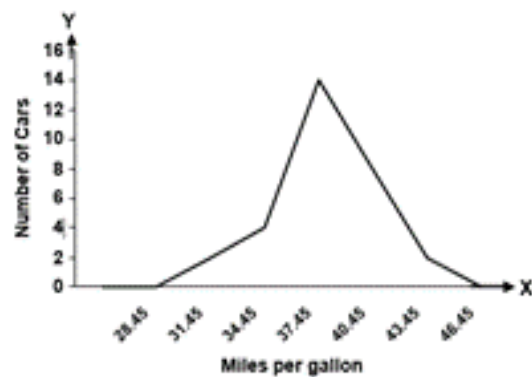
It is evident that the third class is the modal class. The mode lies somewhere between 35.95 and 38.95. In order to apply the formula for the mode, we note that $f_m = 14$, $f_1 = 4$ and $f_2 = 8$ 



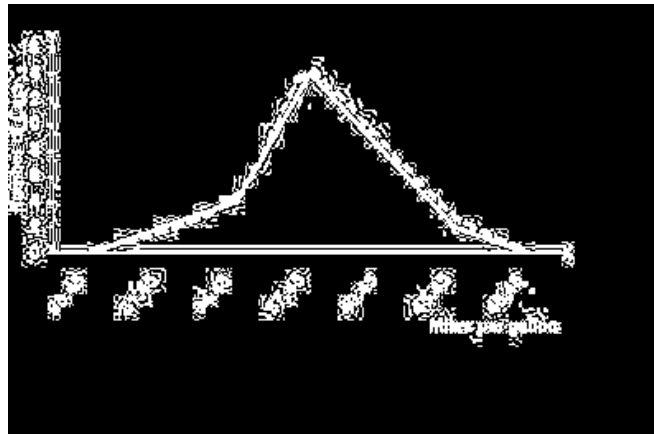
Let us now perceive the mode by considering the graphical representation of our frequency distribution. You will recall that, for the example of EPA Mileage Ratings, the histogram was as shown below:



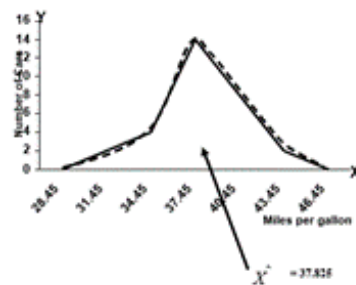
The frequency polygon of the same distribution was:



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.

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.”

$$\bar{X} = \frac{\sum_i X_i}{n}$$

Where n represents the number of observations in the sample that has been the *i*th observation in the sample (*i* = 1, 2, 3,..., n), and represents the mean of the sample.

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:

$$= £ 259.85/7 = £ 37.12 \text{ (To the nearest penny).}$$

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 X_i .

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	Frequency
X	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:

$$\bar{X} = \frac{\sum fx}{\sum f}$$

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:

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



$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{1135.5}{30} = 37.85$$

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.

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.

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

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 upto 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

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

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

$$\text{Median} = l + \left(\frac{\frac{n}{2} - c}{f} \right) \times h$$

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

$n = \sum f$ (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

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

EMPIRICAL RELATION BETWEEN THE MEAN, MEDIAN AND THE MODE

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

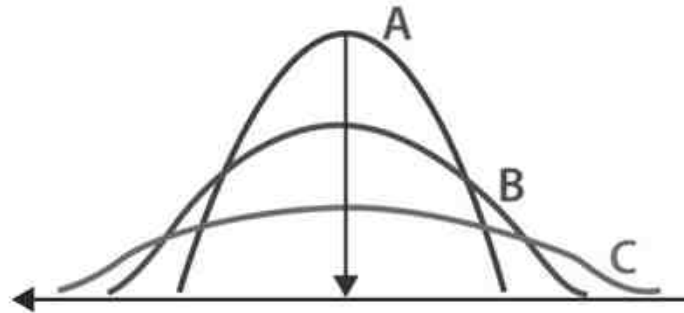
An exactly similar situation holds in case of a moderately negatively skewed distribution.

An important point to note is that this empirical relation does not hold in case of a J-shaped or an extremely skewed distribution.

Topic 31

Measures of Dispersion

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.



Measures of Dispersion

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. The absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3, 5, 6, 7 $\Rightarrow$ Range = 7 - 1 = 6
2. **Variance:** Deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total no of values in the data set to get the variance.
Variance (σ^2) = $\sum(X-\mu)^2/N$

3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. $S.D. = \sqrt{\sigma}$.
4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

1. Co-efficient of Range
2. Co-efficient of Variation
3. Co-efficient of Standard Deviation
4. Co-efficient of Quartile Deviation
5. Co-efficient of Mean Deviation

Co-efficient of Dispersion

The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages. The dispersion coefficient is also used when two series with different measurement units are compared. It is denoted as C.D.

The common coefficients of dispersion are:

C.D. in terms of	Coefficient of dispersion
Range	$C.D. = (X_{\max} - X_{\min}) / (X_{\max} + X_{\min})$
Quartile Deviation	$C.D. = (Q_3 - Q_1) / (Q_3 + Q_1)$
Standard Deviation (S.D.)	$C.D. = S.D. / \text{Mean}$
Mean Deviation	$C.D. = \text{Mean deviation} / \text{Average}$

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.

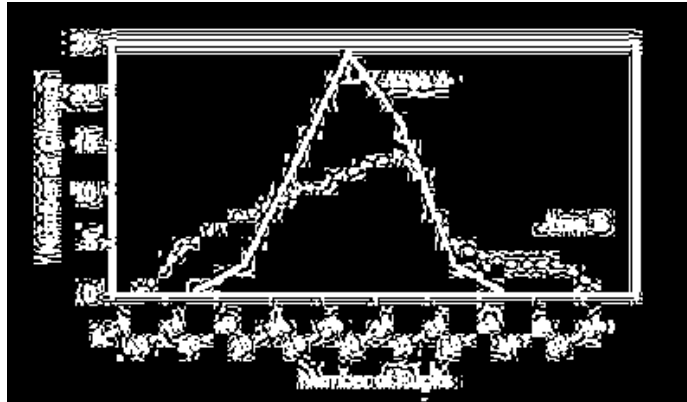
EXAMPLE

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.

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.



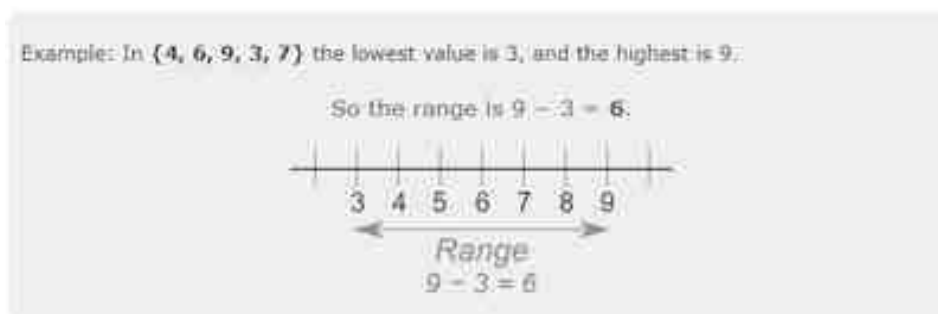
The question which must be posed and answered is ‘In what way can these two situations be distinguished?’ We need a measure of variability or DISPERSION to accompany the relevant measure of position or ‘average’ used.

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’.

RANGE

The range is defined as the difference between the two extreme values of a data-set, i.e. $R = X_m - X_0$ where X_m represents the highest value and X_0 the lowest.

The Range is the difference between the lowest and highest values.



Topic 32

Mean Deviation, Standard Deviation and Variance & Coefficient of variation

Mean Deviation

The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set. The mean deviation of the data values can be easily calculated using the below procedure.

MEAN DEVIATION

$$M.D = \frac{\sum |X - \bar{X}|}{n} \quad (\text{for ungrouped data})$$

$$M.D = \frac{\sum f |X - \bar{X}|}{\sum f} \quad (\text{for grouped data})$$

Σ represents the addition of values

X represents each value in the data set

μ represents the mean of the data set

N represents the number of data values

$||$ represents the absolute value, which ignores the “-” symbol

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”. If we are using median instead of mean then it is called as **median deviation**. When will we have a situation when we will be using the median instead of the mean? 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.

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

$$\text{CO-EFFICIENT OF M.D} = \frac{M.D}{\text{Mean}}$$

Example

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 $\bar{X} = 4$

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

X	$X - \bar{X} = d$	$ d $
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

$$M.D = \frac{\sum |X - \bar{X}|}{n}$$

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

$$M.D = 2$$

STANDARD DEVIATION

Standard deviation is expressed in absolute terms and is given in the same unit of measurement as the variable itself. The standard deviation is an absolute measure of dispersion. Its relative measure called **coefficient of standard deviation or coefficient of variation**. 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.

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

For ungrouped data.

Standard deviation can be calculated as for ungrouped data.

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

THE STANDARD DEVIATION IN CASE OF GROUPED DATA

Formula

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

Variance

Variance is the expected value of the squared variation of a random variable from its mean value, in probability and statistics. Informally, variance estimates how far a set of numbers (random) are spread out from their mean value.

The value of variance is equal to the square of standard deviation, which is another central tool.

Variance is symbolically represented by σ^2 , s^2 , or **Var(X)**.

The formula for variance is given by:

For ungrouped data.

Variance can be calculated as for ungrouped data.

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

THE Variance IN CASE OF GROUPEd DATA

Formula

$$Var = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f} \right)^2$$

Example.

Calculate the standard deviation and variance of:

4,6,8,9 and 11.

Solution:

Let the given values be denoted by X

X	X^2
1	1
3	9
5	25
6	36
8	64
$\sum X = 23$	$\sum X^2 = 135$

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

$$S.D = \sqrt{\frac{135}{5} - \left(\frac{23}{5}\right)^2}$$

$$S.D = 2.42$$

$$Var = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2$$

$$Var = \frac{135}{5} - \left(\frac{23}{5}\right)^2$$

$$Var = 5.85$$

THE STANDARD DEVIATION IN CASE OF GROUPED DATA

Formula

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

Example:

Compute the standard deviation and variance from the following distribution of marks.

Marks	No. of Students
1-3	40
3-5	30
5-7	20
7-9	10

Solution:

Marks	No. of Students	X	fX	fX^2
1-3	40	2	80	160
3-5	30	4	120	480
5-7	20	6	120	720
7-9	10	8	80	640
Total	$\sum f = 100$		$\sum fX = 400$	$\sum fX^2 = 2000$

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

$$S.D = \sqrt{\frac{2000}{100} - \left(\frac{400}{100}\right)^2}$$

$$S.D = \sqrt{20 - 16}$$

$$S.D = \sqrt{4}$$

$$S.D = 2$$

$$Var = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$Var = \frac{2000}{100} - \left(\frac{400}{100}\right)^2$$

$$Var = 4$$

Coefficient of standard deviation

$$\text{Coefficient of standard deviation} = \frac{\text{Standard deviation}}{\text{Mean}}$$

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

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

Topic 33

Chebyshev's Inequality, The Empirical Rule & The Five-Number Summary

Chebyshev's Theorem

Chebyshev's Theorem tells us that no matter what the distribution looks like, the probability that a randomly selected values is in the interval

$$\mu \pm k \text{ is at least } \left[100\left(1 - \frac{1}{k^2}\right) \right] \%$$

The Empirical Rule applies to data sets with frequency distributions that are mound-shaped and symmetric. The Empirical Rule allows us to estimate the answers for intervals that are symmetrical about the mean.

According to this empirical rule:

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

Example:

If mean of a distribution is 20 and standard deviation is 4. Find out the limits by applying Chebyshev's Inequality for $k=2$ and $k=3$. How much data will lie between these two limits?

Solution:

For $k=2$ According to Chebyshev's Inequality

$$\bar{x} \pm k\sigma$$

$$20 \pm 2 \times 4$$

$$20 \pm 8$$

$$12, 28$$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

These limits will contain At least 75% of the data.

For $k=3$,

$$\bar{x} \pm k\sigma$$

$$20 \pm 3 \times 4$$

$$20 \pm 12$$

$$8, 32$$

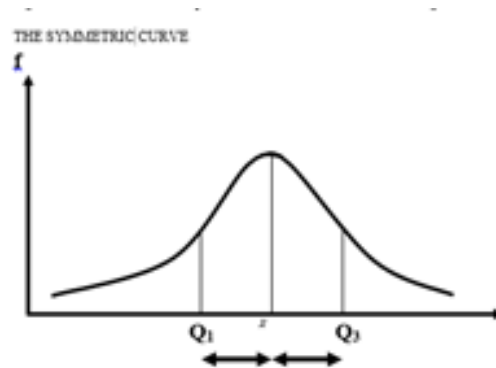
$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 88.89\%$$

These limits will contain at least 88.89% of the data.

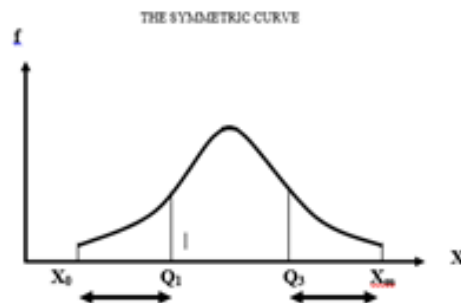
FIVE-NUMBER SUMMARY

A five-number summary consists of X_0, Q_1 , Median, Q_3 , and X_m ; It provides us quite a good idea about the shape of the distribution. If the data were perfectly symmetrical, the following would be true:

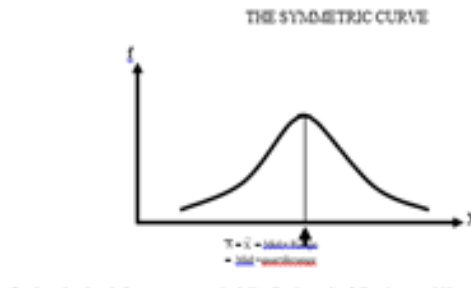
1. The distance from Q_1 to the median would be equal to the distance from the median to Q_3 :



The distance from X_0 to Q_1 would be equal to the distance from Q_3 to X_m .

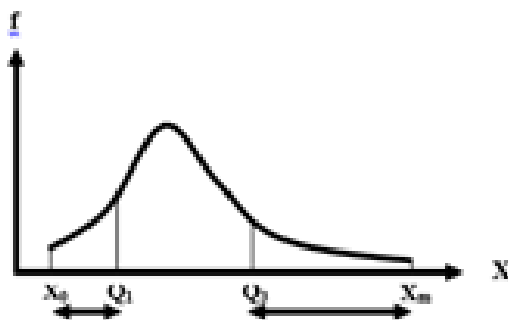


3. The median, the mid-quartile range, and the midrange would all be equal. All these measures would also be equal to the arithmetic mean of the data:



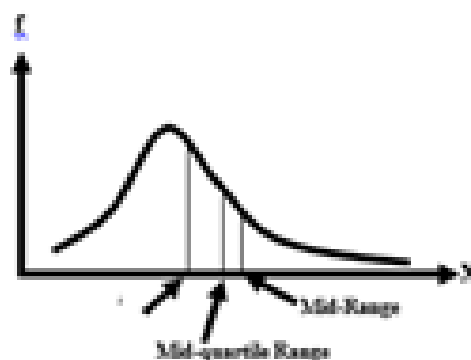
On the other hand, for non-symmetrical distributions, the following would be true:

1. In right-skewed distributions the distance from Q_3 to X_m greatly exceeds the distance from X_0 to Q_1 .



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

THE POSITIVELY SKEWED CURVE



Similarly, in left-skewed distributions, the distance from X_0 to Q_1 greatly exceeds the distance from Q_3 to X_m . Also, in left-skewed distributions, $\text{midrange} < \text{mid-quartile range} < \text{median}$.

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:

X₀	=	14.3	14.9	15.2	15.2	15.4	15.6	16.4	17.0	X_m	=
13.0										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	$\sim X$	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 Q_3 to X_m (i.e., 6.7) greatly exceeds the distance from X_0 to Q_1 (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.

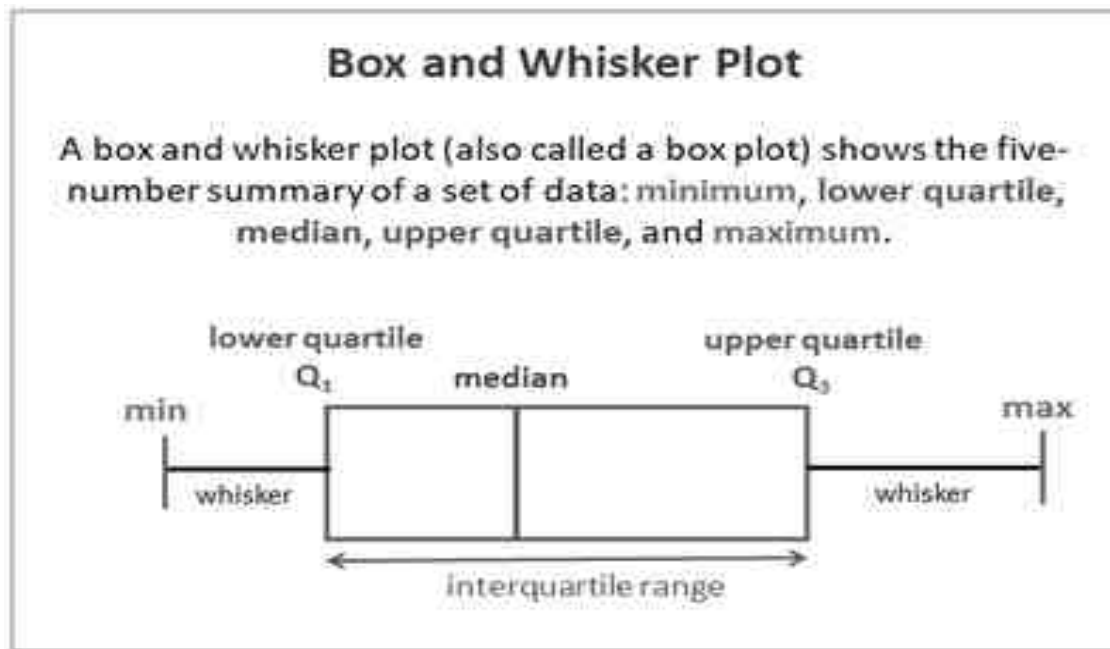
Topic 34

Box and Whisker Plot, Pearson's Coefficient of Skewness

Box and whisker plot

A box plot (also called a box and whisker plot) shows data using the middle value of the data and the quartiles, or 25% divisions of the data.

The following diagram shows a box plot or box and whisker plot.



Drawing A Box And Whisker Plot

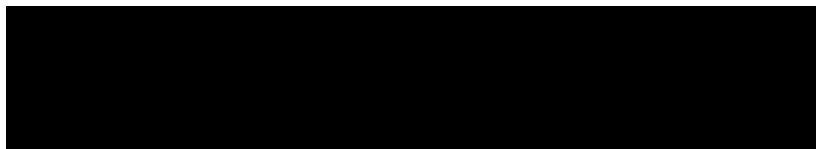
Example:

Construct a box plot for the following data:
12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution:

Step 1: Arrange the data in ascending order.

Step 2: Find the median, lower quartile and upper quartile.



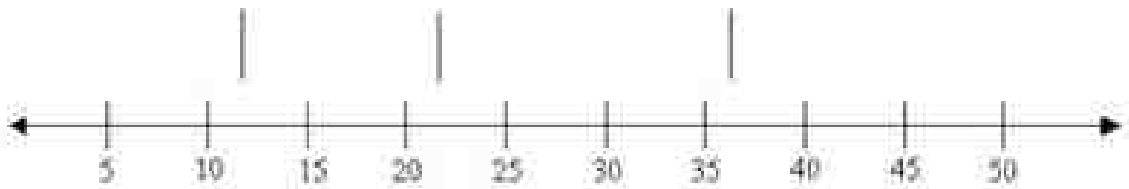
Median (middle value) = 22
 Lower quartile (middle value of the lower half) = 12
 Upper quartile (middle value of the upper half) = 36

(If there is an even number of data items, then we need to get the average of the middle numbers.)

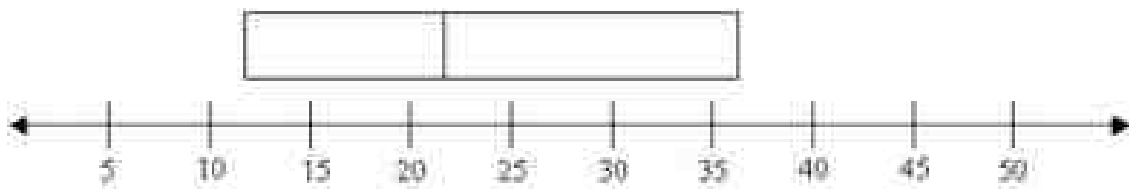
Step 3: Draw a number line that will include the smallest and the largest data.



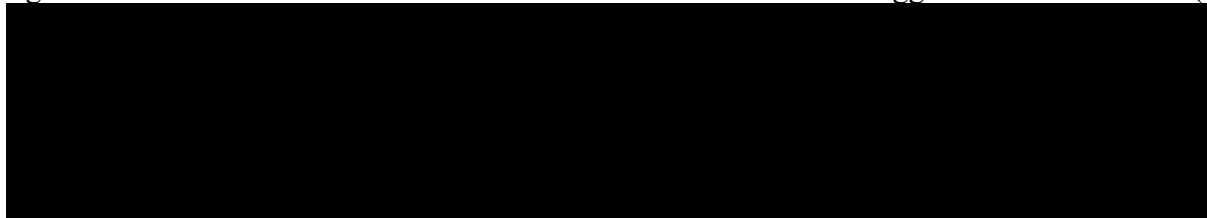
Step 4: Draw three vertical lines at the lower quartile (12), median (22) and the upper quartile (36), just above the number line.



Step 5: Join the lines for the lower quartile and the upper quartile to form a box.



Step 6: Draw a line from the smallest value (5) to the left side of the box and draw a line from the right side of the box to the biggest value (53).



Interquartile range

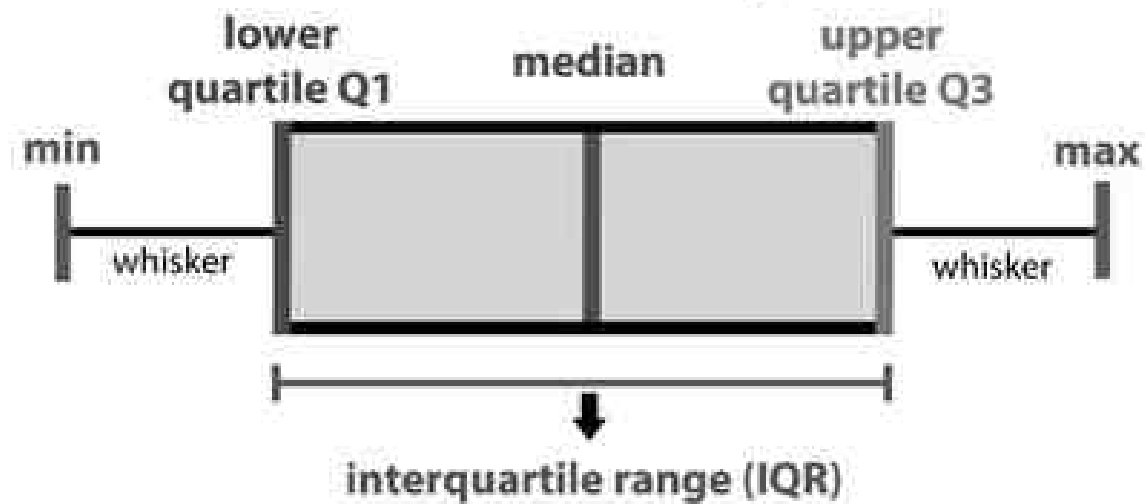
The interquartile range is the difference between the upper quartile and the lower quartile.

$$53 - 5 = 48$$

In descriptive statistics, a box plot or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and Skewness through displaying the data quartiles (or percentiles) and averages.

Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.

introduction to data analysis: Box Plot



Definitions

Minimum Score

The lowest score, excluding outliers (shown at the end of the left whisker).

Lower Quartile

Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile).

Median

The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value and half are less.

Upper Quartile

Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value.

Maximum Score

The highest score, excluding outliers (shown at the end of the right whisker).

Whiskers

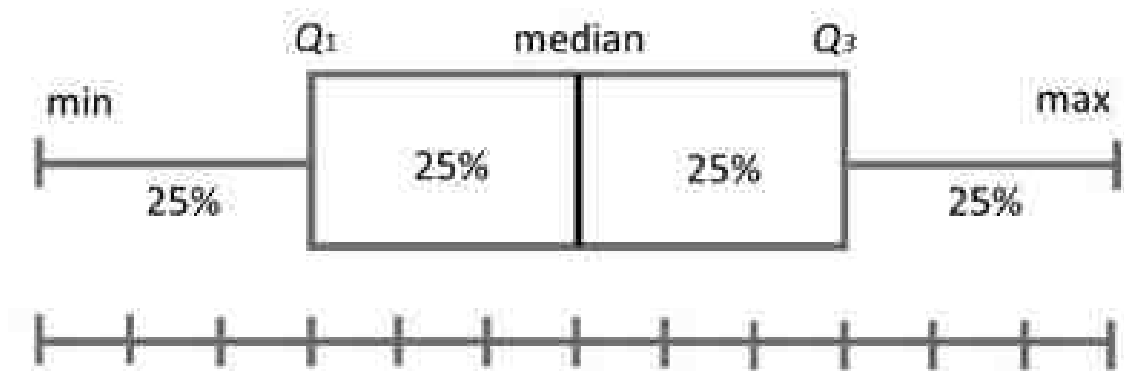
The upper and lower whiskers represent scores outside the middle 50% (i.e. the lower 25% of scores and the upper 25% of scores).

The Interquartile Range (or IQR)

This is the box plot showing the middle 50% of scores (i.e., the range between the 25th and 75th percentile).

Why are box plots useful?

Box plots divide the data into sections that each contain approximately 25% of the data in that set.



Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.

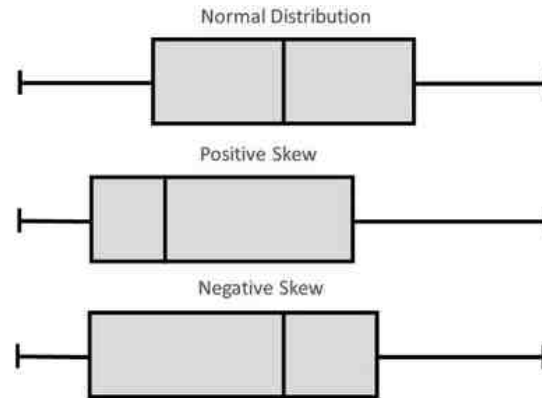
Note the image above represents data which is a perfect normal distribution and most box plots will not conform to this symmetry (where each quartile is the same length).

Box plots are useful as they show the average score of a data set.

The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

Box plots are useful as they show the skewness of a data set

The box plot shape will show if a statistical data set is normally distributed or skewed.

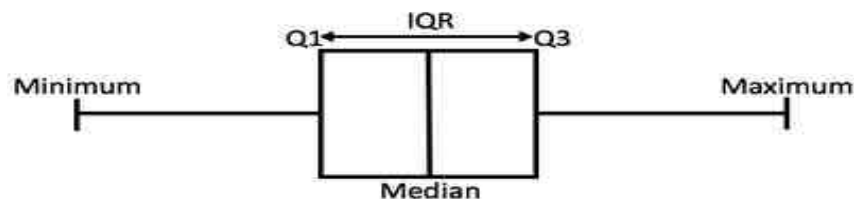


- When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.
- When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).
- When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).

Box plots are useful as they show the dispersion of a data set.

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.

The smallest value and largest value are found at the end of the 'whiskers' and are useful for providing a visual indicator regarding the spread of scores (e.g. the range).

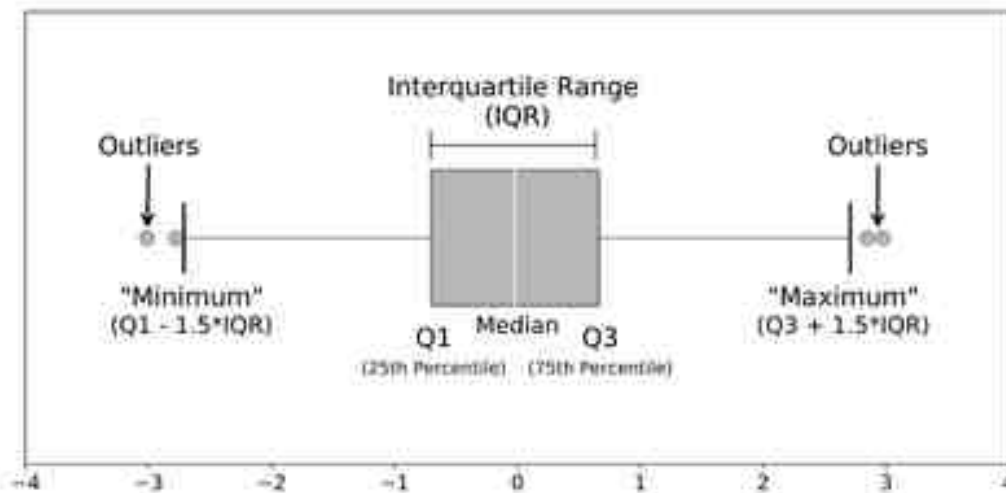


The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. $Q3 - Q1$).

Box plots are useful as they show outliers within a data set.

An outlier is an observation that is numerically distant from the rest of the data.

When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.



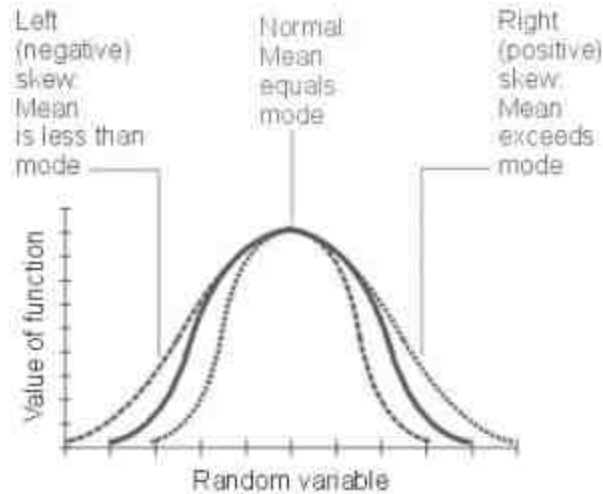
For example, outside 1.5 times the interquartile range above the upper quartile and below the lower quartile ($Q1 - 1.5 * IQR$ or $Q3 + 1.5 * IQR$).

SKEWNESS

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or average, and the mode, or maximum point on the curve, are equal.

- In a perfect normal distribution (green solid curve in the illustration below), the tails on either side of the curve are exact mirror images of each other.
- When a distribution is skewed to the left (red dashed curve), the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative Skewness.
- When a distribution is skewed to the right (blue dotted curve), the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive Skewness.



Pearson's Coefficient of Skewness

Coefficient of Skewness. The coefficient of skewness measures the skewness of a distribution. It is based on the notion of the moment of the distribution. This coefficient is one of the measures of skewness.

Pearson's coefficient of skewness is calculated as:

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

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):

$$\begin{array}{c} \text{Ages of Children} \\ \text{of Manual Workers} \\ 3 \left(\frac{8.50 - 8.50}{3.61} \right) \\ = 0 \end{array}$$

$$\begin{array}{c} \text{Ages of Children} \\ \text{of Non-Manual Workers} \\ 3 \left(\frac{8.50 - 9.16}{3.61} \right) \\ = -0.55 \end{array}$$

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.

Topic 35

Set Theory, Counting Rules

“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.

Examples of Sets:

$$A = \{a, b, c, d\} \text{ or}$$

$$B = \{1, 2, 3, 7\}$$

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.

$$A = \{1, 3, 5, 7, 9, 11\} ;$$

$$B = \{\text{a book, a city, a clock, a teacher}\};$$

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.

$A = \{x \mid x \text{ is an odd number and } x < 12\}$ 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) $A = \{1, 2, 3, \dots, 99, 100\};$
- ii) $B = \{x \mid x \text{ is a month of the year}\};$
- iii) $C = \{x \mid x \text{ is a printing mistake in a book}\};$
- iv) $D = \{x \mid x \text{ is a living citizen of Pakistan}\};$

Examples of infinite sets:

- i) $A = \{x \mid x \text{ is an even integer}\};$
- ii) $B = \{x \mid x \text{ is a real number between 0 and 1 inclusive}\},$
i.e. $B = \{x \mid 0 \leq x \leq 1\}$
- iii) $C = \{x \mid x \text{ is a point on a line}\};$
- iv) $D = \{x \mid x \text{ is a sentence in a English language}\}; \text{ etc}$

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

$$B \subset A \text{ and } B \neq A$$

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 2^n 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 $2^3 = 8$ subsets of the set A .

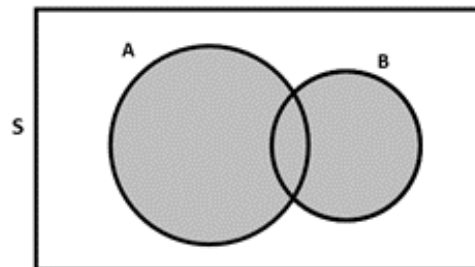
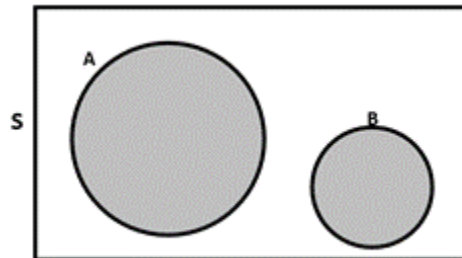
VENN DIAGRAM

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.

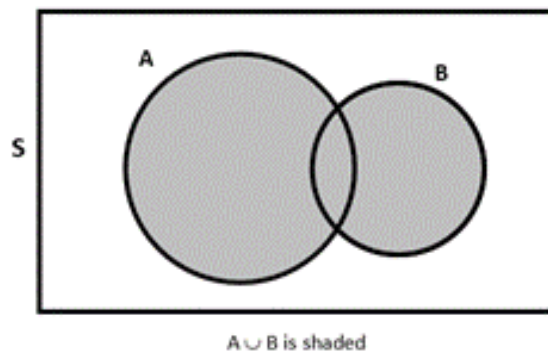
The basic operations are union, intersection, difference and complementation.

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:



Example

Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$

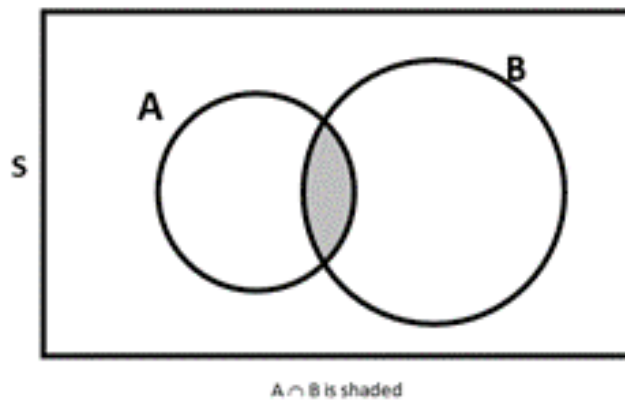
Then $A \cup B = \{1, 2, 3, 4, 5, 6\}$

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

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

Diagrammatically, $A \cap B$ is shown by the shaded area as below:



Example

$$\text{Let } A = \{1, 2, 3, 4\} \text{ and } B = \{3, 4, 5, 6\}$$

$$\text{Then } A \cap B = \{3, 4\}$$

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

$$\text{i.e. } A \cap B = \phi.$$

On the other hand, two sets A and B are said to be conjoint when they have at least one element in common.

SET DIFFERENCE

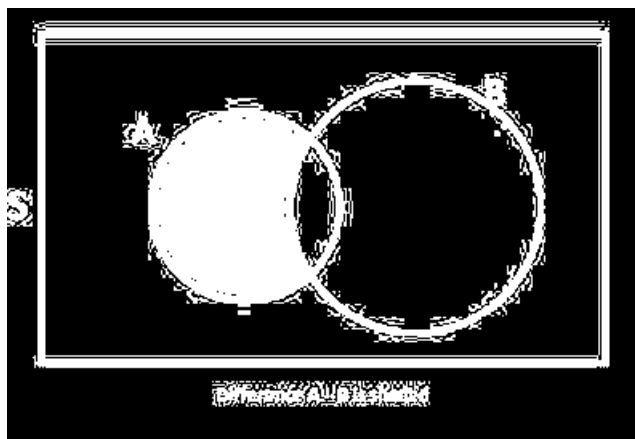
The difference of two sets A and B , denoted by $A - B$ or by $A - (A \cap B)$, is the set of all elements of A which do not belong to B .

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 $A - B$:



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

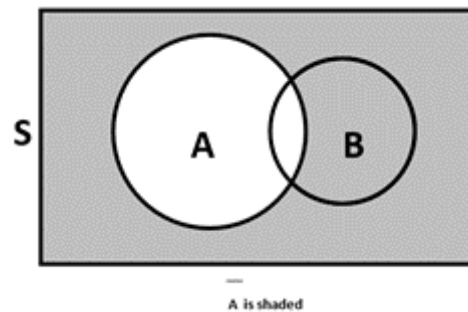
The particular difference $S - A$, that is, the set of all those elements of S which do not belong to A , is called the complement of A and is denoted by $\bar{A}$ or by A^c .

In symbols:

$$\bar{A} = \{x \mid x \in S \text{ and } x \notin A\}$$

The complement of S is the empty set ϕ .

The complement of A is shown by the shaded portion in the following Venn diagram.



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:

Counting Rules

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

$$A \cup S = S,$$

$$A \cap S = A,$$

$$A \cup \phi = A, \text{ and}$$

$$A \cap \phi = \phi.$$

6. Complementation laws

$$A \cup \bar{A} = S,$$

$$A \cap \bar{A} = \phi,$$

$$(A^c)^c = A,$$

$$\bar{\bar{S}} = \phi, \text{ and}$$

$$\bar{\phi} = S$$

7. De Morgan's laws:

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B}$$

$$\overline{(A \cap B)} = \bar{A} \cup \bar{B}$$

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) $A_1 \cap A_2 \cup \dots \cup A_n = S$.

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 $P(A) = \{\phi, \{H\}, \{T\}, \{H, T\}\}$.

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 , 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.

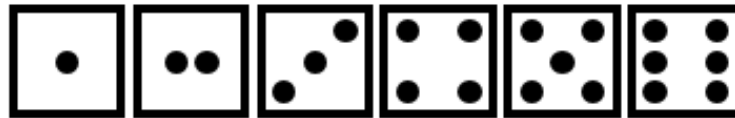
EXAMPLE

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:

$$A \times B = \{(H, 1); (H, 2); (H, 3); (H, 4); (H, 5); (H, 6); (T, 1); (T, 2); (T, 3); (T, 4); (T, 5); (T, 6)\}$$

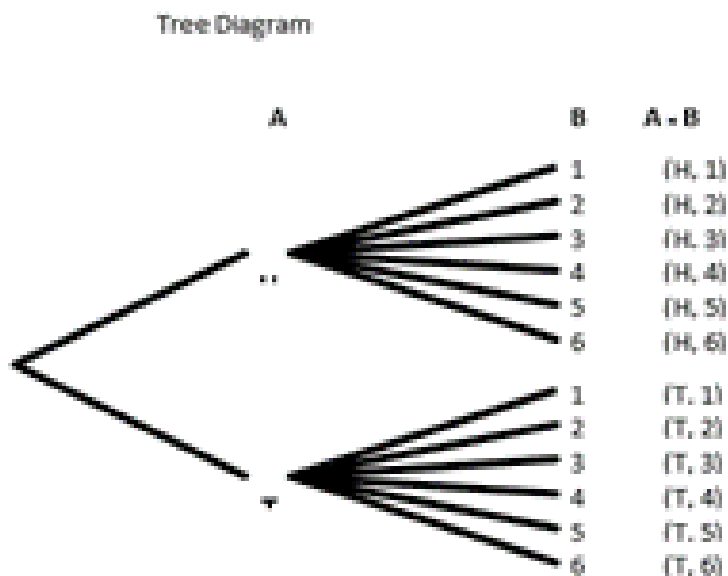
Clearly, these twelve elements together make up the universal set S when a COIN and a DIE are tossed together.

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 $A \times B$ may conveniently be found by means of the so-called tree diagram shown below:



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:

$$A \times B = \{ (H, 1); (H, 2); (H, 3); (H, 4);$$

$$(H, 6); (H, 6); (T, 1); (T, 2);$$

$$(T, 3); (T, 4); (T, 5); (T, 6) \}$$

Topic 36

Permutations, Combinations

- Permutations
- Combinations

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

$$\begin{aligned} {}^nP_r &= n(n-1)(n-2) \dots (n-r+1) \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

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 $0! = 1$.

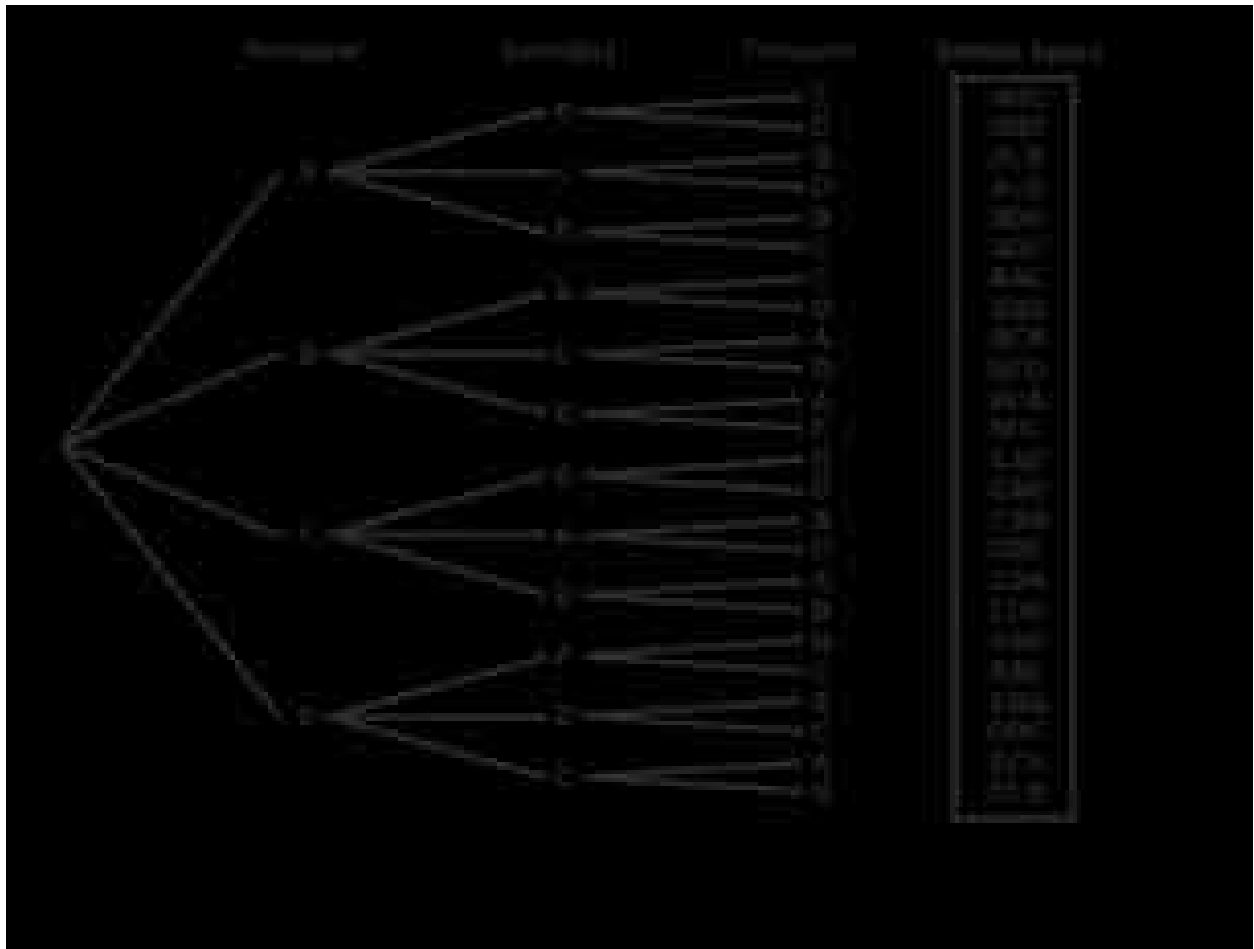
Example

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 ${}^nP_r$, if we put $r = n$, we obtain:

$$\begin{aligned} {}^nP_n &= n(n-1)(n-2) \dots 3 \times 2 \times 1 \\ &= n! \end{aligned}$$

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

$${}^3P_3 = 3! = 6$$

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 n_1 of one kind, n_2 of a second kind, ..., n_k of a K th kind,

$$is \ P = \frac{n!}{n_1! n_2! \dots n_k!}.$$

(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)

$n_1 = 1$ (because there is one c)

$n_2 = 1$ (because there is one o)

$n_3 = 2$ (because there are two m’s)

$n_4 = 1$ (because there is one i)

$n_5 = 2$ (because there are two t’s)

and

$n_6 = 2$ (because there are two e’s)

Hence, the total number of (meaningless) words (permutations) is:

$$\begin{aligned} P &= \frac{n!}{n_1! n_2! \dots n_k!} \\ &= \frac{9!}{1! 1! 2! 1! 2! 2!} \\ &= \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{1 \times 1 \times 2 \times 1 \times 1 \times 2 \times 1 \times 2 \times 1} \\ &= 45360 \end{aligned}$$

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:

$${}^nC_r \text{ or } \binom{n}{r},$$

and is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

where $r < n$.

It should be noted that 

$${}^nP_r = r! \binom{n}{r}$$

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 ${}^nC_r$ 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.

$$i): \quad \binom{n}{r} = \binom{n}{n-r}$$

$$ii): \quad \binom{n}{n-r} + \binom{n}{r} = \binom{n+1}{r}$$

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

$$\begin{aligned} \binom{n}{r} &= \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!} \\ &= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \\ &= 120 \end{aligned}$$

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.

Topic 37

Definitions of Probability

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:

$$\begin{aligned} S = & \{ (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \} \end{aligned}$$

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 A^c 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 2^n 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 $2^3 = 8$ possible subsets are

$$\begin{aligned} &\phi, \{a\}, \{b\}, \{c\}, \{a, b\}, \\ &\{a, c\}, \{b, c\}, \{a, b, c\} \end{aligned}$$

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 cards, 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.

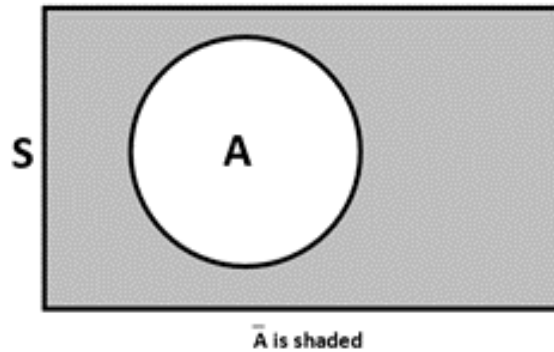
EXAMPLES

- 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.



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 $1/52$.

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 cards, it *can* be both a king and a diamond.

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

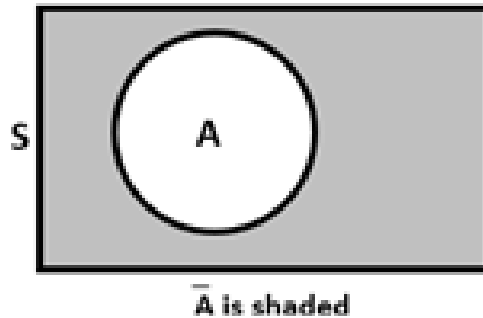
- In the coin-tossing experiment, ‘head’ and ‘tail’ are collectively exhaustive events.
- In the die-tossing experiment, ‘even number’ and ‘odd number’ are collectively exhaustive events.

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



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

If a random experiment can produce n mutually exclusive and equally likely outcomes, and if m out of 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

$$\begin{aligned} P(A) &= \frac{m}{n} \\ &= \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}} \end{aligned}$$

This definition was formulated by the French mathematician P.S. Laplace (1749-1827) 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*).

$$\begin{aligned} P(A) &= \frac{m}{n} \\ &= \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}} \\ &= \frac{26}{52} = \frac{1}{2} \end{aligned}$$

$$P(B) = \frac{4}{52} = \frac{1}{13}$$

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.

$$n(A) = \binom{3}{0} \binom{9}{4} + \binom{3}{1} \binom{9}{3}$$

= 126 + 252 = 378 sample points.

Then

$$\therefore P(A) = \frac{m}{n} = \frac{378}{495} = 0.76$$

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.

Topic 38

Application of Addition Theorem, Conditional Probability & Multiplication Theorem

Applications of Probability:

- 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

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &= 4/36 + 2/36 = 6/36 \\ &= 16.67\%. \end{aligned}$$

COROLLARY-2

If $A_1, A_2, \dots, A_k$ 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.

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

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

$$p + 2p + 4p = 1$$
$$\text{or } p = 1/7$$

$$\therefore P(C) = 1/7,$$

$$P(B) = 2(1/7) = 2/7,$$

$$\text{and } P(A) = 4(1/7) = 4/7.$$

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.

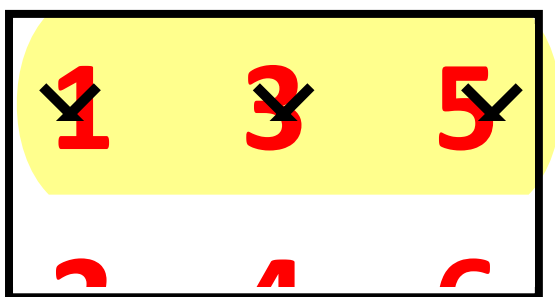
The following example illustrates the concept of conditional probability

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,

$$P(\text{die shows 6} / \text{die shows even numbers}) = 1/3,$$

(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)** $P(S/B) = 1$
- $P(A1 \cup A2/B) = P(A1/B) + P(A2/B)$ (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

$$P(B) = 0.75 \text{ and } P(A \cap B) = 0.20.$$

Thus, we obtain

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.20}{0.75} = 0.27$$

Multiplication Theorem of probability

It is interesting to note that the multiplication theorem is obtained very conveniently from the formula of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

As discussed earlier, the conditional probability of A given that B has occurred has already been defined as:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \text{ Where } P(B) > 0$$

Multiplying both sides by P (B), we get

$$P(A \cap B) = P(B) \cdot P(A/B).$$

And if we interchange the roles of A and B, we obtain:

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

Provided $P(A) > 0$.

MULTIPLICATION LAW

If A and B are any two events defined in a sample space S, then

$$P(A \cap B)$$

$$= P(A) P(B/A), \text{ provided } P(A) > 0,$$

$$= P(B) P(A/B) \text{ provided } P(B) > 0.$$

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

Type of Item	No. of Items
Defective	4
Good	11
Total	15

We have:

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. 

- 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. $\binom{13}{2}$.

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 $\binom{10}{2}$.

And, the total number of ways in which 2 white balls can be drawn out of 10 white balls is

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

$$P(A_1) = \frac{\binom{10}{2}}{\binom{13}{2}} = \frac{45}{78},$$

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

$$P(A_2) = \frac{\binom{10}{1} \binom{3}{1}}{\binom{13}{2}} = \frac{30}{78},$$

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

$$P(A_3) = \frac{\binom{3}{2}}{\binom{13}{2}} = \frac{3}{78}.$$

AFTER having transferred 2 balls from the first bag, the second bag contains

5 white and 5 black balls (if 2 white balls are transferred)

Colour of Ball	No. of	No. of
	Balls in Bag A	Balls in Bag B
White	$10 - 2 = 8$	$3 + 2 = 5$
Black	3	5
Total	$13 - 2 = 11$	$8 + 2 = 10$

Hence: $P(W/A_1) = 5/10$

ii) 4 white and 6 black balls (if 1 white and 1 black ball are transferred)

Colour of Ball	No. of	No. of
	Balls in Bag A	Balls in Bag B
White	$10 - 1 = 9$	$3 + 1 = 4$
Black	$3 - 1 = 2$	$5 + 1 = 6$
Total	$13 - 2 = 11$	$9 + 2 = 11$

Hence: $P(W/A_2) = 4/10$

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:

$$P(A \cap B) = P(A) \cdot P(B/A)$$

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:

$$S = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

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.

$$S = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

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 (C)	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-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

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, whereas 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.

$$P(\text{male birth}) = P(\text{male live-born or male stillborn})$$

$$= P(\text{male live-born}) + P(\text{male stillborn})$$

$$= 0.5021 + 0.0120$$

$$= 0.5141$$

EXAMPLE

$$P(\text{stillbirth/male birth})$$

$$P(\text{male birth and stillbirth})/P(\text{male birth})$$

$$= 0.0120/0.5141 = 0.0233$$

Topic 39

Bayes' Theorem, Discrete Random Variable

- 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 $A_1, A_2, \dots, A_k$ form a PARTITION of a sample space S (that is, the events A_i 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 A_i OCCURS, then

$$\text{for any } i, P(A_i / B) = \frac{P(A_i)P(B / A_i)}{\sum_{i=1}^k P(A_i)P(B / A_i)}, \quad \text{for } i = 1, 2, \dots, k.$$

Stated differently:

BAYES' THEOREM:

If $A_1, A_2, \dots$ and A_k are mutually exclusive events of which one must occur, then

$$P(A_i | B) = \frac{P(A_i) \cdot P(B | A_i)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2) + \dots + P(A_k) \cdot P(B | A_k)}$$

If $k = 2$, we obtain:

Bayes' Theorem for two mutually exclusive events A_1 and A_2 :

$$P(A_i | B) = \frac{P(A_i) \cdot P(B | A_i)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)}$$

Where $i = 1, 2$.

In other words $P(A_1 | B) = \frac{P(A_1) \cdot P(B | A_1)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)}$

$$P(A_2 | B) = \frac{P(A_2) \cdot P(B | A_2)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)} \text{ and}$$

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 A_1 denote the event that it emits EXCESSIVE amounts of pollutants, and let A_2 denote the event that a car does NOT emit excessive amounts of pollutants. (In other words, A_2 is the complement of A_1 .)

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, A_1 and A_2 are mutually exclusive and exhaustive events i.e. A_1 and A_2 form a PARTITION of the sample space S .

Hence, we are in a position to apply the Bayes' theorem.

We need to calculate $P(A_1|B)$, and, according to the Bayes' theorem:

$$P(A_1 | B) = \frac{P(A_1) \cdot P(B | A_1)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)}$$

Now, according to the data given in this problem:

$$P(A_1) = 0.25,$$

$$P(A_2) = 0.75 \text{ (as } A_2 \text{ is simply the complement of } A_1\text{),}$$

$$P(B|A_1) = 0.99,$$

$$\text{and } P(B|A_2) = 0.17$$

Substituting the above values in the Bayes' theorem, we obtain:

$$\begin{aligned}
P(A_1 | B) &= \frac{P(A_1) \cdot P(B | A_1)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)} \\
&= \frac{(0.25)(0.99)}{(0.25)(0.99) + (0.75)(0.17)} \\
&= \frac{0.2475}{0.2475 + 0.1275} \\
&= \frac{0.2475}{0.3750} \\
&= 0.66
\end{aligned}$$

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 A_1 and A_2 .

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 A_1 , A_2 and A_3 . The event “defective bolt” represents the event B. Hence, in this example, we need to determine $P(A_1/B)$.

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 $f/\sum f$ 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

$$0 \leq P(X_i) \leq 1$$

for each X_i ($i = 1, 2, \dots, 7$)

$$\sum p(X_i) = 1$$

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 $\sum_{i=1}^{\infty} P(X_i) = 1$)

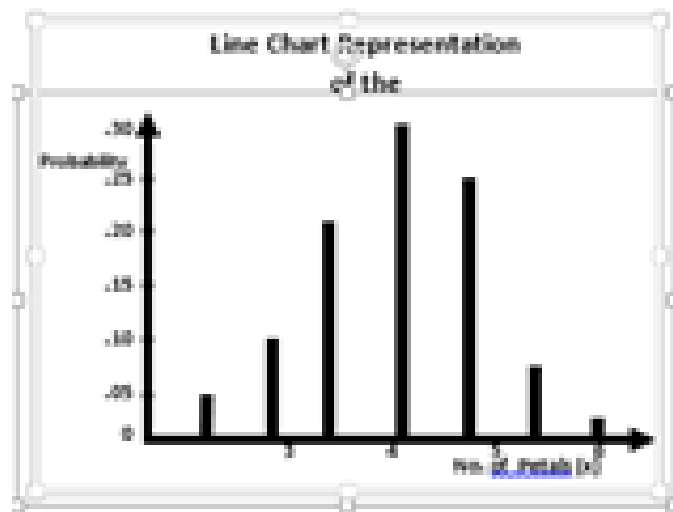
is known as a Discrete Random Variable, and its probability distribution is known as a Discrete Probability Distribution. Just as we can depict a frequency distribution graphically, we can draw the GRAPH of a probability distribution.

EXAMPLE

Going back to the probability distribution of the number of petals on the flowers of a particular 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.



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

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

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

$$\begin{aligned} S &= \sqrt{\frac{\sum f(X - \bar{X})^2}{\sum f}} \\ &= \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2} \\ &= \sqrt{\frac{\sum X^2 f}{\sum f} - \left(\frac{\sum Xf}{\sum f}\right)^2} \end{aligned}$$

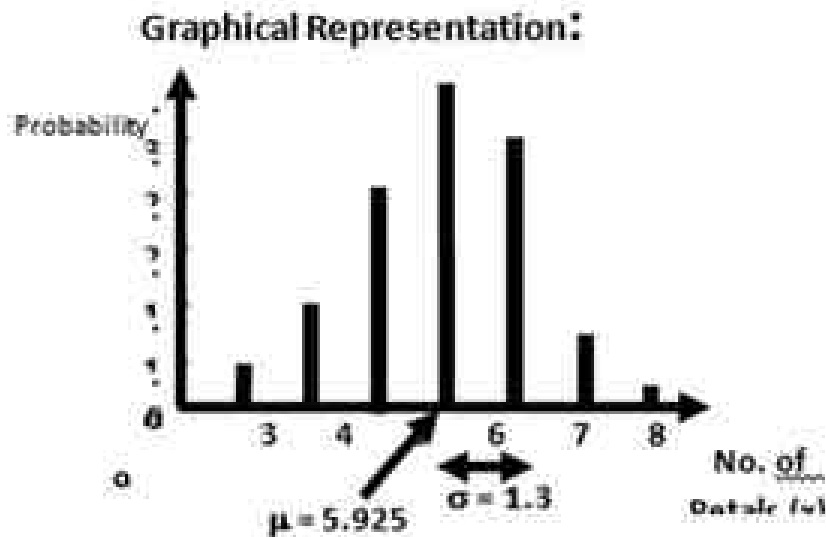
Similarly, in case of a probability distribution, we have

$$\begin{aligned} &= \sqrt{\frac{\sum X^2 P(X)}{\sum P(X)} - \left[\frac{\sum XP(X)}{\sum P(X)}\right]^2} \\ &= \sqrt{\sum X^2 P(X) - [\sum XP(X)]^2} \\ (\because \quad \sum P(X) &= 1) \end{aligned}$$

In the above example

No. of Petals	P(x)	xP(x)	x ² P(x)
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

$$\text{Hence S.D.}(X) = \sqrt{36.925 - (5.925)^2} = \sqrt{36.925 - 35.106} = \sqrt{1.819} = 1.3$$



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

$$\begin{aligned}
 C.V. &= \frac{\sigma}{\mu} \times 100 \\
 &= \frac{1.3}{5.925} \times 100 \\
 &= 21.9 \%
 \end{aligned}$$

Let us consider another example to understand the concept of discrete probability distribution.

Example

- Find the probability distribution of the sum of the dots when two fair dice are thrown
- 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

$$f(5) = \frac{4}{36}, f(6) = \frac{5}{36}, f(7) = \frac{6}{36}, f(8) = \frac{5}{36}, f(9) = \frac{4}{36},$$

$$f(10) = \frac{3}{36}, f(11) = \frac{2}{36} \text{ and } f(12) = \frac{1}{36}.$$

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:

$$\begin{aligned}
 & i) P(\text{a sum that is greater than } 8) \\
 & \quad = P(X > 8) \\
 & \quad = P(X=9) + P(X=10) + P(X=11) + P(X=12) \\
 & \quad = f(9) + f(10) + f(11) + f(12) \\
 & \quad = \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36}
 \end{aligned}$$

$$\begin{aligned}
 & ii) \quad P(\text{a sum that is greater than } 5 \text{ but less than or equal to } 10) \\
 & \quad = P(5 < X \leq 10) \\
 & \quad = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\
 & \quad = f(6) + f(7) + f(8) + f(9) + f(10) \\
 & \quad = \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} = \frac{23}{36}.
 \end{aligned}$$

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

$$F(x) = P(X \leq x).$$

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.

	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 $10/36$.

Topic 40

Graphical Representation 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 $F(x) = P(X \leq x)$. 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:

$$f(0) = P(X = 0)$$

$$= P[\{TTT\}] = 1/8$$

$$f(1) = P(X = 1)$$

$$= P[\{HTT, THT, TTH\}] = 3/8$$

$$f(2) = P(X = 2)$$

$$= P[\{HHT, HTH, THH\}] = 3/8$$

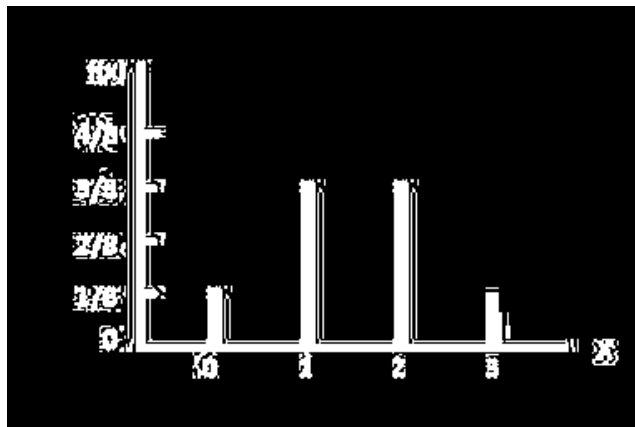
$$f(3) = P(X = 3)$$

$$= P[\{HHH\}] = 1/8$$

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

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{1}{8}, & \text{for } 0 \leq x < 1 \\ \frac{4}{8}, & \text{for } 1 \leq x < 2 \\ \frac{7}{8}, & \text{for } 2 \leq x < 3 \\ 1, & \text{for } x \geq 3 \end{cases}$$

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 $0 < x < 1$,

$$P(X < x) = P(X = 0) = \frac{1}{8};$$

For $1 < x < 2$, we have

$$\begin{aligned} P(X < x) &= P(X = 0) + P(X = 1) \\ &= \frac{1}{8} + \frac{3}{8} = \frac{4}{8}; \end{aligned}$$

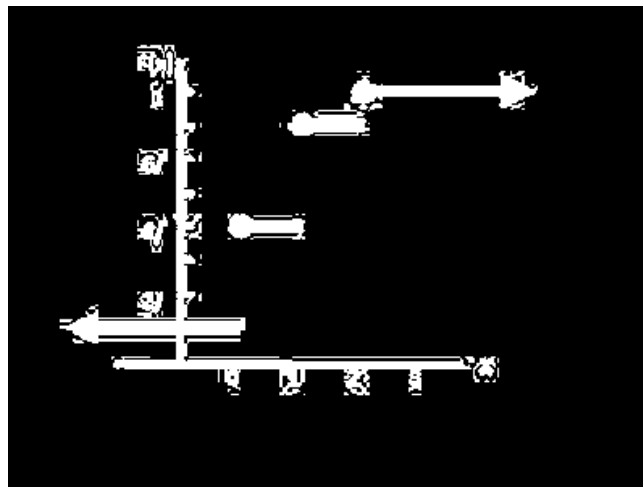
For $2 < x < 3$, we have

$$\begin{aligned} P(X < x) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}; \end{aligned}$$

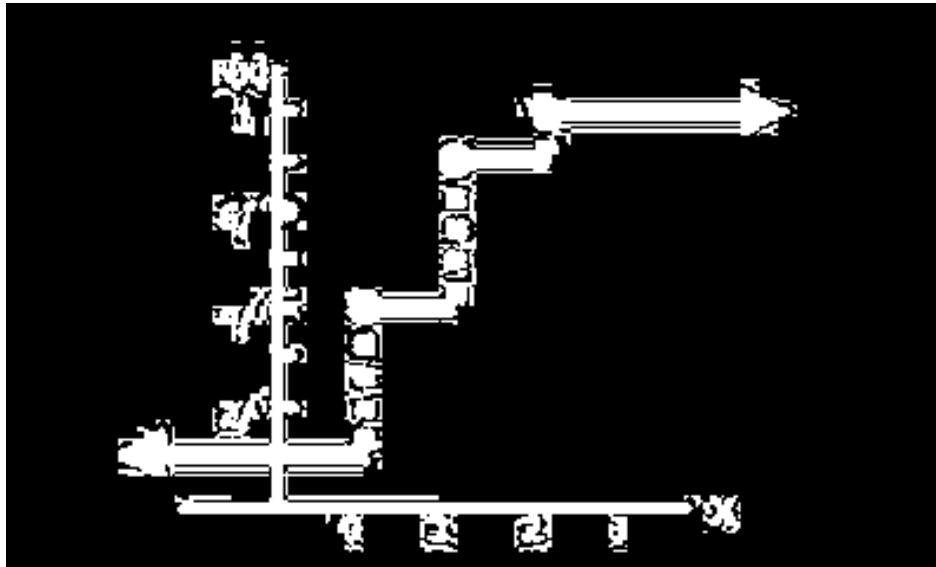
And, finally, for $x > 3$, we have

$$\begin{aligned} P(X < x) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{8}{8} = 1. \end{aligned}$$

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:

Topic 41

Properties of Expected Values, Covariance & Correlation & Discrete Probability Distributions

Properties of expected values, Covariance, Correlation:

- 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 $\sigma^2 > 0$, 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 $1/k^2$. 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 :

$$m = E(X) = \sum XP(X) = 5.925 = 5.9$$

And the standard deviation of this distribution is :

$$\begin{aligned} S.D.(X) &= \sqrt{36.925 - (5.925)^2} \\ &= \sqrt{36.925 - 35.106} \\ &= \sqrt{1.819} = 1.3 \end{aligned}$$

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

$$\mu - 2\sigma \text{ and } \mu + 2\sigma \quad \text{i.e. between}$$

$$5.9 - 2(1.3) \text{ and } 5.9 + 2(1.3) \text{ i.e.}$$

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

$$\sigma = \sqrt{E[(X - \mu)^2]}.$$

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,

$$\begin{aligned}\mu &= E(X) = \sum xP(x) \\ &= 0(.002) + 1(.029) + 2(.132) + 3(.309) + 4(.360) + 5(.168) \\ &= 3.50\end{aligned}$$

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-μ) ²	(x-μ) ² P(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

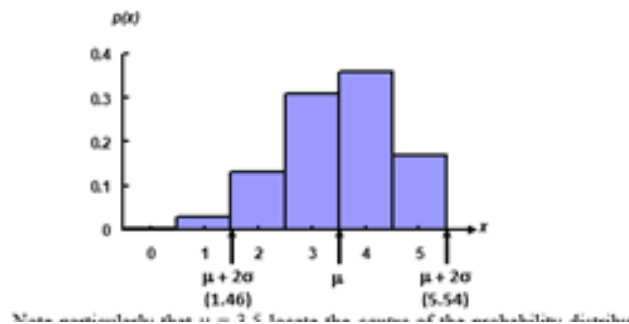
Thus, the variance is $\sigma^2 = 1.05$ and the standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} = 1.02$$

This value measures the spread of the probability distribution of X, the number of successful ventures out of five.

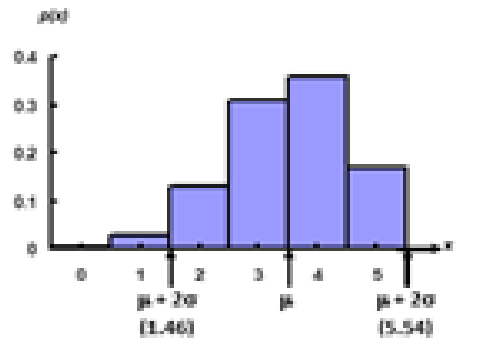
c) The graph of P(x) is shown in the following figure with the mean μ and the interval

$$\begin{aligned}\mu + 2\sigma &= 3.50 + 2(1.02) \\ &= 3.50 + 2.04 \\ &= (1.46, 5.54) \text{ shown on the graph.}\end{aligned}$$



Note particularly that $\mu = 3.5$ locates 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 $\mu + 2\sigma$ ----- 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 $\mu + 2\sigma$ 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, $P(x < 2)$, 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

Topic 42

Binomial Distribution, Introduction to the Hyper geometric Distribution

Discrete Uniform Distribution, Binomial Distribution:

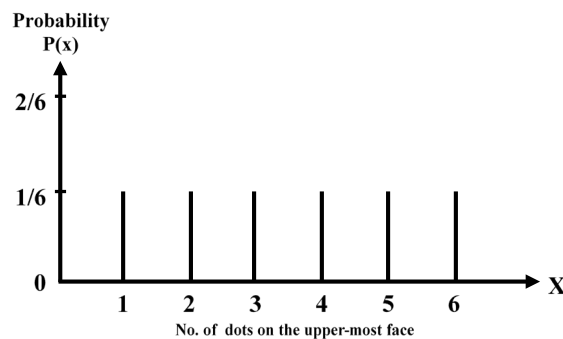
DISCRETE UNIFORM DISTRIBUTION

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

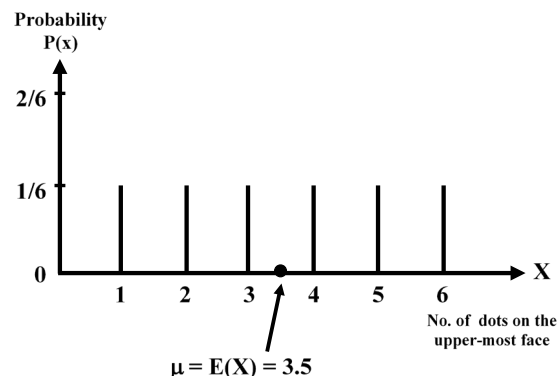
If we draw the line chart of this distribution, we obtain Line Chart Representation of the Discrete Uniform Probability Distribution



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.

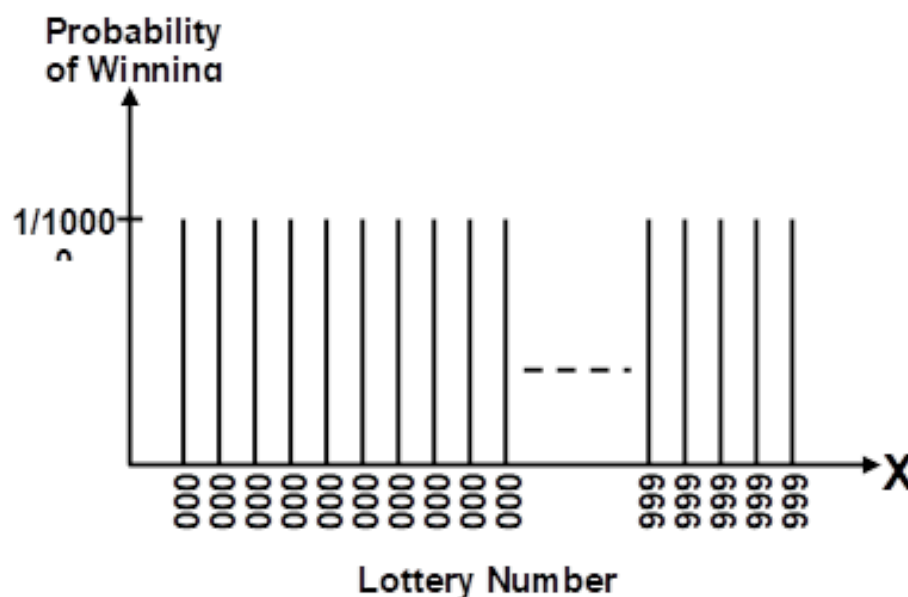
Line Chart Representation of the Discrete Uniform Probability Distribution



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:



Discrete Uniform Distribution

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 $0, 1, 2, \dots$ or, as in the above example, $0000, 0001 \dots, 9999$, 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 $\frac{1}{2}$ 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 represent 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 advance.

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 $\frac{1}{2}$ 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.

$x = 0, 1, 2, \dots, n$

The binomial distribution has two parameters, n and p . In this example, $n = 5$ since the coin was thrown 5 times, $p = \frac{1}{2}$ since it is a fair coin, $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$ Hence

$$P(X = x) = \binom{5}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}$$

Putting $x = 0$

$$P(X = 0) = \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{5-0}$$

$$= \frac{5!}{0!5!} (1) \left(\frac{1}{2}\right)^5$$

$$= 1(1) \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

Putting $x = 1$

$$P(X = 1) = \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{5-1}$$

$$P(X = 1) = \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{5-1}$$

$$= \frac{(5)}{1} \left(\frac{1}{2}\right)$$

$$= 5 \left(\frac{1}{2}\right)^5 = 5 \left(\frac{1}{32}\right) = \frac{5}{32}$$

$$P(X = 2) = \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{5-2} = \frac{10}{32}$$

$$P(X = 3) = \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{5-3} = \frac{10}{32}$$

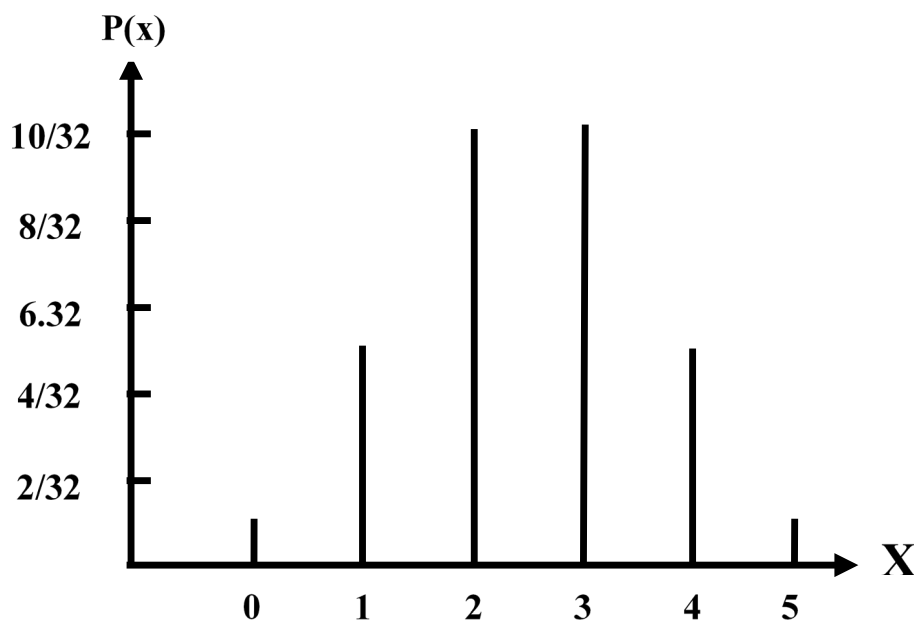
$$P(X = 4) = \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{5-4} = \frac{5}{32}$$

$$P(X = 5) = \binom{5}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{5-5} = \frac{1}{32}$$

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 $X = E(X) = \sum XP(X)$

$\text{Var}(X) = \sum X^2 P(X) - [\sum XP(X)]^2$

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 $\text{Var}(X) = npq$

$S.D.(X) = \sqrt{npq}$ so that

For the above example, $n = 5$, $p = \frac{1}{2}$ and $q = \frac{1}{2}$

Hence

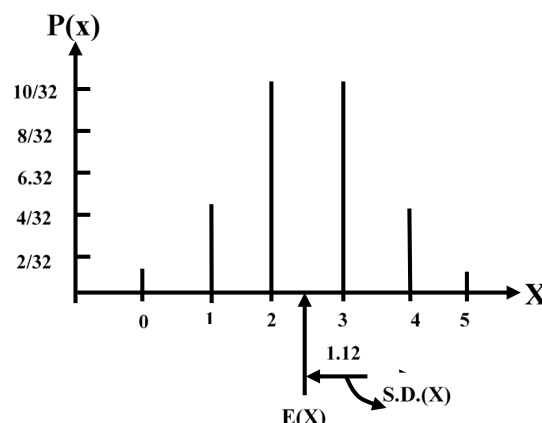
Mean = $E(X) = np = 5(\frac{1}{2}) = 2.5$

and $S.D.(X) = \sqrt{npq} = \sqrt{5(\frac{1}{2})(\frac{1}{2})} = \sqrt{\left(\frac{5}{4}\right)} = \underline{1.12}$

We would have got exactly the same answers if we had applied the LENGTHIER procedure.

$E(X) = \sum XP(X)$ and $\text{Var } X = \sum X^2 P(X) - [\sum XP(X)]^2$

Graphical Representation of the Mean and Standard Deviation of the Binomial Distribution ($n=5$, $p=1/2$)



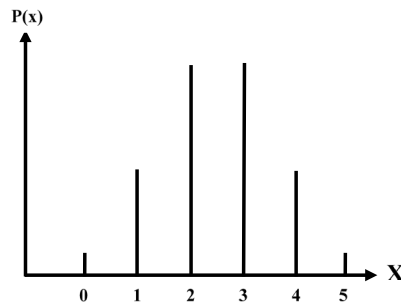
What does this mean?

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).

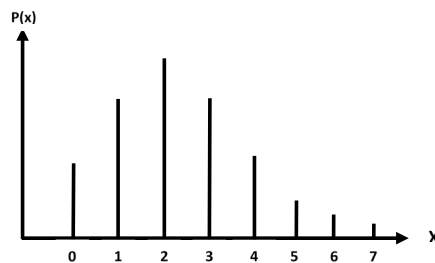
Coefficient of Variation:

$$C.V. = \frac{\sigma}{\mu} \times 100 = \frac{1.12}{2.5} \times 100 = 44.8\%$$

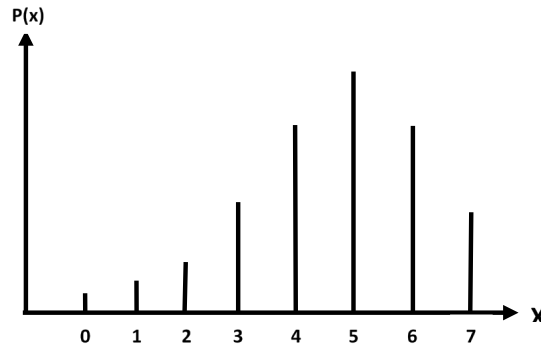
Note that the binomial distribution is not always symmetrical as in the above example. It will be symmetrical *only* when $p = q = \frac{1}{2}$ (as in the above example).



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 $\bar{X} = np$.

$$\begin{aligned}\bar{x} &= \frac{\sum f_i x_i}{\sum f_i} \\ &= \frac{0+56+148+117+72+5}{200} \text{ Now, we have:} \\ &= \frac{398}{200} = 1.99\end{aligned}$$

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:

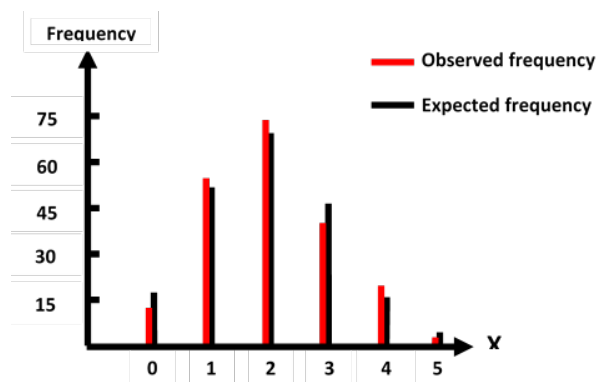
No. of Sixes (x)	Probability f(x)	Expected frequency
0	$\binom{5}{0} q^5 = (0.602)^5 = 0.07907$	15.8
1	$\binom{5}{1} q^4 p = 5 \cdot (0.602)^4 (0.398) = 0.26136$	52.5
2	$\binom{5}{2} q^3 p^2 = 10 \cdot (0.602)^3 (0.398)^2 = 0.34559$	69.1
3	$\binom{5}{3} q^2 p^3 = 10 \cdot (0.602)(0.398)^3 = 0.22847$	45.7
4	$\binom{5}{4} q p^4 = (0.602)(0.398)^4 = 0.07553$	15.1
5	$\binom{5}{5} p^5 = (0.398)^5 = .00998$	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_o	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:



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.

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).

Topic 43

Hypergeometric Distribution, Poisson Distribution & Continuous Uniform Distribution

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,

$$\begin{aligned} P(X = 2) &= \frac{\binom{5}{2} \binom{5}{4-2}}{\binom{10}{4}} \\ &= \frac{\binom{5}{2} \binom{5}{2}}{\binom{10}{4}} \quad \text{i.e. } P(X = 2) \text{ is} \\ &= \frac{10 \times 10}{210} \\ &= \frac{10}{21} \end{aligned}$$

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	Finite	Infinite
Sampling		
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.

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 $np = \mu$ 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 $\mu = np$ remains constant, then the limiting form of the binomial probability distribution is

$$\lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} b(x; n, p) = \frac{e^{-\mu} \mu^x}{x!}, \quad x = 0, 1, 2, \dots, \infty$$

where $e = 2.71828$.

The Poisson distribution has only one parameter $m > 0$.

The parameter m 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

$$\mu = np = (200)(0.01) = 2.$$

Therefore, if X represents the number of successes (not showing up), we have

$$\begin{aligned} P(X = 3) &= \frac{e^{-2}(2)^3}{3!} \\ &= \frac{(0.1353)(8)}{3 \times 2 \times 1} = 0.1804 \\ \left(\because e^{-2} = \frac{1}{(2.71828)^2} = 0.1353 \right) \end{aligned}$$

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:

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 $\text{Var}(X) = \mu$.
- (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 $p < 0.05$.

Topic 44

Appropriate Use of Measures of Location

Measures of location

Measures of location summarize a list of numbers by a "typical" value.

The three most common measures of location are the mean, the median, and the mode.

- **Mean:** The average value in a dataset. The mean is the sum of the values, divided by the number of values. It has the smallest possible sum of squared differences from members of the list.
- **Median:** The median is the middle value in the sorted list. It is the smallest number that is at least as big as at least half the values in the list. It has the smallest possible sum of absolute differences from members of the list.
- **Mode:** The most frequently occurring value(s) in a dataset (or one of the most frequent values, if there are more than one). It differs from the fewest possible members of the list.

Real Life Examples: Using Mean, Median, & Mode

Individuals and companies use these metrics all the time in different fields to gain a better understanding of datasets.

The following examples explain how the mean, median, and mode are used in different real life scenarios.

Example 1: Mean, Median, & Mode in Healthcare

The mean, median, and mode are widely used by insurance analysts and actuaries in the healthcare industry.

For example:

- **Mean:** Insurance analysts often calculate the mean age of the individuals they provide insurance for so they can know the average age of their customers.
- **Median:** Actuaries often calculate the median amount spend on healthcare each year by individuals so they can know how much insurance they need to be able to provide to individuals.
- **Mode:** Actuaries also calculate the mode of their customers (the most commonly occurring age) so they can know which age group uses their insurance the most.

Example 2: Mean, Median, & Mode in Real Estate

The mean, median, and mode are also used often by real estate agents.

For example:

- **Mean:** Real estate agents calculate the mean price of houses in a particular area so they can inform their clients of what they can expect to spend on a house.
- **Median:** Real estate agents also calculate the median price of houses to gain a better idea of the “typical” home price, since the median is less influenced by outliers (like multi-million dollar homes) compared to the mean.
- **Mode:** Real estate agents also calculate the mode of the number of bedrooms per house so they can inform their clients on how many bedrooms they can expect to have in houses in a particular area.

Example 3: Mean, Median, & Mode in Human Resources

The mean, median, and mode are often used by individuals who work in Human Resource departments at companies.

For example:

- **Mean:** Human Resource managers often calculate the mean salary of individuals in a certain field so that they can know what type of “average” salary to offer to new employees.
- **Median:** Human Resource managers also often calculate the median salary in certain fields so that they can be informed of what the typical “middle” salary is for a particular field.
- **Mode:** Human Resource managers also calculate the mode of different positions in the company so that they can be aware of the most common position of employees at their company.

Topic 45

Quantiles (Cut Points)

Quantiles

The word “quantile” comes from the word quantity. In simple terms, a quantile is where a sample is divided into equal-sized, adjacent, subgroups (that’s why it’s sometimes called a “fractile”).

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)

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

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.

Topic 46

Examples of Quantiles

Find the quartiles for the following frequency distribution.

Class Boundaries	Frequency
145 – 150	4
150 – 155	6
155 – 160	28
160 – 165	58
165 – 170	64
170 – 175	30

Solution:

Class Boundaries	Frequency	Cumulative Frequency
145 – 150	4	4
150 – 155	6	10
155 – 160	28	38
160 – 165	58	96
165 – 170	64	160
170 – 175	30	190

f = 58

$$Q1 = l + \frac{h}{f} \left(\frac{n}{4} - C \right)$$

$$n = 190, \frac{n}{4} = \frac{190}{4} = 47.5, C = 38, l = 160, h = 5, f = 58$$

$$Q1 = 160 + \frac{5}{58} (47.5 - 38)$$

$$= 160 + \frac{47.5}{58}$$

$$= 160 + 0.81896 \quad \mathbf{Q1 = 160.819.}$$

Class Boundaries	Frequency	Cumulative Frequency
145 – 150	4	4
150 – 155	6	10
155 – 160	28	38
160 – 165	58	96
165 – 170	64	160
170 – 175	30	190

f = 58

$$Q2 = l + \frac{h}{f} \left(\frac{2n}{4} - C \right) = l + \frac{h}{f} \left(\frac{n}{2} - C \right)$$

$$n = 190, \frac{n}{2} = \frac{190}{2} = 95, C = 38, l = 160, h = 5, f = 58$$

$$Q2 = 160 + \frac{5}{58} (95 - 38)$$

$$= 160 + \frac{285}{58}$$

$$= 160 + 4.913793 \quad \mathbf{Q2 = 164.9138}$$

Class Boundaries	Frequency	Cumulative Frequency
145 – 150	4	4
150 – 155	6	10
155 – 160	28	38
160 – 165	58	96
165 – 170	64	160
170 – 175	30	190

f = 58

$$Q3 = l + \frac{h}{f} \left(\frac{3n}{4} - C \right)$$

$$n = 190, \frac{3n}{4} = \frac{3(190)}{4} = \frac{570}{4} = 142.5, C = 96, l = 165, h = 5, f = 64$$

$$Q3 = 165 + \frac{5}{64} (142.5 - 96)$$

$$= 165 + \frac{232.5}{64}$$

$$= 165 + 3.633$$

$$= 168.633.$$

Quartiles:

If we want to consider the raw data then our first step will be to arrange the data in ascending order of magnitude, then:

- The median is the middle value of the data set.



- The **lower quartile** (Q_1) is the first quartile of the data set.



- The **upper quartile** (Q_3) is the third quartile of the data set.



Example

Find the median, lower quartile and upper quartile of the following data set of scores:

19 21 24 21 24 28 25 24 30

Solution:

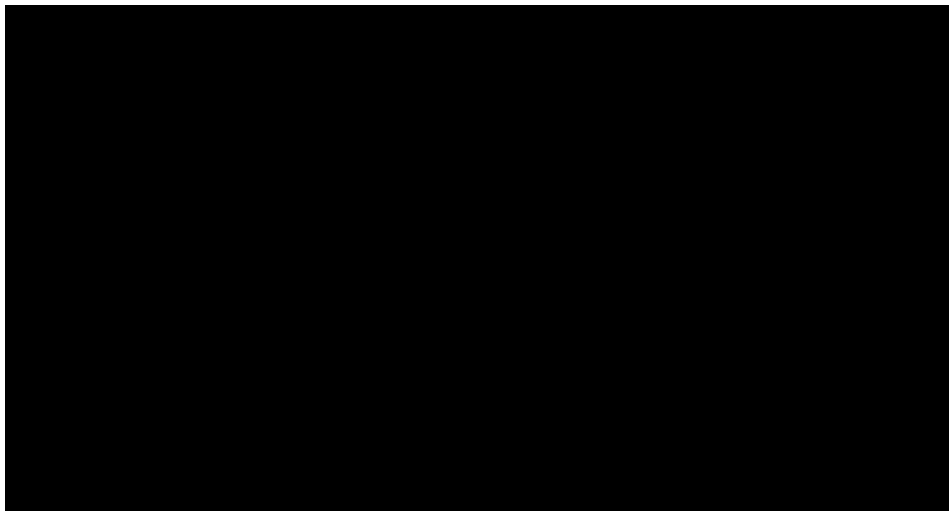
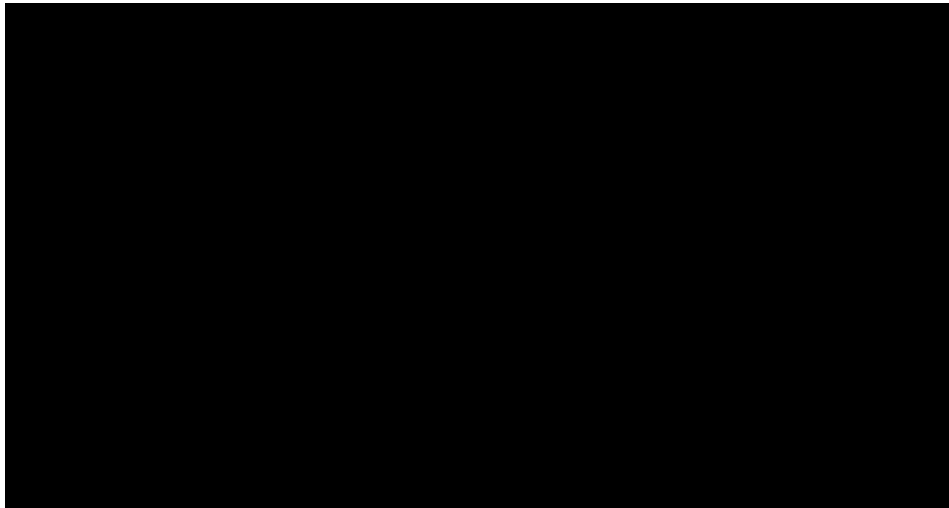
Arrange the score values in ascending order of magnitude:

19 21 21 24 24 24 25 28 30

There are 9 values in the data set.

$\therefore n = 9$

$$\begin{aligned}\text{Now, median} &= \left(\frac{n+1}{2} \right) \text{th value} \\ &= \left(\frac{9+1}{2} \right) \text{th value} \\ &= \frac{10}{2} \text{th value} \\ &= 5\text{th value} \\ &= 24\end{aligned}$$



Topic 47

Measures of Dispersion

What Is Dispersion?

Dispersion refers to the ‘distribution’ of objects over a large region. The degree to which numerical data are dispersed or squished around an average value is referred to as dispersion in statistics. It is, in a nutshell, the dispersion of data. A vast amount of data will always be widely dispersed or firmly packed. Data that is widely dispersed – 0, 30, 60, 90, 120, With tiny data grouped densely – 1, 2, 2, 3, 3, 4, 4....

Understanding Dispersion

The term “dispersion” refers to how dispersed a set of data is. The measure of dispersion is always a non-negative real number that starts at zero when all the data is the same and rises as the data gets more varied. The homogeneity or heterogeneity of the scattered data is defined by dispersion measures. It also refers to how data differs from one another.

Measures of Dispersion

As the name suggests, the measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

Suppose you have four datasets of the same size and the mean is also the same, say, m . In all the cases the sum of the observations will be the same. Here, the measure of central tendency is not giving a clear and complete idea about the distribution for the four given sets.

Can we get an idea about the distribution if we get to know about the dispersion of the observations from one another within and between the datasets? The main idea about the measure of dispersion is to get to know how the data are spread. It shows how much the data vary from their average value.

Characteristics of Measures of Dispersion

- A measure of dispersion should be rigidly defined
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations
- Based on all observations

Classification of Measures of Dispersion

The measure of dispersion is categorized as:

(i) An absolute measure of dispersion:

- The measures express the scattering of observation in terms of distances i.e., range, quartile deviation.
- The measure expresses the variations in terms of the average of deviations of observations like mean deviation and standard deviation.

(ii) A relative measure of dispersion:

We use a relative measure of dispersion for comparing distributions of two or more data set and for unit free comparison. They are the coefficient of range, the coefficient of mean deviation, the coefficient of quartile deviation, the coefficient of variation, and the coefficient of standard deviation.

Topic 48

Absolute Measures of Dispersion

Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. The absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3, 5, 6, 7 $\Rightarrow$ Range = $7 - 1 = 6$
2. **Variance:** Deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total no of values in the data set to get the variance.
$$\text{Variance } (\sigma^2) = \sum (X - \mu)^2 / N$$
3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.
4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Topic 49

Examples for Absolute Measures of Dispersion

Range

Range refers to the difference between each series' minimum and maximum values. The range offers us a good indication of how dispersed the data is, but we need other measures of variability to discover the dispersion of data from central tendency measurements. A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If $X_{\max}$ and $X_{\min}$ are the two extreme observations then

$$\text{Range} = X_{\max} - X_{\min}$$

Merits of Range

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

Example:

In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9.

So the range is $9 - 3 = 6$.

Variance

Variance: Deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total no of values in the data set to get the variance.

$$\sigma^2 = \frac{\sum_{i=1}^N (X - \mu)^2}{N}$$

Find the variance and standard deviation of the data.

X	$x - \bar{X}$	$(X - \bar{X})^2$
5	5-12.4=-7.4	$(-7.4)^2 = 54.76$
9	9-12.4=-3.4	$(-3.4)^2 = 11.56$
13	13-12.4=0.6	$(0.6)^2 = 0.36$
15	15-12.4=2.6	$(2.6)^2 = 6.76$
20	20-12.4=7.6	$(7.6)^2 = 57.76$

1st we need to find the mean $\bar{X}$ of the data as

$$\bar{X} = \frac{5+9+13+15+20}{5} = \frac{62}{5} = 12.4$$

Now the variance denoted by σ^2 is calculated by

$$\sigma^2 = \frac{\sum_{i=1}^5 (X_i - \bar{X})^2}{5} = \frac{54.76+11.56+0.36+6.76+57.76}{5} = \frac{131.2}{5} = 26.24$$

Standard Deviation

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ . It is also referred to as root mean square deviation. The standard deviation is given as

$$S = \sigma = \sqrt{\frac{\sum_{i=1}^N (X - \mu)^2}{N}}$$

Find the standard deviation of the data.

X	$x - \bar{X}$	$(X - \bar{X})^2$
5	$5 - 12.4 = -7.4$	$(-7.4)^2 = 54.76$
9	$9 - 12.4 = -3.4$	$(-3.4)^2 = 11.56$
13	$13 - 12.4 = 0.6$	$(0.6)^2 = 0.36$
15	$15 - 12.4 = 2.6$	$(2.6)^2 = 6.76$
20	$20 - 12.4 = 7.6$	$(7.6)^2 = 57.76$

1st we need to find the mean $\bar{X}$ of the data as

$$\bar{X} = \frac{5 + 9 + 13 + 15 + 20}{5} = \frac{62}{5} = 12.4$$

Now the variance denoted by σ^2 is calculated by

$$\sigma^2 = \frac{\sum_{i=1}^5 (X_i - \bar{X})^2}{5} = \frac{54.76 + 11.56 + 0.36 + 6.76 + 57.76}{5} = \frac{131.2}{5} = 26.24$$

Positive square root of variance is known as standard deviation. Which is?

$$\sigma = \sqrt{\sigma^2} = \sqrt{26.24} = 5.1225$$

Merits of Standard Deviation

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations
- Suitable for further mathematical treatment
- Least affected by the fluctuation of the observations
- The standard deviation is zero if all the observations are constant
- Independent of change of origin

Demerits of Standard Deviation

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale

Examples of Quantiles

Find the quartiles for the following frequency distribution.

Class Boundaries	Frequency
145 – 150	4
150 – 155	6
155 – 160	28
160 – 165	58
165 – 170	64
170 – 175	30

Solution:

Class Boundaries	Frequency	Cumulative Frequency
145 – 150	4	4
150 – 155	6	10
155 – 160	28	38
160 – 165	58	96
165 – 170	64	160
170 – 175	30	190

f = 58

$$Q1 = l + \frac{h}{f} \left(\frac{n}{4} - C \right)$$

$$n = 190, \frac{n}{4} = \frac{190}{4} = 47.5, C = 38, l = 160, h = 5, f = 58$$

$$Q1 = 160 + \frac{5}{58} (47.5 - 38)$$

$$= 160 + \frac{47.5}{58}$$

$$= 160 + 0.81896 \quad Q1 = 160.819.$$

Class Boundaries	Frequency	Cumulative Frequency
145 – 150	4	4
150 – 155	6	10
155 – 160	28	38
160 – 165	58	96
165 – 170	64	160
170 – 175	30	190

f = 58

$$Q2 = l + \frac{h}{f} \left(\frac{2n}{4} - C \right) = l + \frac{h}{f} \left(\frac{n}{2} - C \right)$$

$$n = 190, \frac{n}{2} = \frac{190}{2} = 95, C = 38, l = 160, h = 5, f = 58$$

$$Q2 = 160 + \frac{5}{58} (95 - 38)$$

$$= 160 + \frac{285}{58}$$

$$= 160 + 4.913793 \quad \mathbf{Q2 = 164.9138}$$

Class Boundaries	Frequency	Cumulative Frequency
145 – 150	4	4
150 – 155	6	10
155 – 160	28	38
160 – 165	58	96
165 – 170	64	160
170 – 175	30	190

f = 58

$$Q3 = l + \frac{h}{f} \left(\frac{3n}{4} - C \right)$$

$$n = 190, \frac{3n}{4} = \frac{3(190)}{4} = \frac{570}{4} = 142.5, C = 96, l = 165, h = 5, f = 64$$

$$Q3 = 165 + \frac{5}{64} (142.5 - 96)$$

$$= 165 + \frac{232.5}{64}$$

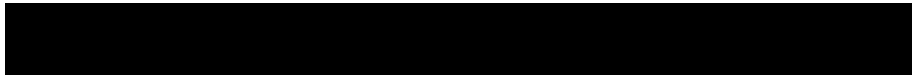
$$= 165 + 3.633$$

$$= 168.633.$$

- The median is the middle value of the data set.



- The **lower quartile** (Q_1) is the first quartile of the data set.



- The **upper quartile** (Q_3) is the third quartile of the data set.



Example

Find the median, lower quartile and upper quartile of the following data set of scores:

19 21 24 21 24 28 25 24 30

Solution:

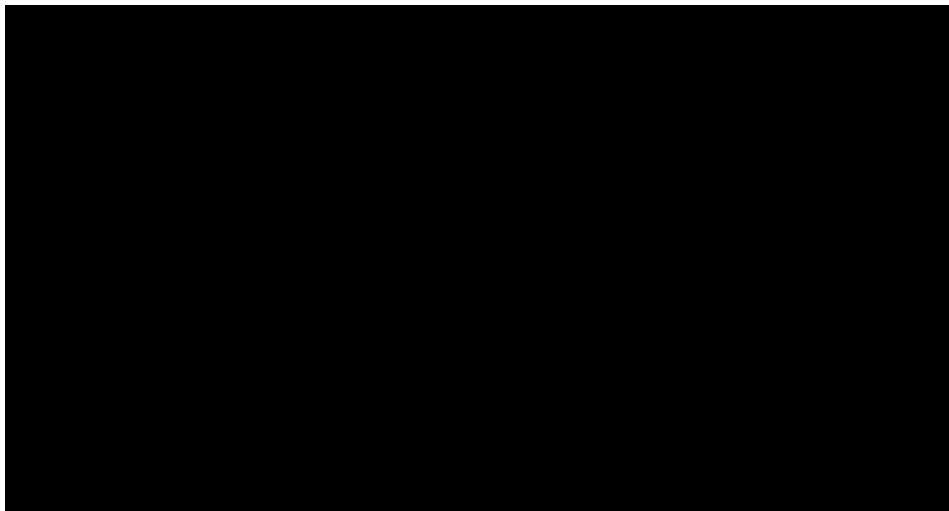
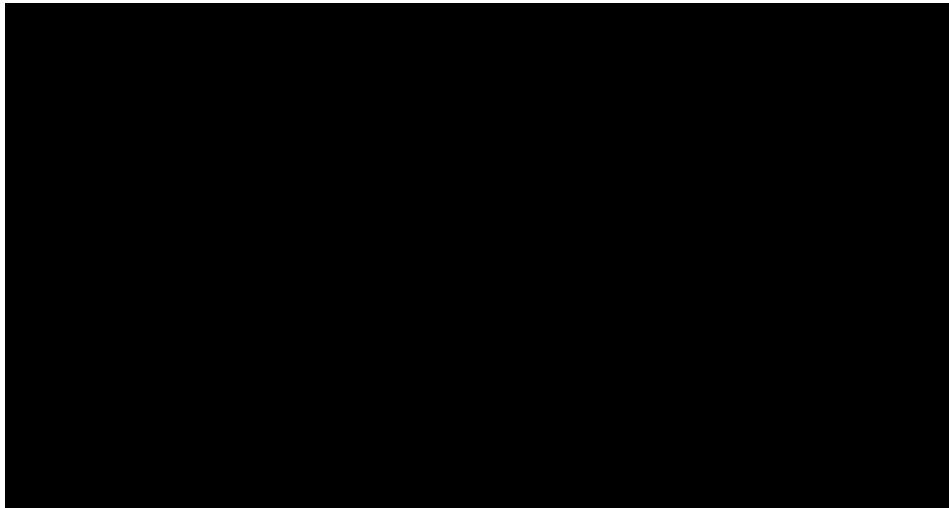
Arrange the score values in ascending order of magnitude:

19 21 21 24 24 24 25 28 30

There are 9 values in the data set.

$\therefore n = 9$

$$\begin{aligned}\text{Now, median} &= \left(\frac{n+1}{2} \right) \text{th value} \\ &= \left(\frac{9+1}{2} \right) \text{th value} \\ &= \frac{10}{2} \text{th value} \\ &= 5\text{th value} \\ &= 24\end{aligned}$$



Quartile Deviation

The quartiles divide a data set into quarters. The first quartile, (Q_1) is the middle number between the smallest number and the median of the data. The second quartile, (Q_2) is the median of the data set. The third quartile, (Q_3) is the middle number between the median and the largest number.

Quartile deviation or semi-inter-quartile deviation is

$$Q = \frac{1}{2} \times (Q_3 - Q_1)$$

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores 50% of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

EXAMPLE

The shareholding structure of two companies is given below:

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

The quartile deviation for company X is

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

For company Y, it is

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

Mean Deviation

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \dots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is

$$M.D. = \frac{\sum |x_i - \bar{x}|}{n}, \quad \text{for ungrouped data}$$

$$M.D. = \frac{\sum f |x_i - \mu|}{\sum f}, \quad \text{for grouped data}$$

Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

Merits of Mean Deviation

- Based on all observations
- It provides a minimum value when the deviations are taken from the median
- Independent of change of origin

Demerits of Mean Deviation

- Not easily understandable
- Its calculation is not easy and time-consuming
- Dependent on the change of scale
- Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

Example:

The Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

Step 1: Find the mean:

$$\text{Mean} = 3 + 6 + 6 + 7 + 8 + 11 + 15 + 16 \div 8 = 9$$

Step 2: Find the distance of each value from that mean:

Value	Distance from 9
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7

Step 3. Find the mean of those distances:

$$\text{Mean Deviation} = 6 + 3 + 3 + 2 + 1 + 2 + 6 + 7 \div 8 = 3.75$$

So, the mean = 9, and the mean deviation = 3.75

Example 1: Find the mean deviation about the median for the following data.

x	f
12	7
9	3
6	8
18	1
10	2

Solution:

x	f	x.f	$\frac{ x - \bar{x} }{n}$	$f \cdot \frac{ x - \bar{x} }{n}$
12	7	84	2.619	18.33
9	3	27	0.381	1.143
6	8	48	3.381	27.048
18	1	18	8.619	8.619
10	2	20	0.619	1.238
Total	21	197		56.378

Mean is given by $M.D. = \frac{\sum f |x_i - \mu|}{\sum f}$

Substituting values in the equation

Answer: The mean deviation about mean is 2.684

Topic 50

Relative Measures of Dispersion

Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

1. Co-efficient of Range
2. Co-efficient of Variation
3. Co-efficient of Standard Deviation
4. Co-efficient of Quartile Deviation
5. Co-efficient of Mean Deviation

Co-efficient of Dispersion

The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages. The dispersion coefficient is also used when two series with different measurement units are compared. It is denoted as C.D.

The common coefficients of dispersion are:

C.D. in terms of	Coefficient of dispersion
Range	$C.D. = (X_{\max} - X_{\min}) / (X_{\max} + X_{\min})$
Quartile Deviation	$C.D. = (Q_3 - Q_1) / (Q_3 + Q_1)$
Standard Deviation (S.D.)	$C.D. = S.D. / \text{Mean}$
Mean Deviation	$C.D. = \text{Mean deviation} / \text{Average}$

Topic 51

Examples of Relative Measures of Dispersion

Co-efficient of Range

Example:

The following are the wages of 8 workers in a factory. Find the range and coefficient of range.

Wages are in dollars: 1400, 1450, 1520, 1380, 1485, 1495, 1575, 1440.

$$\text{range} = x_m - x_0$$

Solution: $\text{range} = 1575 - 1380$
 $\text{range} = 195$

$$\text{coefficient of range} = \frac{x_m - x_0}{x_m + x_0}$$

$$\text{coefficient of range} = \frac{1575 - 1380}{1575 + 1380}$$

$$\text{coefficient of range} = 0.066$$

Co-efficient of Variation

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.

Co-efficient of Quartile Deviation

The Quartile Deviation(QD) is the product of half of the difference between the upper and lower quartiles. Mathematically we can define as:

$$\text{Quartile Deviation} = (Q_3 - Q_1) / 2$$

Quartile Deviation defines the absolute measure of dispersion. Whereas the relative measure corresponding to QD, is known as the coefficient of QD, which is obtained by applying the certain set of the formula:

$$\text{Coefficient of Quartile Deviation} = (Q_3 - Q_1) / (Q_3 + Q_1)$$

A Coefficient of QD is used to study & compare the degree of variation in different situations.

Example

Find the coefficient of Quartile Deviation for the given data Q1=13.75 and Q3=22.

$$\text{coefficient of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$\text{coefficient of } Q.D = \frac{22 - 13.75}{22 + 13.75}$$

$$\text{coefficient of } Q.D = \frac{8.25}{35.75}$$

$$\text{coefficient of } Q.D = 0.23$$

Co-efficient of Mean Deviation

Example

Find the coefficient of Mean Deviation for the given data Mean=9.3 and M.D=2.684.

$$\text{coefficient of } M.D = \frac{M.D}{\text{Mean}}$$

$$\text{coefficient of } M.D = \frac{2.684}{9.3}$$

$$\text{coefficient of } M.D = 0.288$$

Topic 52

Appropriate usage of the Measures of Dispersion

Why is it important to measure the spread of data?

There are many reasons why the measure of the spread of data values is important, but one of the main reasons regards its relationship with measures of central tendency. A measure of spread gives us an idea of how well the mean, for example, represents the data. If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small. This is because a large spread indicates that there are probably large differences between individual scores. Additionally, in research, it is often seen as positive if there is little variation in each data group as it indicates that the similar.

What are the objectives of computing dispersion?

(1) Comparative study

- Measures of dispersion give a single value indicating the degree of consistency or uniformity of distribution. This single value helps us in making comparisons of various distributions.
- The smaller the magnitude (value) of dispersion, higher is the consistency or uniformity and vice-versa.

(2) Reliability of an average

- A small value of dispersion means low variation between observations and average. It means that the average is a good representative of observation and very reliable.
- A higher value of dispersion means greater deviation among the observations. In this case, the average is not a good representative and it cannot be considered reliable.

(3) Control the variability

- Different measures of dispersion provide us data of variability from different angles, and this knowledge can prove helpful in controlling the variation.
- Especially in the financial analysis of business and medicine, these measures of dispersion can prove very useful.

(4) Basis for further statistical analysis

- Measures of dispersion provide the basis for further statistical analysis like computing correlation, regression, test of hypothesis, etc.

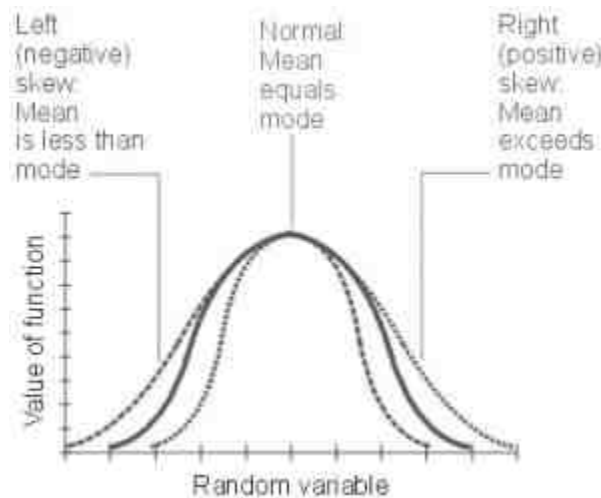
Topic 53

Calculating Skewness and making decision using Numerical measures

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or average, and the mode, or maximum point on the curve, are equal.

- In a perfect normal distribution (green solid curve in the illustration below), the tails on either side of the curve are exact mirror images of each other.
- When a distribution is skewed to the left (red dashed curve), the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative Skewness.
- When a distribution is skewed to the right (blue dotted curve), the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive Skewness.



Coefficient of Skewness

The coefficient of Skewness measures the Skewness of a distribution. It is based on the notion of the moment of the distribution. This coefficient is one of the measures of Skewness.

Pearson's coefficient of Skewness is calculated as:

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

Pearson's First Coefficient

The median is always the middle value, and the mean and mode are the extremes, so you can derive a formula to capture the horizontal distance between mean and mode.

$$\text{Pearson's First Coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

The above formula gives you Pearson's first coefficient. Division by the standard deviation will help you scale down the difference between mode and mean. This will scale down their values in a range of -1 to 1. Now understand the below relationship between mode, mean and median.

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

Substituting this in Pearson's first coefficient gives us Pearson's second coefficient and the formula for skewness:

$$\text{Pearson's Second Coefficient} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

If this value is between:

1. -0.5 and 0.5, the distribution of the value is almost symmetrical
2. -1 and -0.5, the data is negatively skewed, and if it is between 0.5 to 1, the data is positively skewed. The skewness is moderate.
3. If the skewness is lower than -1 (negatively skewed) or greater than 1 (positively skewed), the data is highly skewed.

How to Interpret

- Skewness also includes the extremes of the dataset instead of focusing only on the average. Hence, investors take note of Skewness while estimating the distribution of returns on investments. The average of the data set works out if an investor holds a position for the long term. Therefore, extremes need to be looked at when investors seek short-term and medium-term security positions.
- Usually, a standard deviation is used by investors in forecasting returns, and it presumes a normal distribution with zero Skewness. However, because of Skewness risk, it is better to obtain the performance estimations based on Skewness. Moreover, the occurrence of return distributions coming close to normal is low.
- Skewness risk occurs when a symmetric distribution is applied to the skewed data. The financial models seeking to estimate an asset's future performance consider a normal distribution. However, skewed data will increase the accuracy of the financial model.
- If a return distribution shows a positive skew, investors can expect recurrent small losses and few large returns from investment. Conversely, a negatively skewed distribution implies many small wins and a few large losses on the investment.
- Hence, a positively skewed investment return distribution should be preferred over a negatively skewed return distribution since the huge gains may cover the frequent – but small – losses. However, investors may prefer investments with a negatively skewed return distribution. It may be because they prefer frequent small wins and a few huge losses over frequent small losses and a few large gains.

Example

Calculate Skewness from the following grouped data:

X	Frequency
0	1
1	5
2	10
3	6
4	3

Skewness:

$$\bar{X} = \frac{\sum fx}{\sum f}$$

$$\bar{X} = \frac{55}{25}$$

$$\bar{X} = 2.2$$

x	f	$f \cdot x$	$(x - \bar{x})$	$f \cdot (x - \bar{x})^2$	$f \cdot (x - \bar{x})^3$
-2	-3	(4)=(2)×(3)	-5	(6)=(3)×(5)	(7)=(5)×(6)
0	1	0	-2.2	4.84	-10.65
1	5	5	-1.2	7.2	-8.64
2	10	20	-0.2	0.4	-0.08
3	6	18	0.8	3.84	3.072
4	3	12	1.8	9.72	17.496
---	---	---	---	---	---
--	$n=25$	$\sum f \cdot x=55$	--	26	1.2

Standard deviation

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

$$S = \sqrt{\frac{26}{24}}$$

$$S = 1.0408$$

Skewness

$$Sk = \frac{\sum (x - \bar{x})^3}{n - 1 \cdot S^3}$$

$$Sk = \frac{12}{24(1.04080)^3}$$

$$Sk = 0.0443$$

Topic 54

Calculating Kurtosis and making decision using Numerical measures

Kurtosis

A statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution

What is Kurtosis?

Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.



Along with skewness, kurtosis is an important descriptive statistic of data distribution. However, the two concepts must not be confused with each other. Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.

In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high risk for an investment because it indicates high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

What is Excess Kurtosis?

Excess kurtosis is a metric that compares the kurtosis of a distribution against the kurtosis of a normal distribution. The kurtosis of a normal distribution equals 3. Therefore, the excess kurtosis is found using the formula below:

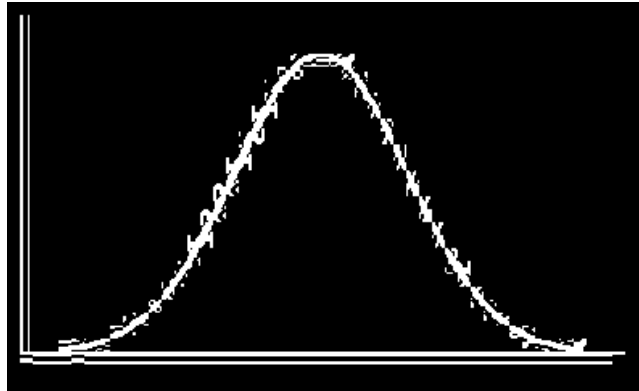
$$\text{Excess Kurtosis} = \text{Kurtosis} - 3$$

Types of Kurtosis

The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values, as well as values close to zero.

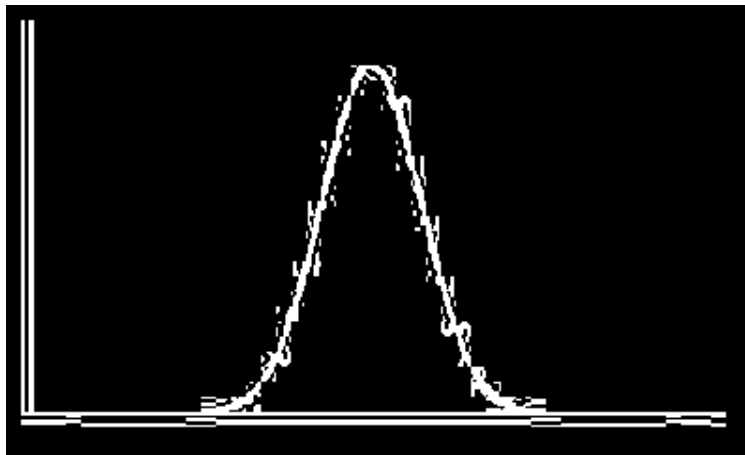
1. Mesokurtic

Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. This means that if the data follows a normal distribution, it follows a mesokurtic distribution.



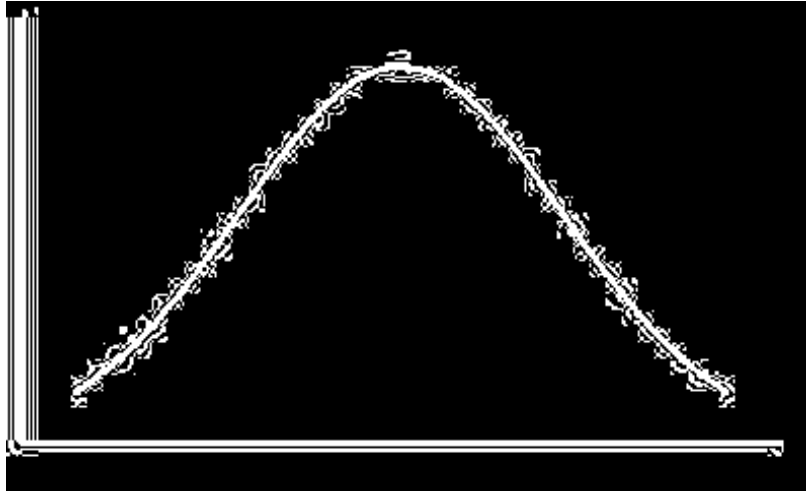
2. Leptokurtic

Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.



3. Platykurtic

A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.



Example

Calculate Kurtosis from the following grouped data:

X	Frequency
0	1
1	5
2	10
3	6
4	3

Kurtosis:

$$\bar{X} = \frac{\sum fx}{\sum f}$$

$$\bar{X} = \frac{55}{25}$$

$$\bar{X} = 2.2$$

x	f	$f \cdot x$	$(x - \bar{x})$	$f \cdot (x - \bar{x})^2$	$f \cdot (x - \bar{x})^3$	$f \cdot (x - \bar{x})^4$
-2	-3	(4)=(2)×(3)	-5	(6)=(3)×(5)	(7)=(5)×(6)	(8)=(5)×(7)
0	1	0	-2.2	4.84	-10.65	23.426
1	5	5	-1.2	7.2	-8.64	10.368
2	10	20	-0.2	0.4	-0.08	0.016
3	6	18	0.8	3.84	3.072	2.4576
4	3	12	1.8	9.72	17.496	31.493
---	---	---	---	---	---	---
--	$n=25$	$\sum f \cdot x=55$	--	26	1.2	67.76

Standard deviation

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

$$S = \sqrt{\frac{26}{24}}$$

$$S = 1.0408$$

Kurtosis

$$Kurtosis = \frac{\sum (x - \bar{x})^4}{n - 1 \cdot S^4}$$

$$Kurtosis = \frac{67.76}{24(1.04080)^4}$$

$$Kurtosis = 2.4057$$

Topic 55

Counting Rules

When selecting elements of a set, the number of possible outcomes depends on the conditions under which the selection has taken place. There are at least 4 rules to count the number of possible outcomes:

Multiplicative rule

Suppose you have j sets of elements, n_1 in the first set, n_2 in the second set, ... and n_j in the j th set. Suppose you wish to form a sample of j elements by taking one element from each of the j sets. The number of possible sets is then defined by

$$n_1 \cdot n_2 \cdot \dots \cdot n_j.$$

Permutation rule

The arrangement of elements in a distinct order is called permutation. Given a single set of n distinctively different elements, you wish to select k elements from the n and arrange them within k positions. The number of different permutations of the n elements taken k at a time is denoted P_k^n and is equal to

$$P_k^n = \frac{n!}{(n-k)!}.$$

Partitions rule

Suppose a single set of n distinctively different elements exists. You wish to partition them into k sets, with the first set containing n_1 elements, the second containing n_2 elements, ..., and the k^{th} set containing n_k elements. The number of different partitions is

$$\frac{n!}{n_1! n_2! \dots n_k!},$$

where $n_1 + n_2 + \dots + n_k = n$.

The numerator gives the permutations of the n elements. The terms in the denominator remove the duplicates due to the same assignments in the k sets (multinomial coefficients).

Combinations rule

A sample of k elements is to be chosen from a set of n elements. The number of different samples of k samples that can be selected from n is equal to

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The combination rule is a special application of the partition rule, with $j=2$ and $n_1=k$. From $n=n_1+n_2$ it follows that n_2 can be replaced by $(n-n_1)$. Usually the two groups refer to the two different groups of selected and non-selected samples. The order in which the n_1 elements are drawn is not important, therefore there are fewer combinations than permutations (binomial theorem).

Topic 56

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

$$\begin{aligned} {}^nP_r &= n(n-1)(n-2) \dots (n-r+1) \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

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 $0! = 1$.

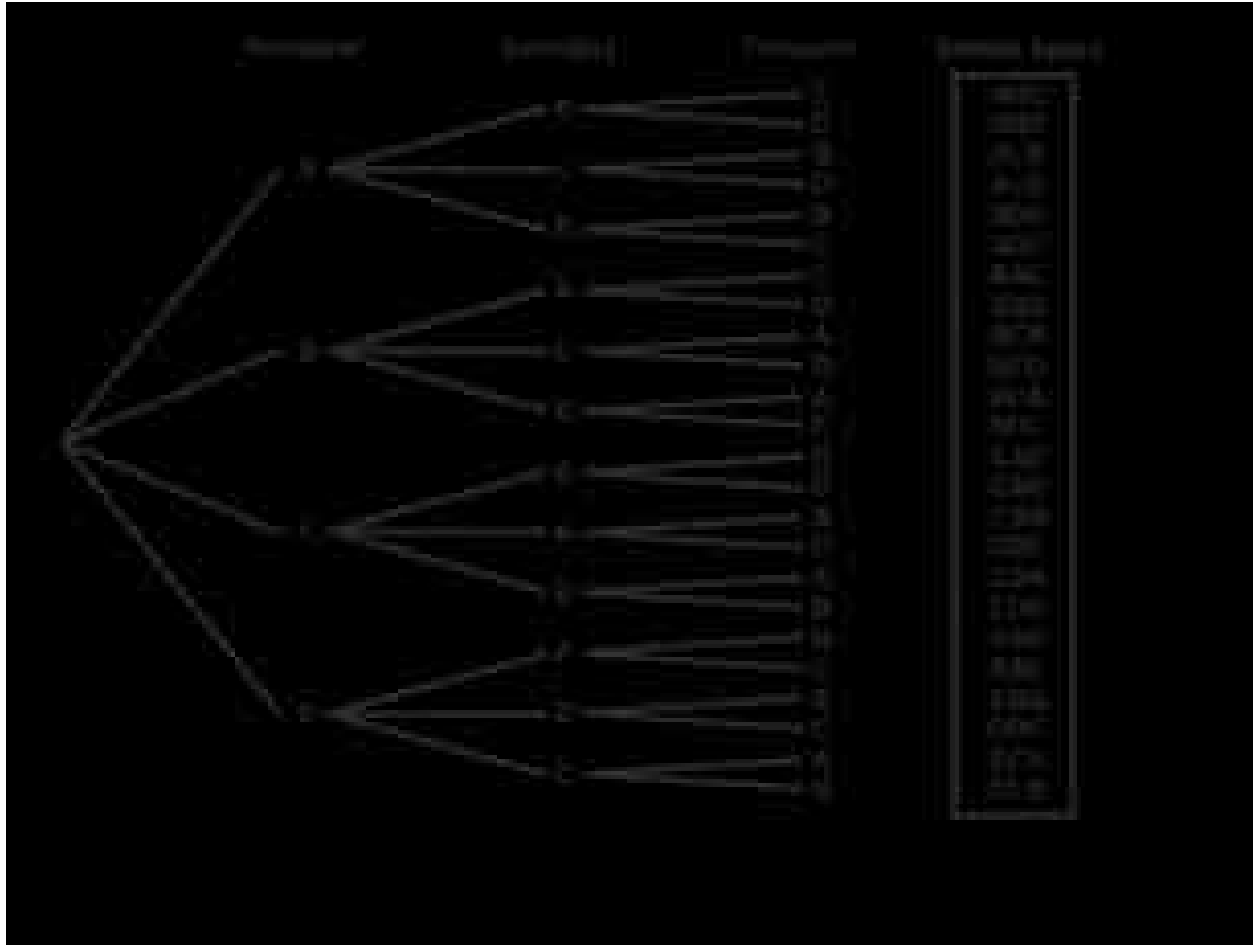
Example

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 ${}^nP_r$, if we put $r = n$, we obtain:

$$\begin{aligned} {}^nP_n &= n(n-1)(n-2) \dots 3 \times 2 \times 1 \\ &= n! \end{aligned}$$

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

$${}^3P_3 = 3! = 6$$

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 n_1 of one kind, n_2 of a second kind, ..., n_k of a K th kind,

$$\text{is } P = \frac{n!}{n_1! n_2! \dots n_k!}.$$

(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)

$n_1 = 1$ (because there is one c)

$n_2 = 1$ (because there is one o)

$n_3 = 2$ (because there are two m’s)

$n_4 = 1$ (because there is one i)

$n_5 = 2$ (because there are two t’s)

and

$n_6 = 2$ (because there are two e’s)

Hence, the total number of (meaningless) words (permutations) is:

$$\begin{aligned} P &= \frac{n!}{n_1! n_2! \dots n_k!} \\ &= \frac{9!}{1! 1! 2! 1! 2! 2!} \\ &= \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{1 \times 1 \times 2 \times 1 \times 1 \times 2 \times 1 \times 2 \times 1} \\ &= 45360 \end{aligned}$$

Topic 57

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:

$${}^nC_r \text{ or } \binom{n}{r},$$

and is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

where $r < n$.

It should be noted that 

$${}^nP_r = r! \binom{n}{r}$$

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 ${}^nC_r$ 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.

$$i): \quad \binom{n}{r} = \binom{n}{n-r}$$

$$ii): \quad \binom{n}{n-r} + \binom{n}{r} = \binom{n+1}{r}$$

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

$$\begin{aligned} \binom{n}{r} &= \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!} \\ &= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \\ &= 120 \end{aligned}$$

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.

Topic 58

Bernoulli Distribution

A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial — a random experiment that has only two outcomes (usually called a “Success” or a “Failure”).

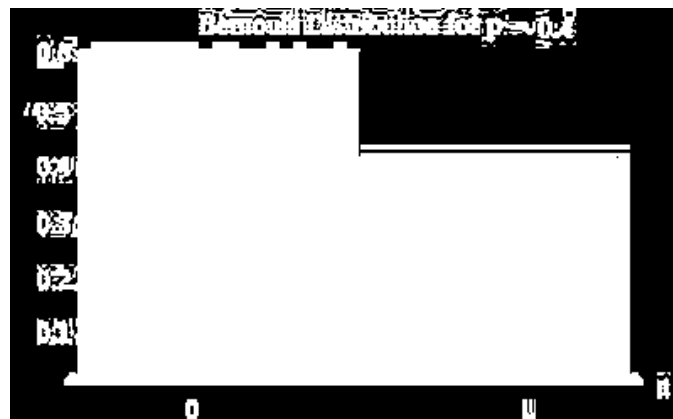
For example,

The probability of getting a heads (a “success”) while flipping a coin is 0.5.

The probability of “failure” is $1 - P$ (1 minus the probability of success, which also equals 0.5 for a coin toss).

It is a special case of the binomial distribution for $n = 1$. In other words, it is a binomial distribution with a single trial (e.g. a single coin toss).

The probability of a failure is labeled on the x-axis as 0 and success is labeled as 1. In the following Bernoulli distribution, the probability of success (1) is 0.4, and the probability of failure (0) is 0.6:



The probability density function (pdf) for this distribution is $p^x (1 - p)^{1-x}$, which can also be written as:

$$P(n) = \begin{cases} 1 - p & \text{for } n = 0 \\ p & \text{for } n = 1 \end{cases}$$

The expected value for a random variable, X , for a Bernoulli distribution is:

$$E[X] = p.$$

For example, if $p = .04$, then $E[X] = 0.04$.

The variance of a Bernoulli random variable is:

$$\text{Var}[X] = p(1 - p).$$

Topic 59

Bernoulli Trials and Bernoulli Distribution

What is a Bernoulli Trial?

A Bernoulli trial is one of the simplest experiments you can conduct. It's an experiment where you can have one of two possible outcomes. For example, "Yes" and "No" or "Heads" and "Tails." A few examples:

Coin tosses: record how many coins land heads up and how many land tails up.

Births: how many boys are born and how many girls are born each day.

Rolling Dice: the probability of a roll of two die resulting in a double six.

Bernoulli trials are usually phrased in terms of success and failure. Success doesn't mean success in the usual way—it just refers to an outcome you want to keep track of. For example, you might want to find out how many boys are born each day, so you call a boy birth a "success" and a girl birth a "failure." In the dice rolling example, a double six die roll would be your "success" and everything else rolled would be considered a "failure."

Independence

An important part of every Bernoulli trial is that each action must be independent. That means the probabilities must remain the same throughout the trials; each event must be completely separate and have nothing to do with the previous event.

Winning a scratch off lottery is an independent event. Your odds of winning on one ticket are the same as winning on any other ticket. On the other hand, drawing lotto numbers is a dependent event. Lotto numbers come out of a ball (the numbers aren't replaced) so the probability of successive numbers being picked depends upon how many balls are left; when there's a hundred balls, the probability is $1/100$ that any number will be picked, but when there are only ten balls left, the probability shoots up to $1/10$. While it's possible to find those probabilities, it isn't a Bernoulli trial because the events (picking the numbers) are connected to each other.

- The binomial distribution,
- The geometric distribution,
- The negative binomial distribution.

The Bernoulli distribution is closely related to the Binomial distribution. As long as each individual Bernoulli trial is independent, then the number of successes in a series of Bernoulli trials has a Binomial Distribution. The Bernoulli distribution can also be defined as the Binomial distribution with $n = 1$.

Use of the Bernoulli Distribution in Epidemiology

In experiments and clinical trials, the Bernoulli distribution is sometimes used to model a single individual experiencing an event like death, a disease, or disease exposure. The model is an excellent indicator of the probability a person has the event in question.

- 1 = “event” ($P = p$)
- 0 = “non event” ($P = 1 - p$)

Bernoulli distributions are used in logistic regression to model disease occurrence.

Assumptions for Bernoulli Trials

The three assumptions for Bernoulli trials are:

1. Each trial has two possible outcomes: Success or Failure. We are interested in the number of Successes X ($X = 0, 1, 2, 3, \dots$).
2. The probability of Success (and of Failure) is constant for each trial; a “Success” is denoted by the letter p and “Failure” is $q = 1 - p$.
3. Each trial is independent; The outcome of previous trials has no influence on any subsequent trials.

Bernoulli Distribution

A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial — a random experiment that has only two outcomes (usually called a “Success” or a “Failure”).

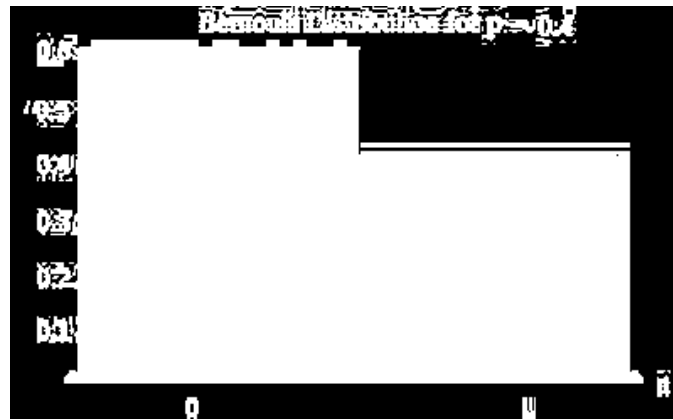
For example,

The probability of getting a heads (a “success”) while flipping a coin is 0.5.

The probability of “failure” is $1 - P$ (1 minus the probability of success, which also equals 0.5 for a coin toss).

It is a special case of the binomial distribution for $n = 1$. In other words, it is a binomial distribution with a single trial (e.g. a single coin toss).

The probability of a failure is labeled on the x-axis as 0 and success is labeled as 1. In the following Bernoulli distribution, the probability of success (1) is 0.4, and the probability of failure (0) is 0.6:



The probability density function (pdf) for this distribution is $p^x (1 - p)^{1-x}$, which can also be written as:

$$P(n) = \begin{cases} 1 - p & \text{for } n = 0 \\ p & \text{for } n = 1 \end{cases}$$

The expected value for a random variable, X, for a Bernoulli distribution is:

$$E[X] = p.$$

For example, if $p = .04$, then $E[X] = 0.04$.

The variance of a Bernoulli random variable is:

$$Var[X] = p(1 - p).$$

Topic 60

Standard Error

The standard error is one of the mathematical tools used in statistics to estimate the variability. It is abbreviated as SE. The standard error of a statistic or an estimate of a parameter is the standard deviation of its sampling distribution. We can define it as an estimate of that standard deviation.

Standard Error Formula

The accuracy of a sample that describes a population is identified through the SE formula. The sample mean which deviates from the given population and that deviation is given as;

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

Standard Error of the Mean (SEM)

The standard error of the mean also called the standard deviation of mean, is represented as the standard deviation of the measure of the sample mean of the population. It is abbreviated as SEM. For example, normally, the estimator of the population mean is the sample mean. But, if we draw another sample from the same population, it may provide a distinct value.

Thus, there would be a population of the sampled means having its distinct variance and mean. It may be defined as the standard deviation of such sample means of all the possible samples taken from the same given population. SEM defines an estimate of standard deviation which has been computed from the sample. It is calculated as the ratio of the standard deviation to the root of sample size, such as:

$$SEM = \frac{s}{\sqrt{n}}$$

Standard Error of Estimate (SEE)

The **standard error** of the **estimate** is the estimation of the accuracy of any predictions. It is denoted as SEE. The regression line depreciates the sum of squared deviations of prediction. It is also known as the sum of squares **error**. **SEE** is the square root of the average squared **deviation**. The deviation of some estimates from intended values is given by standard error of estimate formula.

$$SEE = \sqrt{\frac{\sum(x_i - \bar{x})}{n - 2}}$$

Example

Calculate the standard error of the given data:

y: 5, 10, 12, 15, 20

Solution:

First we have to find the mean of the given data;

$$\text{Mean} = (5+10+12+15+20)/5 = 62/5 = 10.5$$

Now, the standard deviation can be calculated as;

S = Summation of difference between each value of given data and the mean value/Number of values.

Hence,



The image shows a handwritten calculation for the standard deviation S. It starts with the formula $S = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$. Below this, the values are substituted: $S = \sqrt{\frac{(5-10.5)^2 + (10-10.5)^2 + (12-10.5)^2 + (15-10.5)^2 + (20-10.5)^2}{5}}$. The final result is $S = 5.35$.

After solving the above equation, we get;

$$S = 5.35$$

Therefore, SE can be estimated with the formula;

$$SE = S/\sqrt{n}$$

$$SE = 5.35/\sqrt{5} = 2.39$$

Topic 61

Sampling Distribution of the sample mean

The sampling distribution of the mean was defined in the section introducing sampling distributions. This section reviews some important properties of the sampling distribution of the mean introduced in the demonstrations in this chapter.

Mean

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ , then the mean of the sampling distribution of the mean is also μ . The symbol μ_M is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu_M = \mu$$

Variance

The variance of the sampling distribution of the mean is computed as follows:

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

That is, the variance of the sampling distribution of the mean is the population variance divided by N , the sample size (the number of scores used to compute a mean).

Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

Example

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one.

Solution

The following table shows all possible samples with replacement of size two, along with the mean of each:

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
152, 152	152	156, 152	154	160, 152	156	164, 152	158
152, 156	154	156, 156	156	160, 156	158	164, 156	160
152, 160	156	156, 160	158	160, 160	160	164, 160	162
152, 164	158	156, 164	160	160, 164	162	164, 164	164

Topic 62

Expected value and variance of the sampling distribution of the sample mean

Example

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\bar{x}$.

Solution

The following table shows all possible samples with replacement of size two, along with the mean of each:

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
152, 152	152	156, 152	154	160, 152	156	164, 152	158
152, 156	154	156, 156	156	160, 156	158	164, 156	160
152, 160	156	156, 160	158	160, 160	160	164, 160	162
152, 164	158	156, 164	160	160, 164	162	164, 164	164

The table shows that there are seven possible values of the sample mean $\bar{x}$.

The value $\bar{x} = 152$ happens only one way (the rower weighing 152 pounds must be selected both times), as does the value $\bar{x} = 164$, but the other values happen more than one way, hence are more likely to be observed than 152 and 164 are.

Since the 16 samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$\bar{x}$	152	154	156	158	160	162	164
$p(\bar{x})$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

$\bar{x}$	$p(\bar{x})$	$\bar{x}p(\bar{x})$	$\bar{x}^2$	$\bar{x}^2 p(\bar{x})$
152	1/16	152/16	23104	23104/16
154	2/16	308/16	23716	47432/16
156	3/16	468/16	24336	73008/16
158	4/16	632/16	24964	99856/16
160	3/16	480/16	25600	76800/16
162	2/16	324/16	26244	52488/16
164	1/16	164/16	26896	26896/16
	1	2528/16		399584/16

$$\mu_{\bar{x}} = \sum \bar{x} \cdot p(\bar{x})$$

$$\mu_{\bar{x}} = 158$$

$$\sum \bar{x} \cdot p(\bar{x}) = \frac{2528}{16}$$

$$\sum \bar{x}^2 \cdot p(\bar{x}) = \frac{399584}{16}$$

$$\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 \cdot p(\bar{x}) - \left[\sum \bar{x} \cdot p(\bar{x}) \right]^2}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{399584}{16} - \left(\frac{2528}{16} \right)^2}$$

$$\sigma_{\bar{x}} = \sqrt{24974 - 24964}$$

$$\sigma_{\bar{x}} = \sqrt{10}$$

$$\sigma_{\bar{x}} = 3.16$$

Topic 63

Estimation

In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Point Estimate vs. Interval Estimate

Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

- **Point estimate.** A point estimate of a population parameter is a single value of a statistic. For example, the sample mean $\bar{x}$ is a point estimate of the population mean μ . Similarly, the sample proportion p is a point estimate of the population proportion P .
- **Interval estimate.** An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an interval estimate of the population mean μ . It indicates that the population mean is greater than a but less than b .

Confidence Intervals

Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic + margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic + margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level

The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).



Topic 64

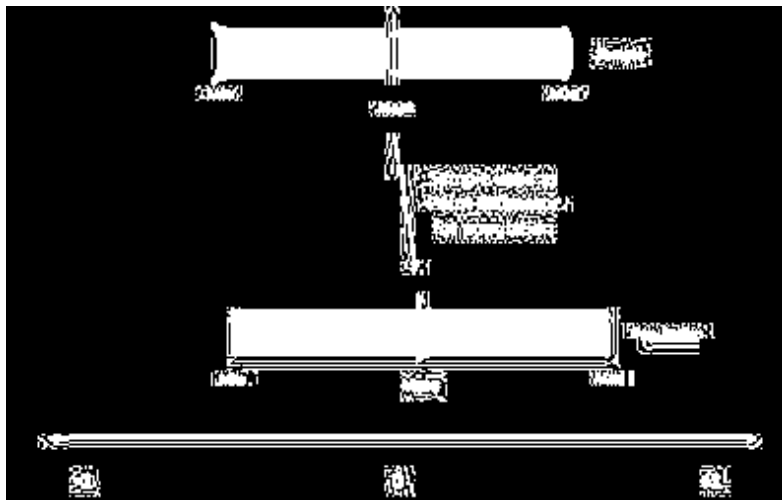
Point Estimation

Point Estimators

A function that is used to find an approximate value of a population parameter from random samples of the population

What are Point Estimators?

Point estimators are functions that are used to find an approximate value of a population parameter from random samples of the population. They use the sample data of a population to calculate a point estimate or a statistic that serves as the best estimate of an unknown parameter of a population.



Most often, the existing methods of finding the parameters of large populations are unrealistic. For example, when finding the average age of kids attending kindergarten, it will be impossible to collect the exact age of every kindergarten kid in the world. Instead, a statistician can use the point estimator to make an estimate of the population parameter.

Properties of Point Estimators

The following are the main characteristics of point estimators:

1. Bias

The bias of a point estimator is defined as the difference between the expected value of the estimator and the value of the parameter being estimated. When the estimated value of the parameter and the value of the parameter being estimated are equal, the estimator is considered unbiased.

Also, the closer the expected value of a parameter is to the value of the parameter being measured, the lesser the bias is.

2. Consistency

Consistency tells us how close the point estimator stays to the value of the parameter as it increases in size. The point estimator requires a large sample size for it to be more consistent and accurate.

You can also check if a point estimator is consistent by looking at its corresponding expected value and variance. For the point estimator to be consistent, the expected value should move toward the true value of the parameter.

3. Most efficient or unbiased

The most efficient point estimator is the one with the smallest variance of all the unbiased and consistent estimators. The variance measures the level of dispersion from the estimate, and the smallest variance should vary the least from one sample to the other.

Generally, the efficiency of the estimator depends on the distribution of the population. For example, in a normal distribution, the mean is considered more efficient than the median, but the same does not apply in asymmetrical distributions.

Common Methods of Finding Point Estimates

The process of point estimation involves utilizing the value of a statistic that is obtained from sample data to get the best estimate of the corresponding unknown parameter of the population. Several methods can be used to calculate the point estimators, and each method comes with different properties.

1. Method of moments

The method of moments of estimating parameters was introduced in 1887 by Russian mathematician Pafnuty Chebyshev. It starts by taking known facts about a population and then applying the facts to a sample of the population. The first step is to derive equations that relate the population moments to the unknown parameters.

The next step is to draw a sample of the population to be used to estimate the population moments. The equations derived in step one are then solved using the sample mean of the population moments. This produces the best estimate of the unknown population parameters.

2. Maximum likelihood estimator

The maximum likelihood estimator method of point estimation attempts to find the unknown parameters that maximize the likelihood function. It takes a known model and uses the values to compare data sets and find the most suitable match for the data.

For example, a researcher may be interested in knowing the average weight of babies born prematurely. Since it would be impossible to measure all babies born prematurely in the population, the researcher can take a sample from one location.

Because the weight of pre-term babies follows a normal distribution, the researcher can use the maximum likelihood estimator to find the average weight of the entire population of pre-term babies based on the sample data.

Topic 65

Interval Estimation

Interval Estimates

Interval estimates give an interval as the estimate for a parameter. This is a new concept which is the focus of this lesson. Such intervals are built around point estimates which is why understanding point estimates is important to understanding interval estimates.

In this course, the interval estimates we find are referred to as **confidence intervals**.

Confidence Interval

An interval of values computed from sample data that is likely to cover the true parameter of interest.

There are many estimators for population parameters. For example, if we want to know the "center" of a distribution, why use the mean? Could we use the median? How about using the middle value, i.e. $(\max + \min)/2$? We choose particular estimators for various reasons with information based on their sampling distributions. Here are some properties of "good" estimators.

General Format of a Confidence Interval

In putting the two properties above together, the center of our interval should be the point estimate for the parameter of interest. With the estimated standard error of the point estimate, we can include a measure of confidence to our estimate by forming a **margin of error**.

This you may have readily seen whenever you have heard or read a sample survey result (e.g. a survey of the current approval rating of the President, or attitude citizens have on some new policy). In such surveys, you may hear reference to the "44% of those surveyed approved of the President's reaction" (this is the sample proportion), and "the survey had a 3.5% margin or error, or $\pm 3.5\%$." This latter number is the margin of error.

With the point estimate and the margin of error, we have an interval for which the group conducting the survey is confident the parameter value falls (i.e. the proportion of U.S. citizens who approve of the President's reaction). In this example, that interval would be from 40.5% to 47.5%.

This example provides the general construction of a confidence interval:

General form of a confidence interval

$$\text{sample statistic} \pm \text{margin of error}$$

The margin of error will consist of two pieces. One is the standard error of the sample statistic. The other is some multiplier, M , of this standard error, based on how confident we want to be in our estimate. This multiplier will come from the same distribution as the sampling distribution of the point estimate; for example, as we will see with the sample proportion this multiplier will come from the standard normal distribution. The general form of the margin of error is shown below.

General form of the margin of error

$$\text{Margin of error} = M \times SE(\text{estimate})$$

*the multiplier, M , depends on our level of confidence

Interpretation of a Confidence Interval

The **interpretation** of a confidence interval has the basic template of: "We are 'some level of percent confident' that the 'population of interest' is from 'lower bound to upper bound'". The phrases in single quotes are replaced with the specific language of the problem. We will discuss more about the interpretation of a confidence interval after we provide a few more examples.

Topic 66

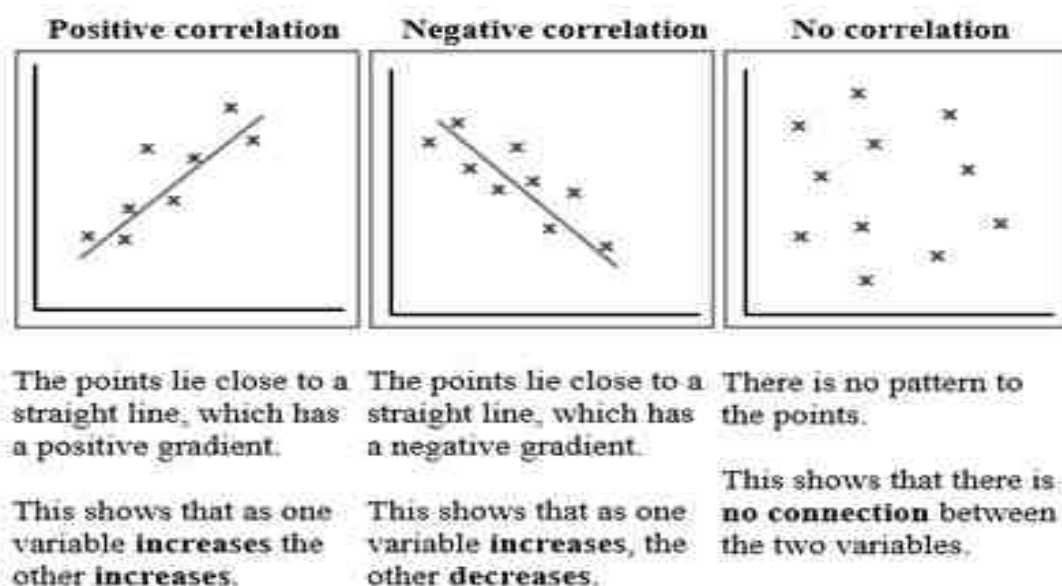
Correlation

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

- A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.
- A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
- A **zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.

Scatter diagrams

A correlation can be expressed visually. This is done by drawing a Scatter diagrams (also known as a scatterplot, scatter graph, scatter chart, or scatter diagram). A Scatter diagrams is a graphical display that shows the relationships or associations between two numerical variables (or co-variables), which are represented as points (or dots) for each pair of score. A scatter graph indicates the strength and direction of the correlation between the co-variables.



How to Find the Correlation?

The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:

$$r_{xy} = \frac{\sum x_i = \sum y_i = n}{\sqrt{\sum x_i^2 = \sum y_i^2}}$$

Where:

- r_{xy} – the correlation coefficient of the linear relationship between the variables x and y
- x_i – the values of the x-variable in a sample
- $\bar{x}$ – the mean of the values of the x-variable
- y_i – the values of the y-variable in a sample
- $\bar{y}$ – the mean of the values of the y-variable

Topic 67

Assumptions used for Correlation

The Five Assumptions for Pearson Correlation

The **Pearson correlation coefficient** (also known as the “product-moment correlation coefficient”) measures the linear association between two variables.

It always takes on a value between -1 and 1 where:

- -1 indicates a perfectly negative linear correlation between two variables
- 0 indicates no linear correlation between two variables
- 1 indicates a perfectly positive linear correlation between two variables

However, before we calculate the Pearson correlation coefficient between two variables we should make sure that five assumptions are met:

1. Level of Measurement: The two variables should be measured at the **interval** or **ratio** level.

2. Linear Relationship: There should exist a linear relationship between the two variables.

3. Normality: Both variables should be roughly normally distributed.

4. Related Pairs: Each observation in the dataset should have a pair of values.

5. No Outliers: There should be no extreme outliers in the dataset.

Assumption 1: Level of Measurement

To calculate a Pearson correlation coefficient between two variables, both of the variables should be measured at the **interval** or **ratio** level.

The following graphic provides a quick explanation of the four levels that variables can be measured at:

Levels of Measurement			
Nominal	Ordinal	Interval	Ratio
"Eye color"	"Level of satisfaction"	"Temperature"	"Height"
Named	Named	Named	Named
	Natural order	Natural order	Natural order
		Equal interval between variables	Equal interval between variables
			Has a "true zero" value, thus ratio between values can be calculated

Some examples of variables that can be measured on an **interval** scale include:

- **Temperature:** Measured in Fahrenheit or Celsius
- **Credit Scores:** Measured from 300 to 850
- **SAT Scores:** Measured from 400 to 1,600

Some examples of variables that can be measured on a **ratio** scale include:

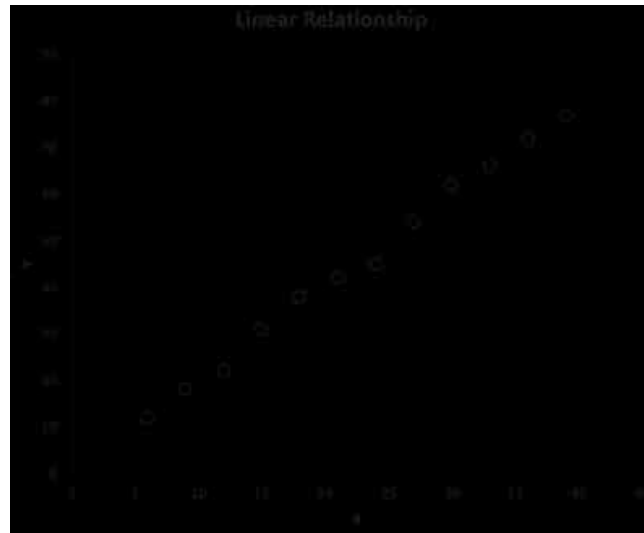
- **Height:** Measured in centimeters, inches, feet, etc.
- **Weight:** Measured in kilograms, pounds, etc.
- **Length:** Measured in centimeters, inches, feet, etc.

If the variables are measured at an **ordinal** level, then you should instead calculate the Spearman correlation coefficient between them.

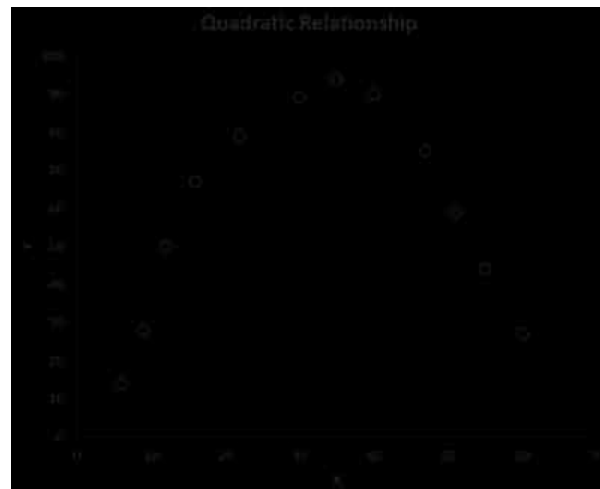
Assumption 2: Linear Relationship

To calculate a Pearson Correlation coefficient between two variables, there should exist a linear relationship between the two variables.

The easiest way to check this assumption is to simply create a scatter plot of the two variables. If the points in the plot fall roughly along a straight line, then a linear relationship exists:



However, if the points are randomly scattered about the plot or if they exhibit some other type of relationship (like quadratic) then a linear relationship does not exist between the variables:



In this case, a Pearson Correlation coefficient won't do a good job of capturing the relationship between the variables.

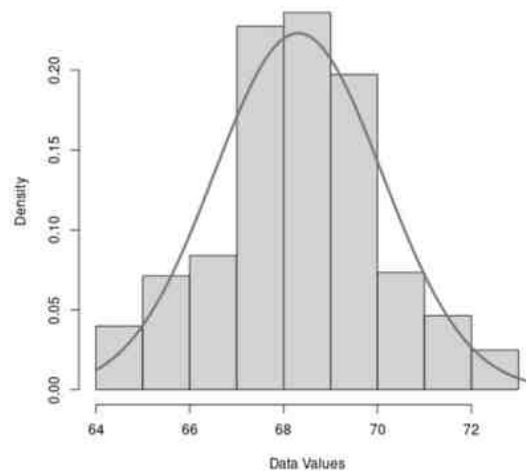
Assumption 3: Normality

A Pearson Correlation coefficient also assumes that both variables are roughly normally distributed.

You can check this assumption visually by creating a histogram or a Q-Q plot for each variable.

1. Histogram

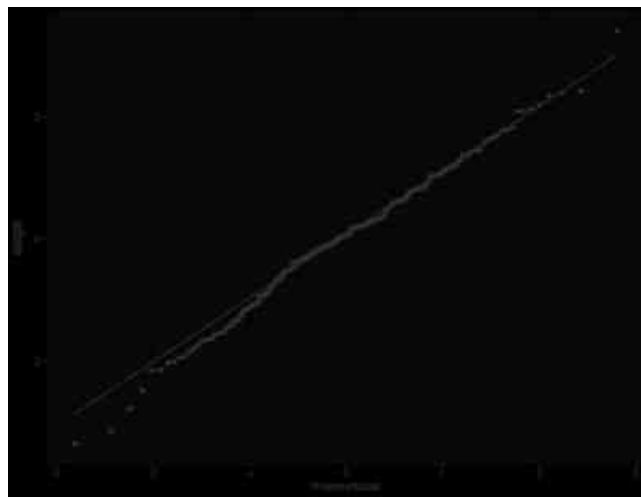
If a histogram for a dataset is roughly bell-shaped, then it's likely that the data is normally distributed.



2. Q-Q Plot

A Q-Q plot, short for “quantile-quantile” plot, is a type of plot that displays theoretical quantiles along the x-axis (i.e. where your data would lie if it did follow a normal distribution) and sample quantiles along the y-axis (i.e. where your data actually lies).

If the data values fall along a roughly straight line at a 45-degree angle, then the data is assumed to be normally distributed.



If the p-value of the test is less than a certain significance level (like $\alpha = 0.05$) then you have sufficient evidence to say that the data is *not* normally distributed.

There are three statistical tests that are commonly used to test for normality:

1. The Jarque-Bera Test
2. The Shapiro-Wilk Test
3. The Kolmogorov-Smirnov Test

Assumption 4: Related Pairs

A Pearson Correlation coefficient also assumes that each observation in the dataset should have a pair of values.

This assumption is easy to check. For example, if you're calculating the correlation between weight and height then simply verify that each observation in the dataset has one measurement for weight and one measurement for height.

Assumption 5: No Outliers

A Pearson Correlation coefficient also assumes that there are no extreme outliers in the dataset since outliers heavily affect the calculation of the correlation coefficient.

Topic 68

Types of correlation and their usage

Types of correlation coefficients

You can choose from many different correlation coefficients based on the linearity of the relationship, the level of measurement of your variables, and the distribution of your data.

For high statistical power and accuracy, it's best to use the correlation coefficient that's most appropriate for your data.

The most commonly used correlation coefficient is Pearson's r because it allows for strong inferences. It's parametric and measures linear relationships. But if your data do not meet all assumptions for this test, you'll need to use a non-parametric test instead.

Non-parametric tests of rank correlation coefficients summarize non-linear relationships between variables. The Spearman's rho and Kendall's tau have the same conditions for use, but Kendall's tau is generally preferred for smaller samples whereas Spearman's rho is more widely used.

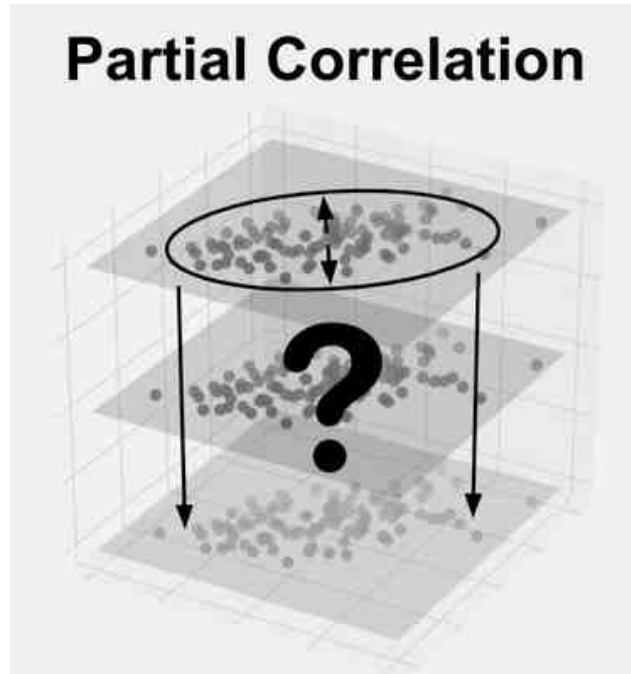
The table below is a selection of commonly used correlation coefficients.

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's ϕ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Topic 69

Partial Correlation

Partial Correlation is used to understand the strength of the relationship between two variables while accounting for the effects of one or more other variables. Your variables of interest should be continuous, be normally distributed, be linearly related, and be outlier free. In addition, your variables should have a similar spread across their individual ranges. See more below.



What is Partial Correlation?

Partial correlation explains the correlation between two continuous variables (let's say X1 and X2) holding X3 constant for both X1 and X2.

Partial Correlation Mathematical Formula

In this case, $r_{12.3}$ is the correlation between variables x1 and x2 keeping x3 constant. r_{13} is the correlation between variables x1 and x3.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

Assumptions for Partial Correlation

Every statistical method has assumptions. Assumptions mean that your data must satisfy certain properties in order for statistical method results to be accurate.

The assumptions for Pearson Correlation include:

1. Continuous
2. Normally Distributed
3. Linearity
4. No Outliers
5. Similar Spread Across Range
6. Covariate(s)

When to use Partial Correlation?

You should use Partial Correlation in the following scenario:

1. You want to know the **relationship** between two variables
2. Your variables of interest are **continuous**
3. You have **covariates**

Partial Correlation Example

Variable 1: Height

Variable 2: Weight

Covariate: Age

In this example, we are interested in the relationship between height and weight while accounting for the effect of age. So to begin, we collect height, weight, and age from a group of people.

First, we check that our variables of interest meet the assumptions of Partial Correlation. After confirming that height and weight are normally distributed, have no outliers, have a similar spread across their range, and are linearly related (see above for details), we move forward with the analysis.

The analysis will result in a correlation coefficient (called “ r ”) and a p-value. R values range from -1 to 1. A negative value of r indicates that the variables are inversely related (i.e. when one variable increases, the other decreases). On the other hand, positive values indicate that when one variable increases, so does the other.

The p-value represents the chance of seeing our results if there was no actual relationship between height and weight while controlling for the effects of age. A p-value less than or equal to 0.05 means that our result is statistically significant and we can trust that the difference is not due to chance alone.

Topic No. 70

Multiple Correlation

We can also calculate the correlation between more than two variables. The coefficient of multiple correlation measures the degree of relationship between a variable and its estimate from the regression equation. In other words, it is product moment correlation coefficient between variable say x_1 and its value estimated by the regression equation $x_1 = b_{12.3}x_2 + b_{13.2}x_3$

Topic No. 71

Calculating and interpreting Correlation

The coefficient of multiple correlation between x_1 and the variable x_2 and x_3 combined is denoted by symbolically by $R_{1.23}$.

$$R_{1.23} = \frac{\text{cov}(x_1, \hat{x}_1)}{\sqrt{\text{var}(x_1) \text{var}(\hat{x}_1)}} = \frac{\sum x_1 \hat{x}_1}{\sqrt{\sum x_1^2 \sum (\hat{x}_1^2)}}$$

Here $\sum x_1 \hat{x}_1 = \sum x_1 (x_1 - x_{1.23})$ ($\because \hat{x}' = x_1 - x_{1.23}$)

$$\begin{aligned} \sum x_1 \hat{x}_1 &= \sum x_1^2 x_1 - \sum x_1 x_{1.23} \\ &= \sum x_1^2 - \sum x_{1.23} x_{1.23} && \because \left(\sum x_1 x_{1.23} = \sum x_{1.23} x_{1.23} \right) \\ &= n(S_1^2 - S_{1.23}^2) \end{aligned}$$

Where, $S_{1.23}^2$ is the sample variance of residuals.

The multiple correlation coefficient measures the strength of association between the independent (explanatory) variables and the dependent variable (the variable we wish to forecast). Its value varies between 0 and 1. The higher value of the coefficient the stronger the association and the vice versa.

Topic No. 72

Testing Hypothesis about Correlation Coefficient

Pearson's correlation coefficient, r , tells us about the strength of the linear relationship between x and y points on a regression plot. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n , together. We perform a hypothesis test of the “significance of the correlation coefficient” to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is “close to 0” or “significantly different from 0”. We decide this based on the sample correlation coefficient r and the sample size n .

If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is “significant.”

Conclusion: “There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.”

What the conclusion means: There is a significant linear relationship between x and y we can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is “not significant.”

Conclusion: “There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from 0”.

What the conclusion means: There is not a significant linear relationship between x and y . Therefore, we can NOT use the regression line to model a linear relationship between x and y in the population.

Performing the Hypothesis Test

Our null hypothesis will be that the correlation coefficient IS NOT significantly different from 0. There IS NOT a significant linear relationship (correlation) between x and y in the population. Our alternative hypothesis will be that the population correlation coefficient IS significantly different from 0. There IS a significant linear relationship (correlation) between x and y in the population.

Steps for Hypothesis Testing for ρ

Step 1: Hypotheses

First, we specify the null and alternative hypotheses:

Null hypothesis $H_0: \rho=0$

Alternative hypothesis $H_A: \rho \neq 0$ or $\rho < 0$ or $\rho > 0$

Step 2: Test Statistic

Second, we calculate the value of the test statistic using the following formula:

$$\text{Test statistic: } t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Step 3: P-Value

Third, we use the resulting test statistic to calculate the P -value. As always, the P -value is the answer to the question "how likely is it that we'd get a test statistic t^* as extreme as we did if

the null hypothesis were true?" The P -value is determined by referring to a t -distribution with $n-2$ degrees of freedom.

Step 4: Decision

Finally, we make a decision:

If the P -value is smaller than the significance level α , we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α level to conclude that there is a linear relationship in the population between the predictor x and response y ."

If the P -value is larger than the significance level α , we fail to reject the null hypothesis. We conclude "there is not enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor x and response y ."

Using a Table of Critical Values to Make a Decision

The 95% critical values of the sample correlation coefficient table shown in gives us a good idea of whether the computed value of r is significant or not. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then we can use the line for prediction.

Degrees of Freedom: $n - 2$	Critical Values: $(+)$ and $(-)$
1	0.987
2	0.950
3	0.917
4	0.881
5	0.854
6	0.827
7	0.800
8	0.782
9	0.765
10	0.756
12	0.730
14	0.712
16	0.697
18	0.682
20	0.668
22	0.656
24	0.644
26	0.633
28	0.623
30	0.613
32	0.604
34	0.596
36	0.588
38	0.581
40	0.574
42	0.567
44	0.561
46	0.555
48	0.549
50	0.543
60	0.529
70	0.515
80	0.502
90	0.491
100	0.482

95% Critical Values of the Sample Correlation Coefficient Table: This table gives us a good idea of whether the computed value of r is significant or not.

As an example, suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are ± 0.632 . If r is less than the negative critical value or r is greater than the positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction.

Topic No. 73

Regression

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables)

There are two basic types of regression: simple linear regression and multiple linear regression. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y , while multiple linear regression uses two or more independent variables to predict the outcome.

The general form of each type of regression is:

Simple linear regression: $Y = \alpha + \beta X + \varepsilon$

Multiple linear regression: $Y = \alpha + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + \varepsilon$

Where:

Y = the variable that you are trying to predict (dependent variable).

X = the variable that you are using to predict Y (independent variable).

α = the intercept.

β = the slope.

ε = the regression residual.

Regression is often used to determine how many specific factors such as the price of a commodity, interest rates, particular industries, or sectors influence the price movement of an asset.

Topic No. 74

Assumptions for Regression Model

Assumptions in Regression

Regression is a parametric approach. 'Parametric' means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver good results with data sets which doesn't fulfill its assumptions. Therefore, for a successful regression analysis, it's essential to validate these assumptions. So, how would you check (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical test.

Let's look at the important assumptions in regression analysis:

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one-unit change in X^1 is constant, regardless of the value of X^1 . An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to heteroskedasticity.
5. The error terms must be normally distributed.

Topic No. 75

Steps in Regression Analysis

Six Steps in Applied Regression Analysis:

This is how we approach regression analysis in the real world.

1. Review the literature and develop a theoretical model.
2. Specify the econometric model: Select the independent variables and functional form.
3. Hypothesize the expected signs of the coefficients.
4. Collect the data. Inspect and clean the data.
5. Estimate and evaluate the equation. 6. Document the results.

Step 1: Literature review and theoretical model

Need to have a clear grasp of what question you are trying to answer. Once this is determined, a solid understanding of the existing literature is required to build a theoretical model. Use economic theory.

Step 2: Model specification

This is the most important step in regression analysis. Specifying the theoretical regression model requires choosing the independent variables and their functional form. 

Step 3: Hypothesized signs

Consider the relationship between the dependent and independent variables and what we might expect. We can use economic theory to ascertain why variables should have certain relationships.

Step 4: Data collection

We typically do use observational data.

Step 5: Estimation

SPSS can do the estimation in a second! We're interested in evaluating the resulting output. How well did we do? Did we get the signs "right" on the coefficients?

Step 6: Documentation

We should always document our results. Write-up should contain an explanation of the model, the data, the estimation results (coefficient estimates, standard errors, $\bar{R}^2$, n , etc.).

Topic No. 76

Introduction to SPSS

SPSS (Statistical Package for the Social Sciences) is a versatile and responsive program designed to undertake a range of statistical procedures. SPSS is a Windows based program that can be used to perform data entry and analysis and to create tables and graphs. SPSS is capable of handling large amounts of data and can perform all of the analyses covered in the text and much more. SPSS is commonly used in the Social Sciences and in the business world, so familiarity with this program should serve you well in the future. SPSS is updated often. This document was written around an earlier version, but the differences should not cause any problems.

The Four Windows of SPSS

There are 4 main windows in SPSS and we thought it would be useful to introduce you to what the 4 windows in SPSS are and what each is used for:

- Data editor
- Output viewer
- Syntax editor
- Script window

Data Editor

The data view is used to store and show your data. It is much like an ordinary spreadsheet although in general the data is structured so that rows are cases and the columns are for the different variables that relate to each case. Spreadsheet-like system for defining, entering, editing, and displaying data. Extension of the saved file will be “sav.”

SPSS Anxiety.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1: subject 1 Variable: 3 of 5 Variables

	subject	anxiety	tension	score	trial
1	1	1	1	18	1
2	1	1	1	14	2
3	1	1	1	12	3
4	1	1	1	6	4
5	2	1	1	19	1
6	2	1	1	12	2
7	2	1	1	8	3
8	2	1	1	4	4
9	3	1	1	14	1
10	3	1	1	10	2
11	3	1	1	6	3
12	3	1	1	2	4
13	4	1	2	16	1
14	4	1	2	12	2
15	4	1	2	8	3
16	4	1	2	4	4

Data View Variable View

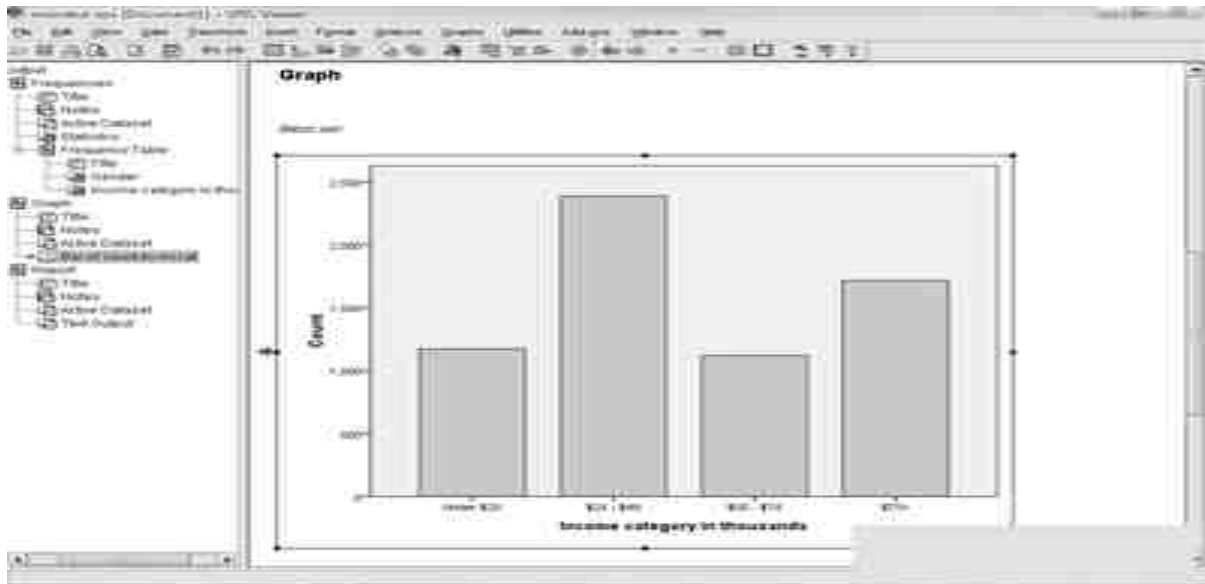
SPSS Processor is ready

Topic No. 77

A Tutorial to SPSS – II the Four Windows of SPSS (Cont...)

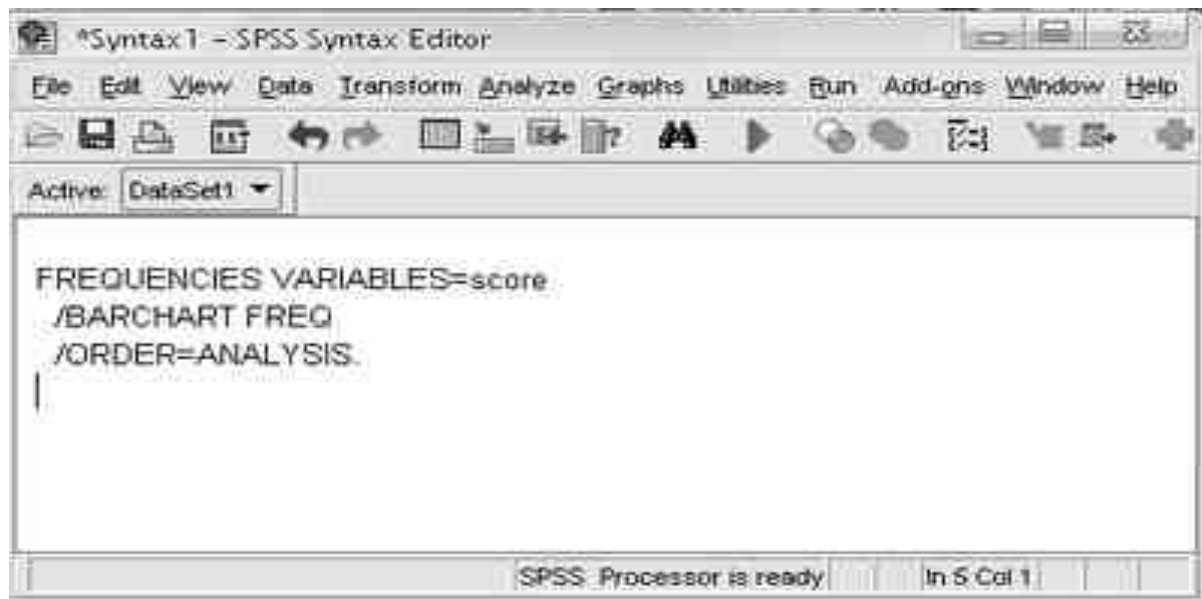
Output Viewer

This window is used to show the results that have been output from your data analysis. Depending on the analysis that you are carrying out this may include the Chart Editor Window or Pivot Table Window. Displays output and errors. Extension of the saved file will be “spv.”



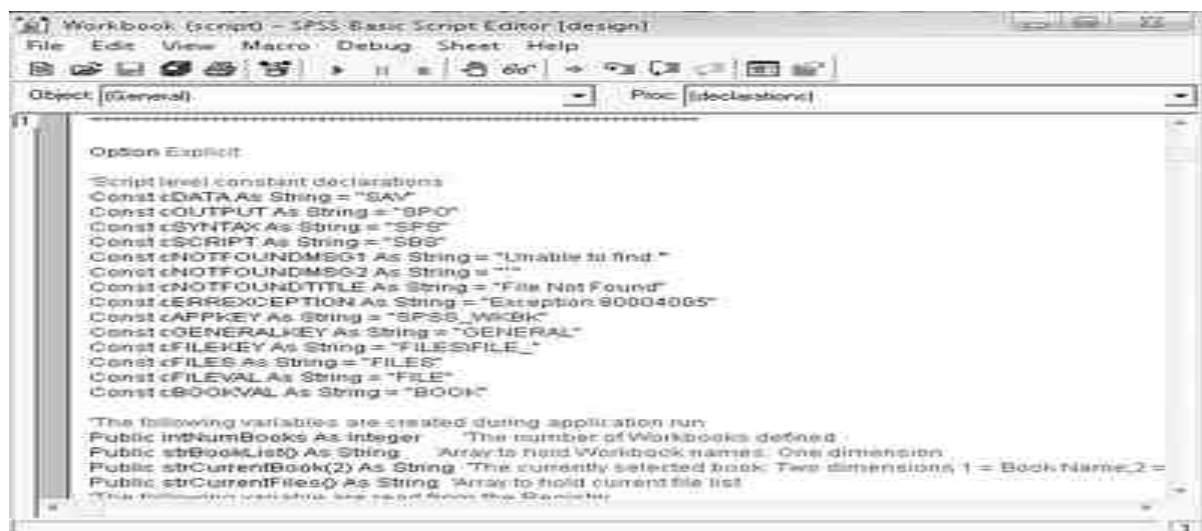
Syntax editor

This window shows the underlying commands that have executed your data analysis. If you are a confident coder this is where you can amend the code, or write your own from scratch, and then run your own custom analysis on your data set Text editor for syntax composition. Extension of the saved file will be “sps.”



Script Window

Provides the opportunity to write full-blown programs, in a BASIC-like language. Text editor for syntax composition. Extension of the saved file will be “sbs.”



Managing data in SPSS

Managing data well is a really important skill, and SPSS is a good place to learn some of the basics of data management.

Three key elements to data management processes include:

1. Importing data into SPSS;
2. Labelling data (variables and values); and
3. Sorting and merging data.

Before and after doing each of these processes, you should also know how to inspect your data.

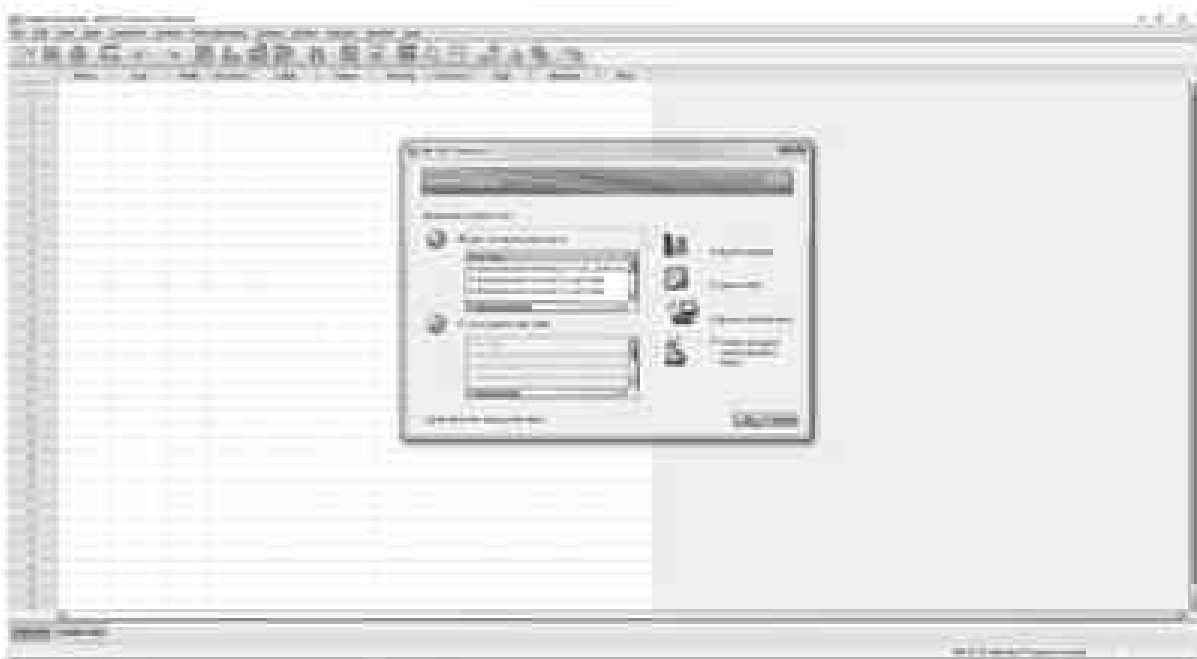
These four processes are covered in this chapter.

Topic No. 78

A Tutorial to SPSS – III Opening SPSS

Click on the **Start menu** () > **All Programs** > **IBM SPSS Statistics** > **IBM SPSS Statistics 21** (or whatever is the latest version number) to open the SPSS program.

When you first open SPSS on your computer, you should see something that looks similar to following screenshot:



SPSS automatically assumes that you want to open an existing file, and immediately opens a dialogue box to ask which file you'd like to open. It'll make it easier to navigate the interface and windows in SPSS if we open a file.

OR

Opening SPSS: Start → All Programs → SPSS Inc→ SPSS 16.0 → SPSS 16.0

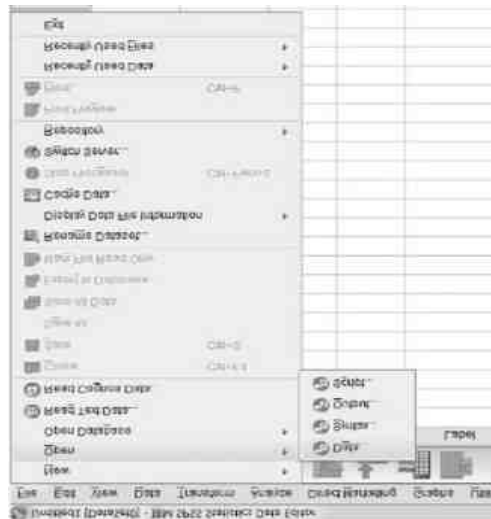


Opening data

It can be handy to know how to open a dataset once you're already in SPSS, using the drop-down menus.

So far, we've learned how to open an SPSS file when we first open SPSS. But sometimes you might already have SPSS open and wish to open a(nother) data file. Later on we will talk about how to use syntax files to work in SPSS, starting with opening data, but for now, here's how to open a file using the menus:

Click on the File menu at the top left of your Data Editor window, then Open > Data ... :



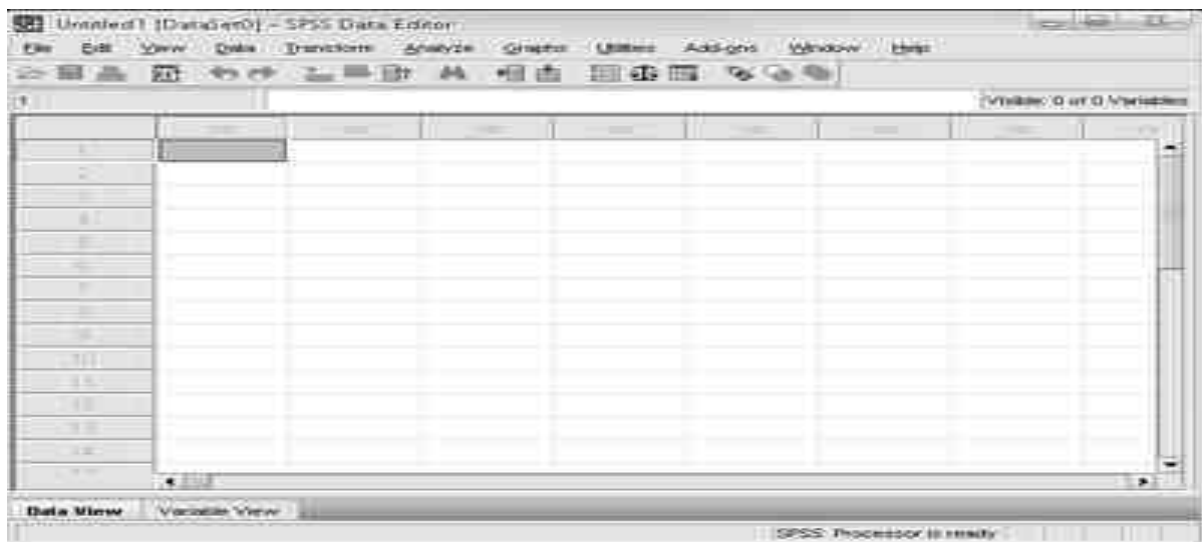
- There are two sheets in the window:

1. Data view

2. Variable view

Data View window

This sheet is visible when you first open the Data Editor and this sheet contains the data



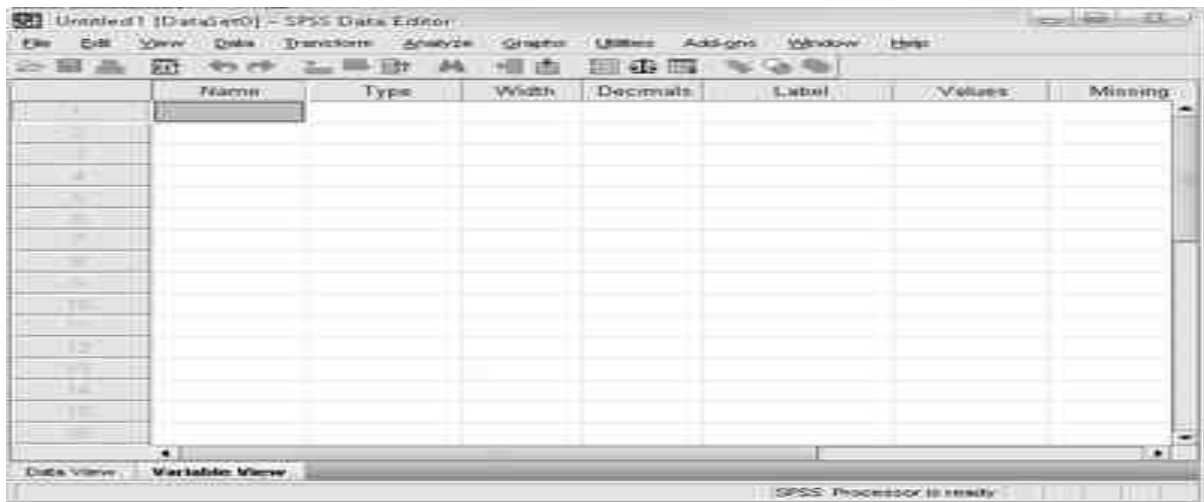
Topic No. 79

A Tutorial to SPSS – IV

Variable View

This sheet contains information about the data set that is stored with the dataset

Click on the tab labeled Variable View.



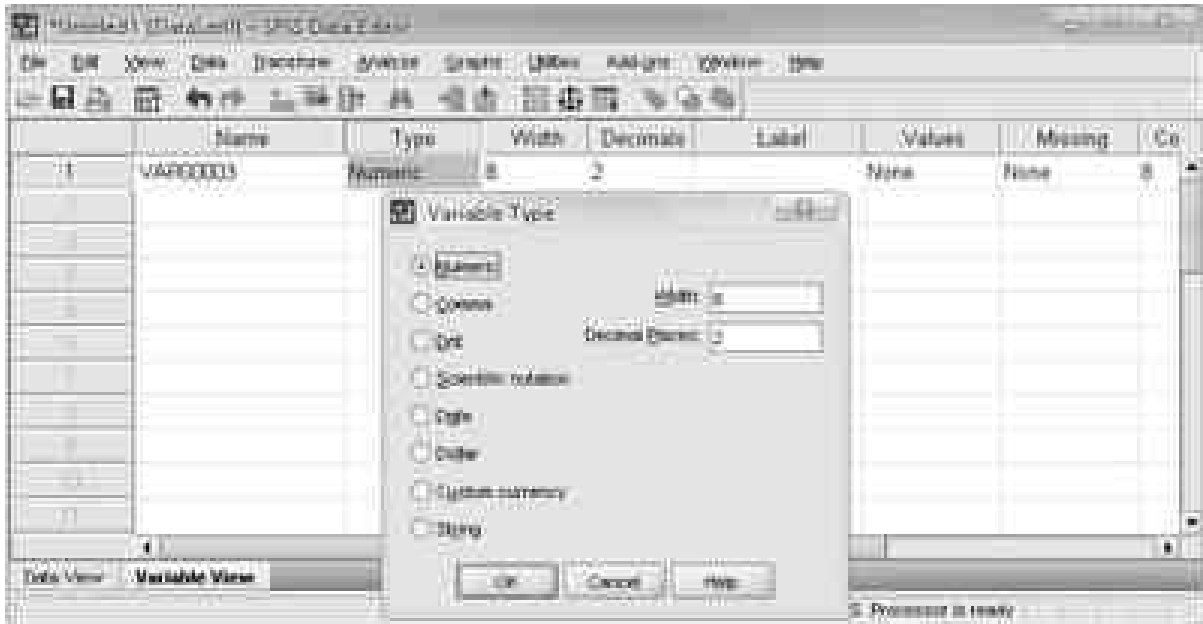
- The variable view contains the variables on your data set, so it defines the properties of your dataset. Each row will define all of the various variables for one set of data. For example, for a numerical piece of data this would show (amongst other things) the number of decimal places that are stored for that piece of data. The variables include - name, type, width, decimals, label, values, missing, columns, align and measure. Ensuring that the 'measure' of your variables is correct is vital. The variable can be Nominal which is for strings of data, Ordinal for data that isn't continuous but can be ranked or ordered or, finally, scale which is used for a variable that is continuous, for example a distance to somewhere.

Name

- The first character of the variable name must be alphabetic
- Variable names must be unique, and have to be less than 64 characters.
- Spaces are NOT allowed.

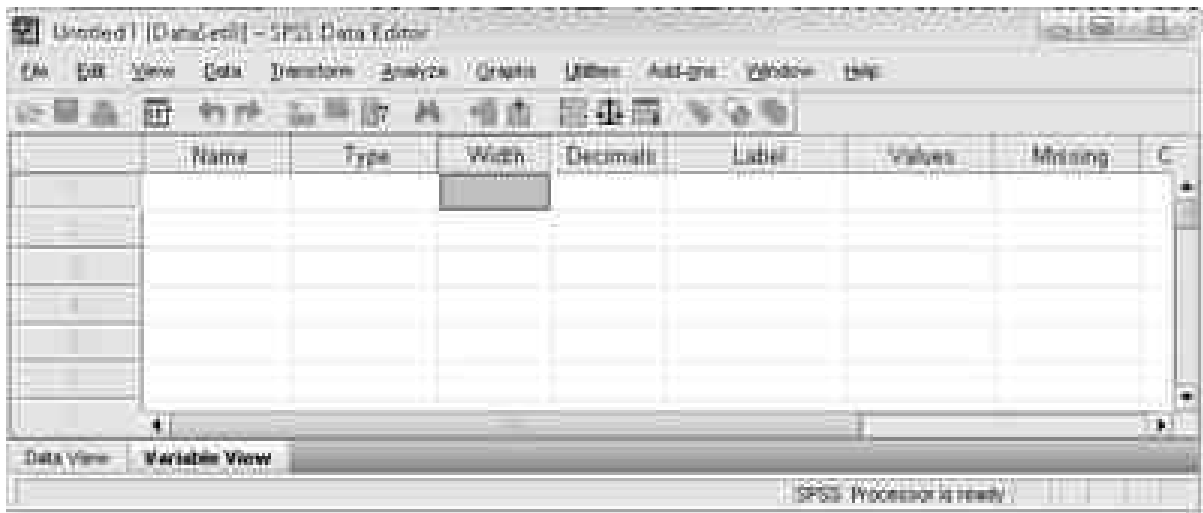
Type

- Click on the ‘type’ box. The two basic types of variables that you will use are numeric and string. This column enables you to specify the type of variable.



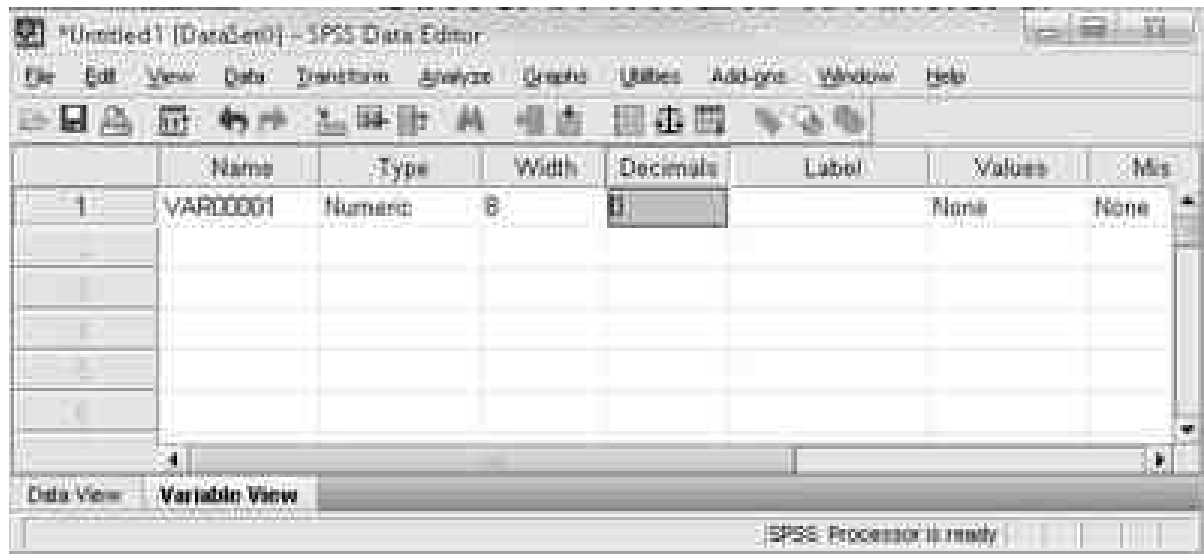
Width

- Width allows you to determine the number of characters SPSS will allow to be entered for the variable



Decimals

- Number of decimals
- It has to be less than or equal to 16

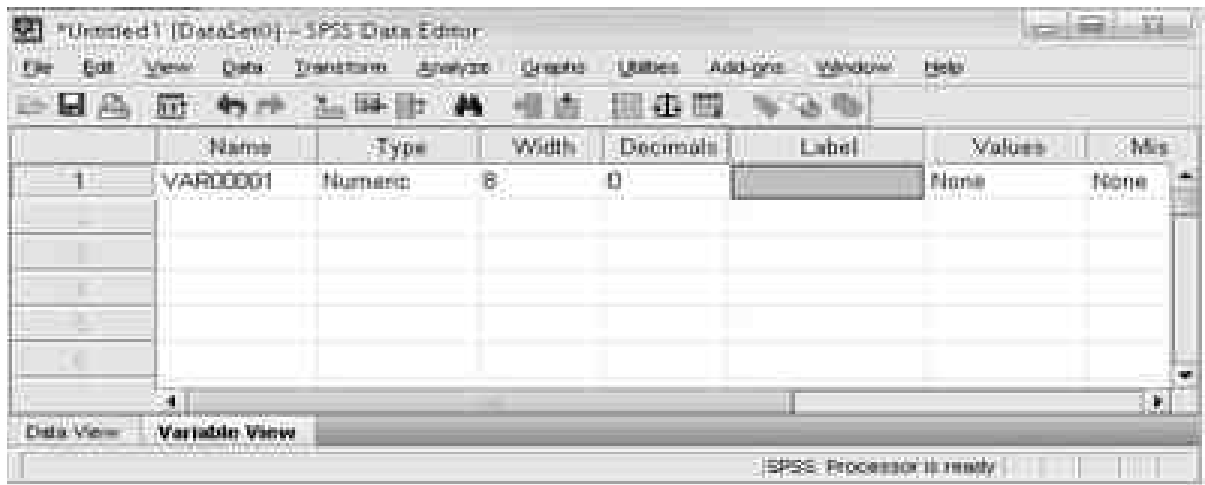


Topic No. 80

A Tutorial to SPSS – V

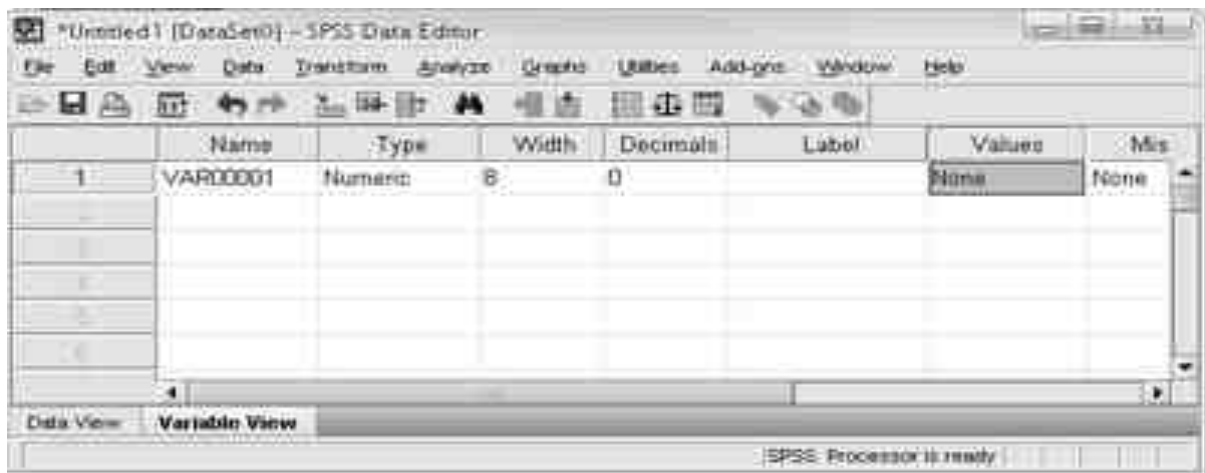
Label

- You can specify the details of the variable
- You can write characters with spaces up to 256 characters



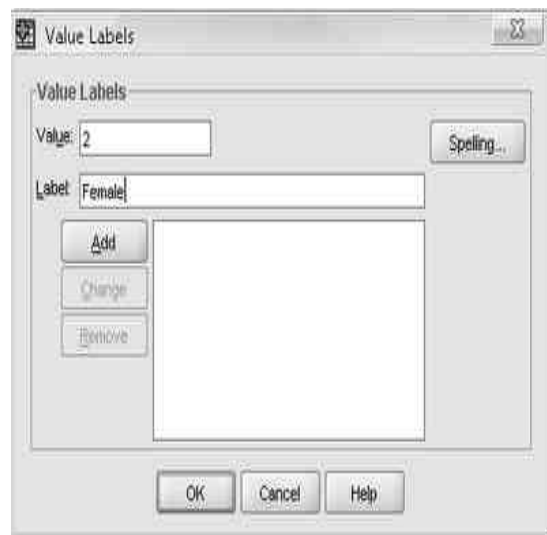
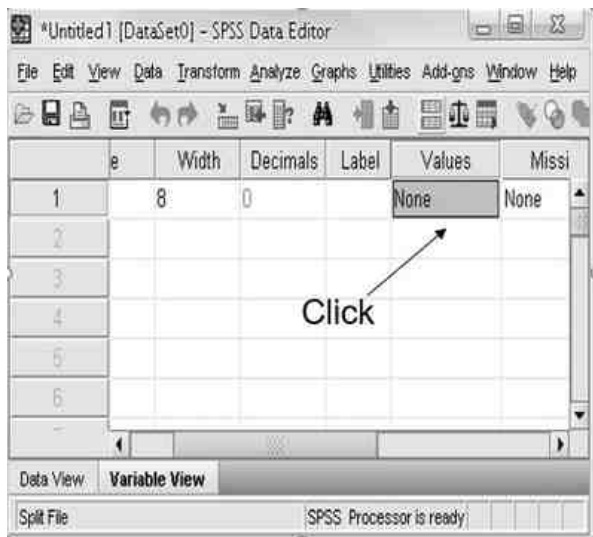
Values

- This is used to suggest which numbers represent which categories when the variable represents a category



Defining the value labels

- Click the cell in the values column as shown below
- For the value, and the label, you can put up to 60 characters.
- After defining the values click add and then click OK.



Topic No. 81

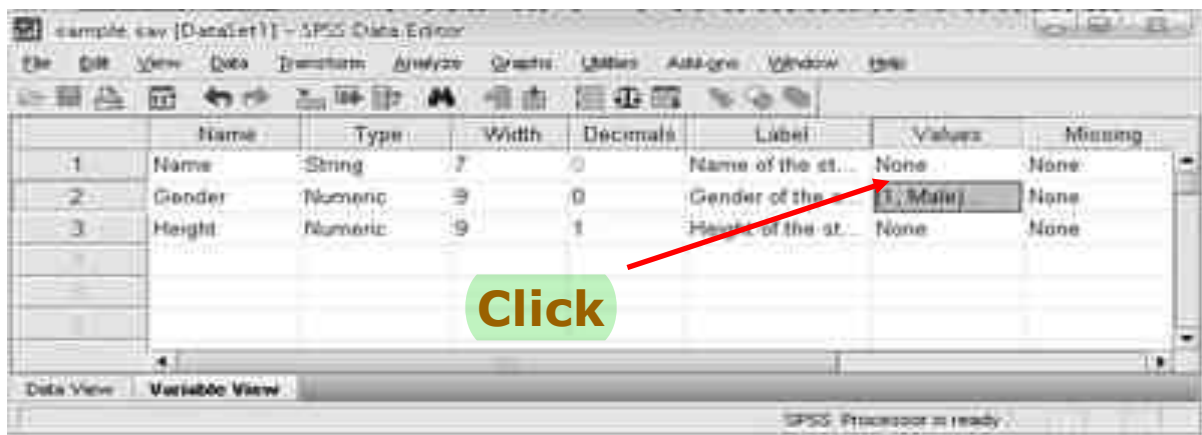
A Tutorial to SPSS – VI

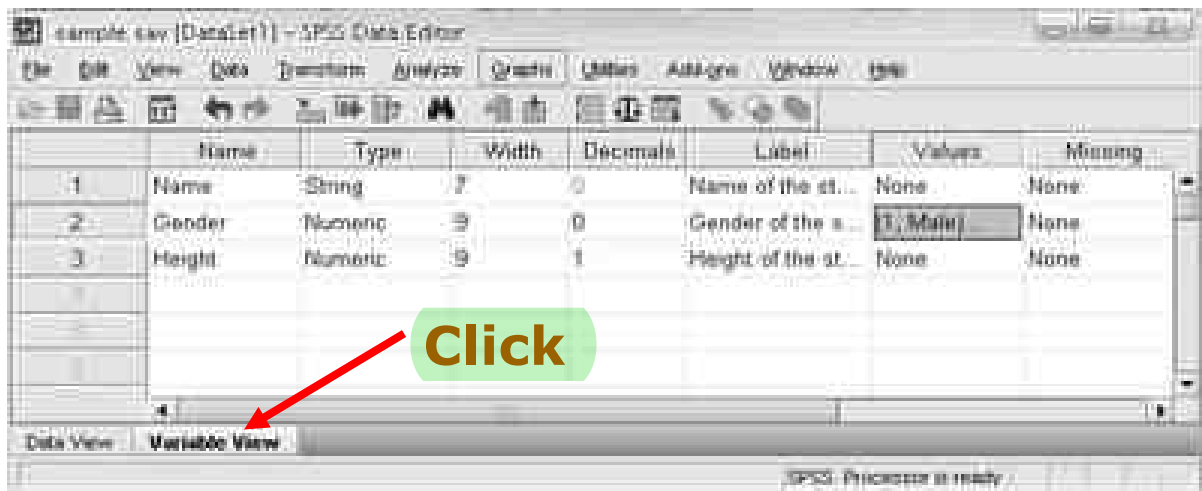
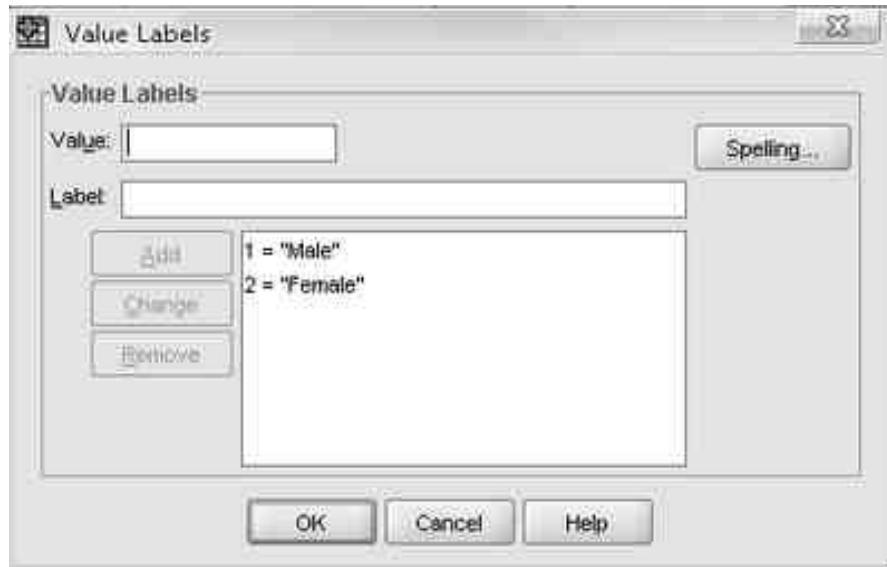
How would you put the following information into SPSS?

Name	Gender	Height
JAUNITA	2	5.4
SALLY	2	5.3
DONNA	2	5.6
SABRINA	2	5.7
JOHN	1	5.7
MARK	1	6
ERIC	1	6.4
BRUCE	1	5.9

Value = 1 represents Male and Value = 2 represents Female

Solution





sample.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

Visible: 3 of 3 Variables

	Name	Gender	Height
1	JAUNITA	2	5.4
2	SALLY	2	5.3
3	DONNA	2	5.6
4	SABRINA	2	5.7
5	JOHN	1	5.7
6	MARK	1	6.0
7	ERIC	1	6.4
8	BRUCE	1	5.9

Data View Variable View

Weight status area SPSS Processor is ready

Topic No. 82

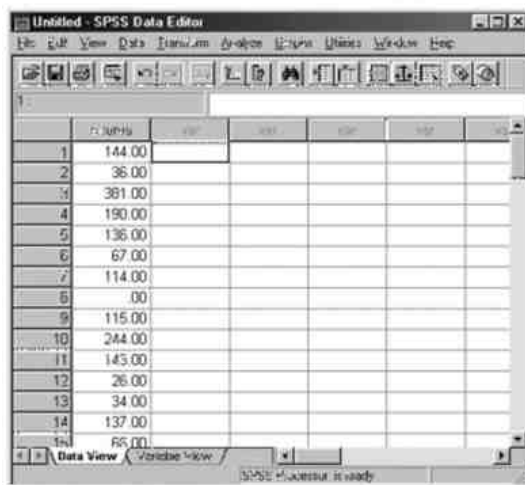
A Tutorial to SPSS – VII

Saving the data

To save the data file you created simply click ‘file’ and click ‘save as.’ You can save the file in different forms by clicking “Save as type.”

In other words When attempting to save files from the SPSS software, it is important to first remember that the only information that is saved is what it in the current window. For example, if the currently displayed window contains the output from an analysis (frequency tables, t-test results, graphs, etc.), the only information that will be contained in the resulting file is the output. The information from the Data Editor will not be saved. In order to save the information from the Data Editor, it must be active window by selecting Window at the top of the screen and then selecting the Data Editor. Once the Data Editor has been made the active window, it can be saved. The following illustrations provide a guide to saving Data and Output files in SPSS. After the data has been entered into the Data Editor, it can be saved by selecting File and Save or Save As.

Data Editor with Scores Entered



Selecting File -> Save or Save As to save the contents of the Data Editor

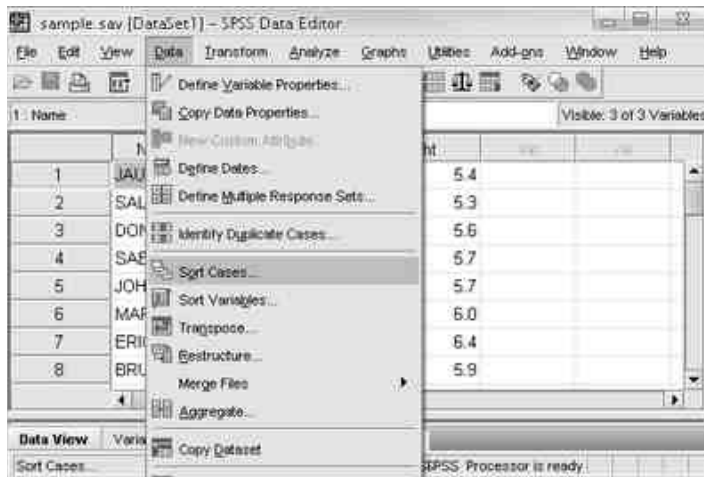


After selecting Save or Save As, the Save dialog box will appear. The name of the data file can be entered in the box labeled File name and the directory in which the file is

saved can be changed by selecting the down arrow on the right side of the box labeled Save in

Sorting the data

Click 'Data' and then click Sort Cases



Double Click 'Name of the students.' Then click ok

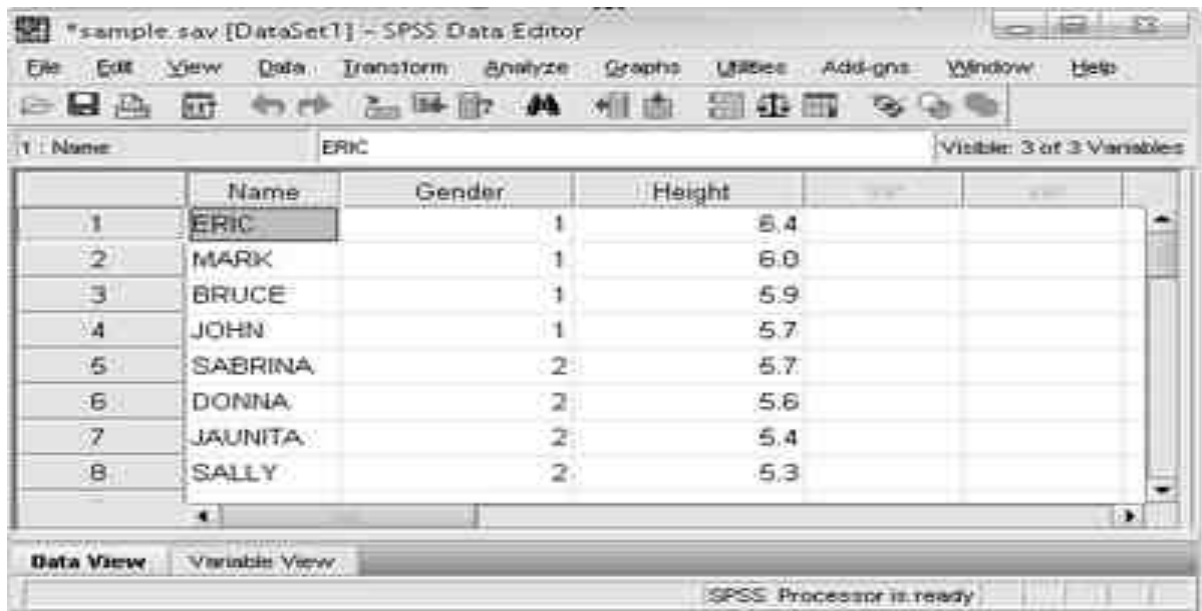


Topic No. 83

A Tutorial to SPSS – VIII

Practice 2

- How would you sort the data by the 'Height' of students in descending order?
- Answer
 - Click data, sort cases, double click 'height of students,' click 'descending,' and finally click ok.



*sample.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1: Name: ERIC Visible: 3 of 3 Variables

	Name	Gender	Height
1	ERIC	1	5.4
2	MARK	1	6.0
3	BRUCE	1	5.9
4	JOHN	1	5.7
5	SABRINA	2	5.7
6	DONNA	2	5.6
7	JAUNITA	2	5.4
8	SALLY	2	5.3

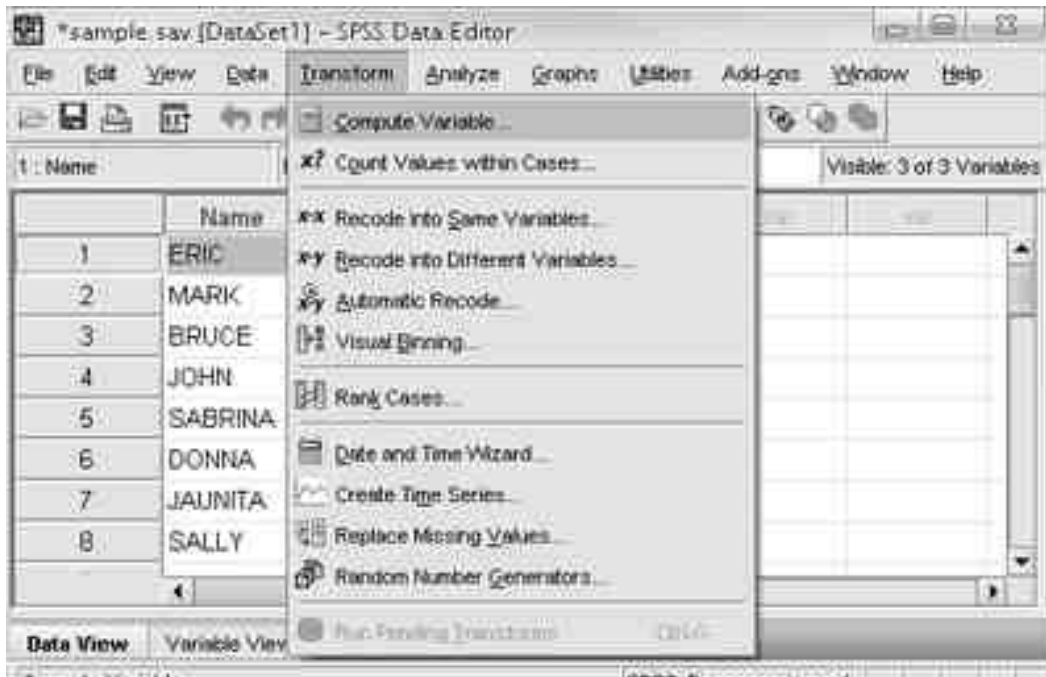
Data View Variable View

SPSS Processor is ready

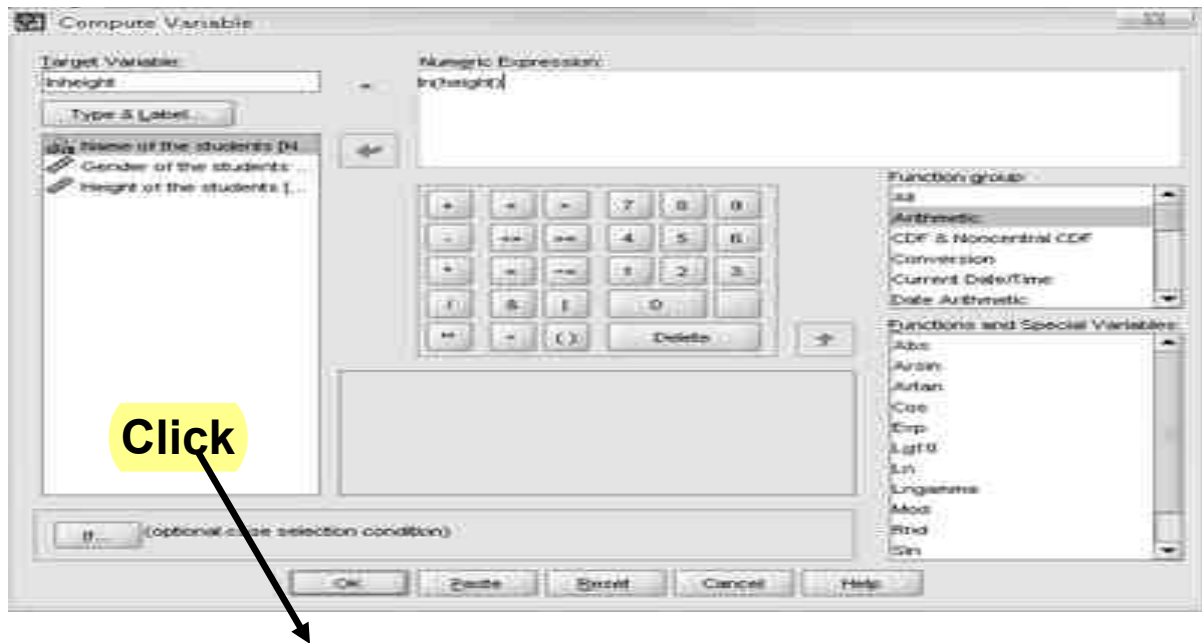
Transforming data

Transforming data is performed for a whole host of different reasons, but one of the most common is to apply a transformation to data that **is not normally distributed** so that the new, transformed data is normally distributed.

Click 'Transform' and then click 'Compute Variable...'



- Example: Adding a new variable named 'lnheight' which is the natural log of height
 - Type in lnheight in the 'Target Variable' box. Then type in 'ln(height)' in the 'Numeric Expression' box. Click OK



- A new variable 'ln height' is added to the table

*sample.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

Visible: 4 of 4 Variables

	Name	Gender	Height	lnheight
1	ERIC	1	6.4	1.86
2	MARK	1	6.0	1.79
3	BRUCE	1	5.9	1.77
4	JOHN	1	5.7	1.74
5	SABRINA	2	5.7	1.74
6	DONNA	2	5.6	1.72
7	JALUNITA	2	5.4	1.69
8	SALLY	2	5.3	1.67

Data View Variable View

[SPSS: Processor is ready]

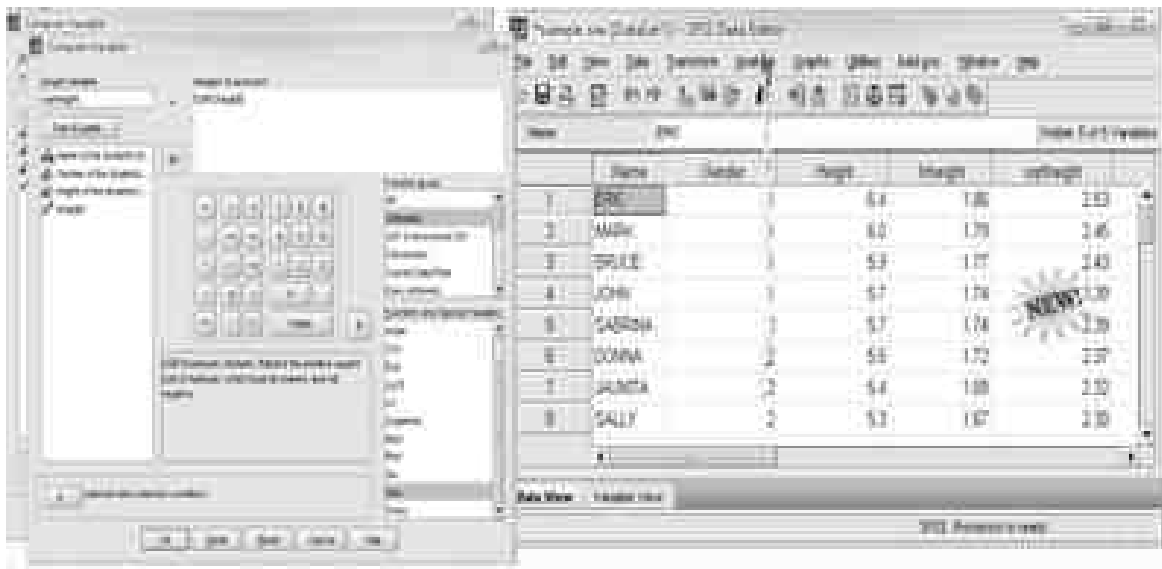
Topic No. 84

A Tutorial to SPSS – IX

Practice 3

Create a new variable named “sqrtheight” which is the square root of height.

Answer



The basic analysis

Creating tables and charts containing frequency counts or summary statistics over (groups of) cases and variables. Running inferential statistics such as ANOVA, regression and factor analysis. Saving data and output in a wide variety of file formats.

The basic analysis of SPSS that will be introduced in this class

Frequencies

- This analysis produces frequency tables showing frequency counts and percentages of the values of individual variables.

Descriptive

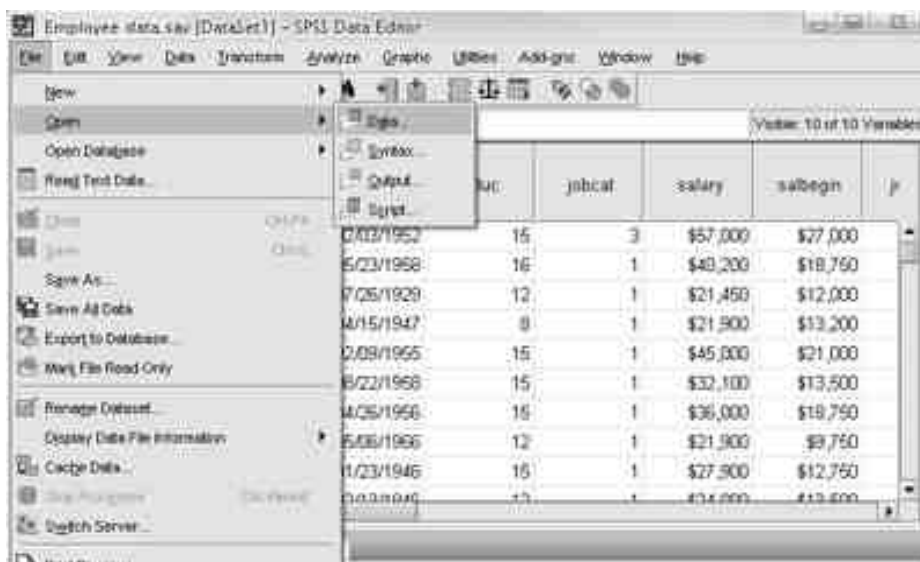
- This analysis shows the maximum, minimum, mean, and standard deviation of the variables

Linear regression analysis

- Linear Regression estimates the coefficients of the linear equation

Opening the sample data

- Open 'Employee data.sav' from the SPSS
- Go to "File," "Open," and Click Data



- Go to Program Files," "SPSSInc," "SPSS16," and "Samples" folder.
- Open "Employee Data.sav" file

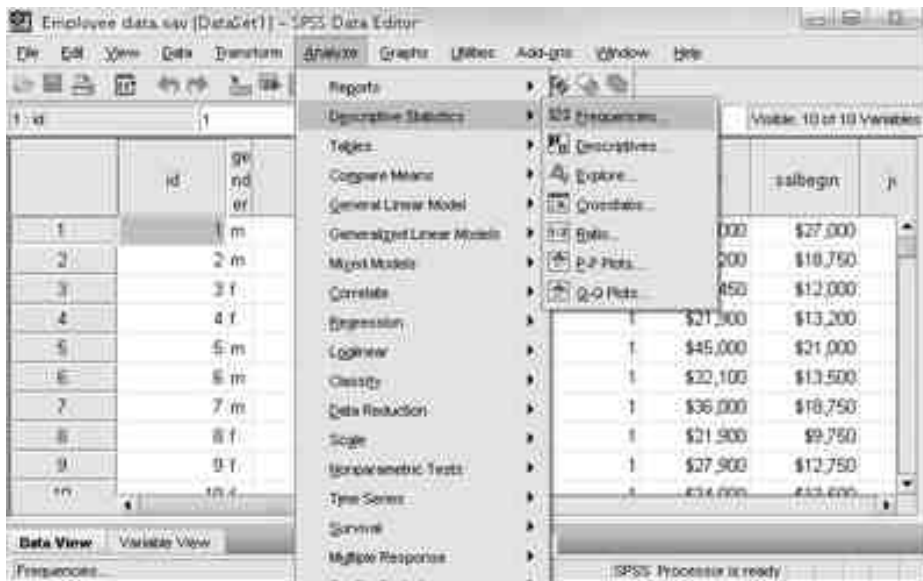


Topic No. 85

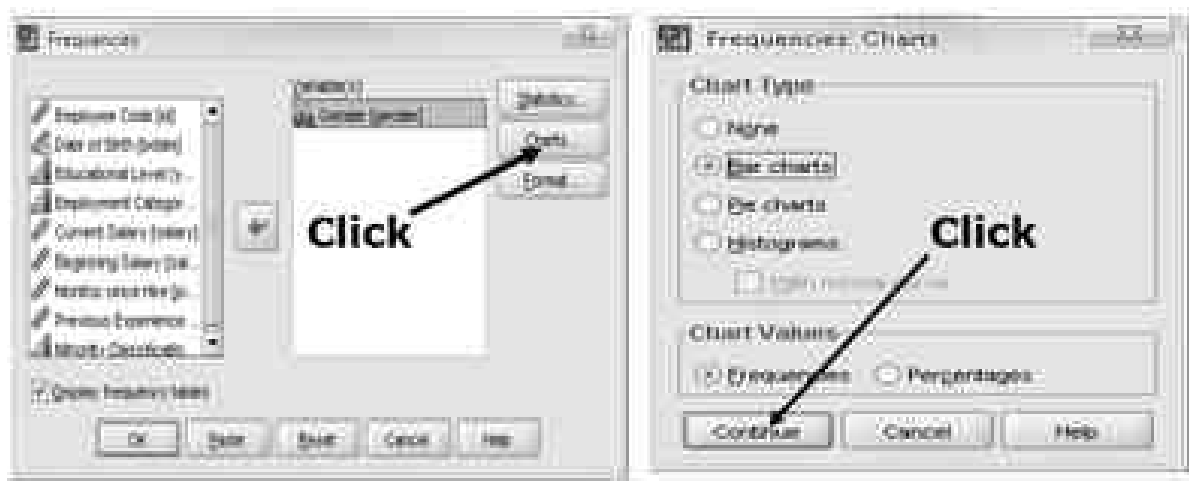
A Tutorial to SPSS - X

Frequencies

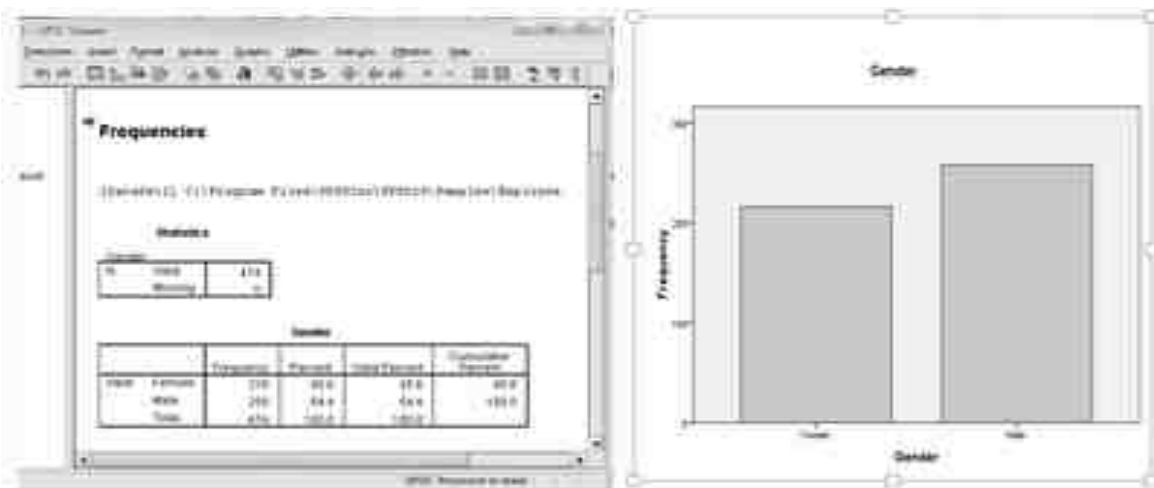
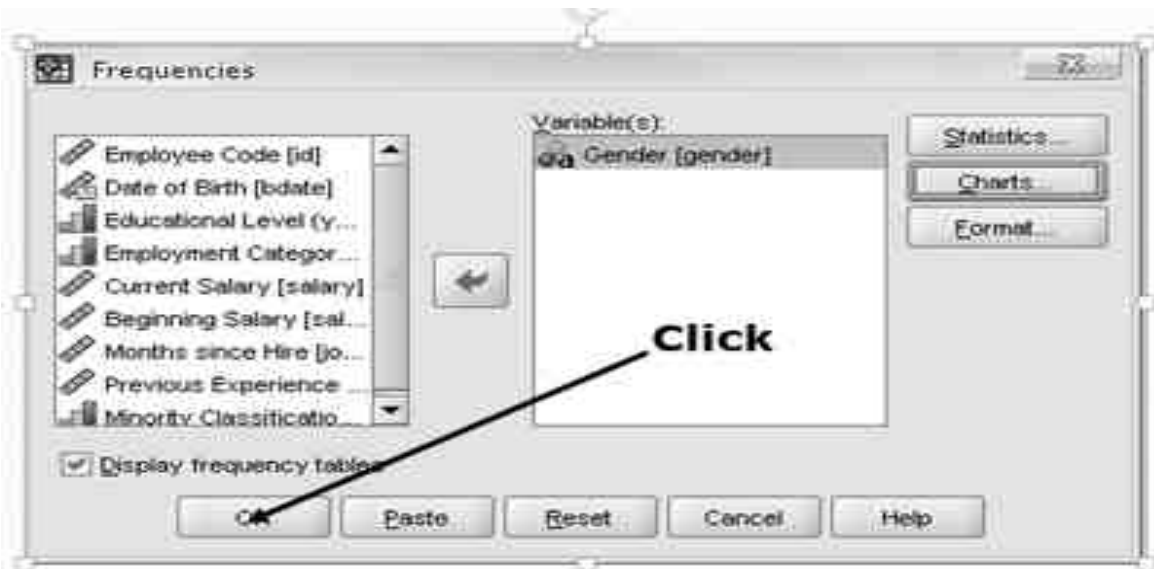
- Click 'Analyze,' 'Descriptive statistics,' then click 'Frequencies'



- Click gender and put it into the variable box.
- Click 'Charts.'
- Then click 'Bar charts' and click 'Continue.'



- Finally Click OK in the Frequencies box.



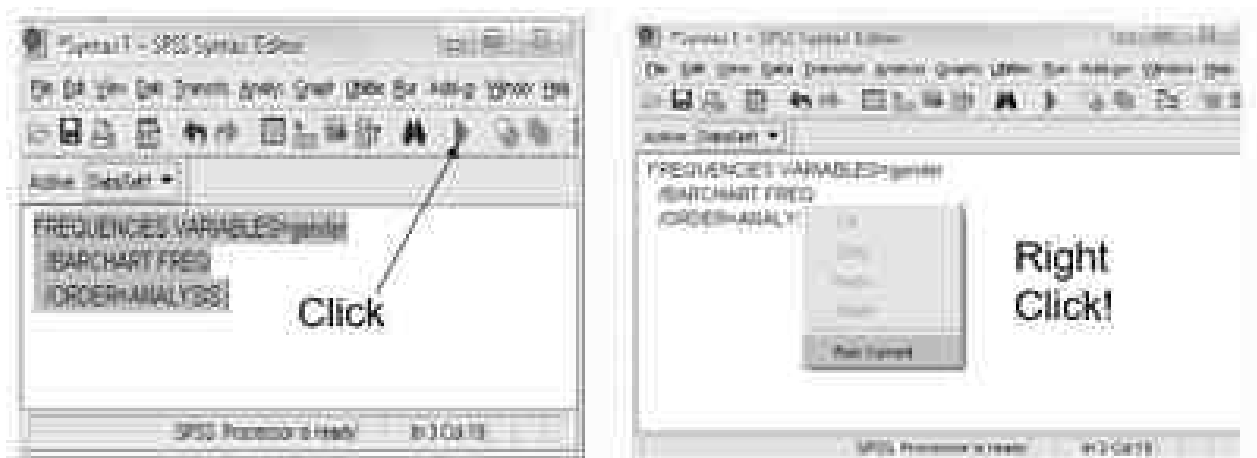
Topic No. 86

A Tutorial to SPSS - XI

Using the Syntax editor

Create SPSS program command lines by creating a new syntax editor window, and key in an SPSS program. To do this, go to the File Menu, go down to “New”, and select “Syntax”. A new syntax editor window is now created. You can enter your own SPSS program.

- Highlight the commands in the Syntax editor and then click the run icon.
- You can do the same thing by right clicking the highlighted area and then by clicking ‘Run Current’



Practice 4

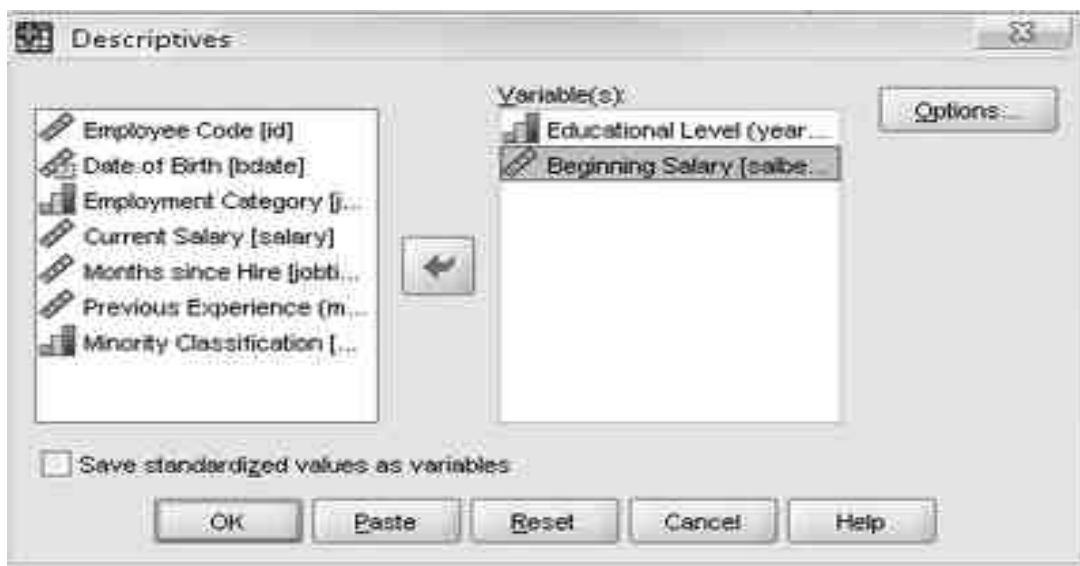
- Do a frequency analysis on the variable “minority”
- Create pie charts for it
- Do the same analysis using the syntax editor

Topic No. 87

A Tutorial to SPSS - XII

Descriptive:

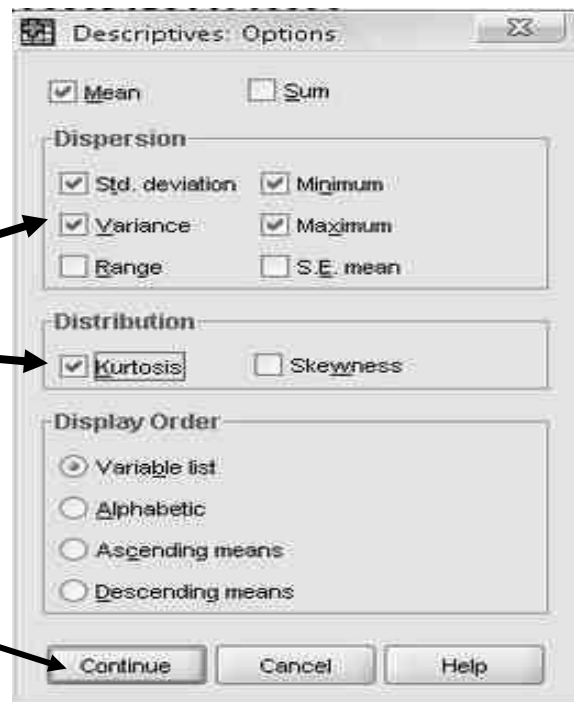
- Click 'Analyze,' 'Descriptive statistics,' then click 'Descriptives...'
- Click 'Educational level' and 'Beginning Salary,' and put it into the variable box.
- Click Options



- The options allows you to analyze other descriptive statistics besides the mean and Std.
- Click 'variance' and 'kurtosis'
- Finally click 'Continue'

Click

Click



- Finally Click OK in the Descriptives box. You will be able to see the result of the analysis.

SPSS (Windows) - SPSS Viewer

File Edit View Data View Variable View Output View Windows Help

Output

- Descriptives
- Frequencies
- Crosstabs
- Cluster Analysis
- Regression
- Model Diagnostics
- File (File)
- Descriptives
- File (File)
- File (File)

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Kurtosis	Skewness
Education at Level (years)	474	0	21	13.48	3.895	15.172	-1.394	.334
Beginning Salary	474	10,000	87,500	31,214.38	17,879.836	319,867.7	11.586	.274
Level of Education	474							

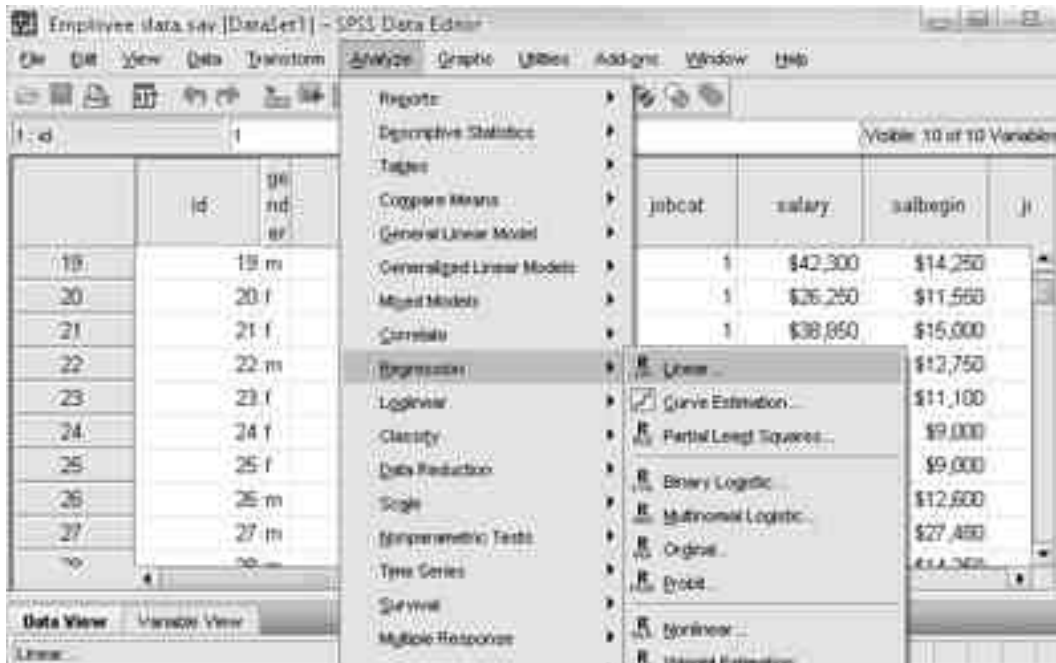
SPSS Viewer - Descriptives

Topic No. 88

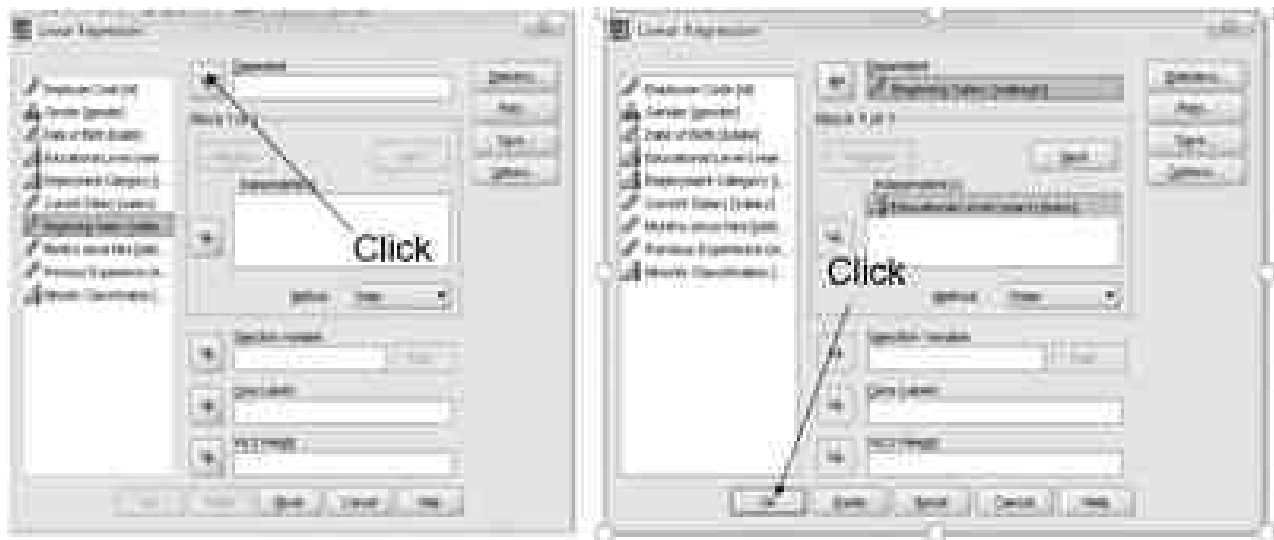
A Tutorial to SPSS - XIII

Regression Analysis

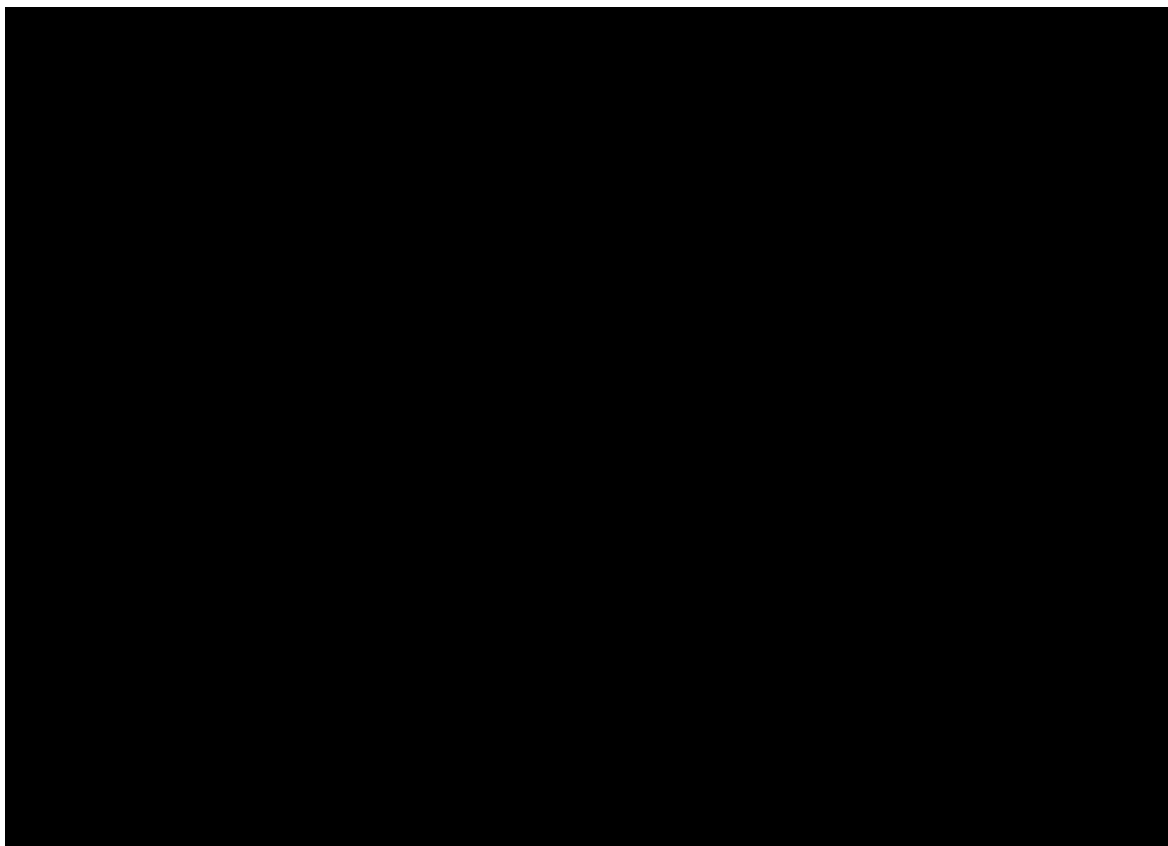
- Click 'Analyze,' 'Regression,' then click 'Linear' from the main menu.



- For example let's analyze the model $\text{salbegin} = \beta_0 + \beta_1 \text{edu} + \varepsilon$
- Put 'Beginning Salary' as Dependent and 'Educational Level' as Independent.

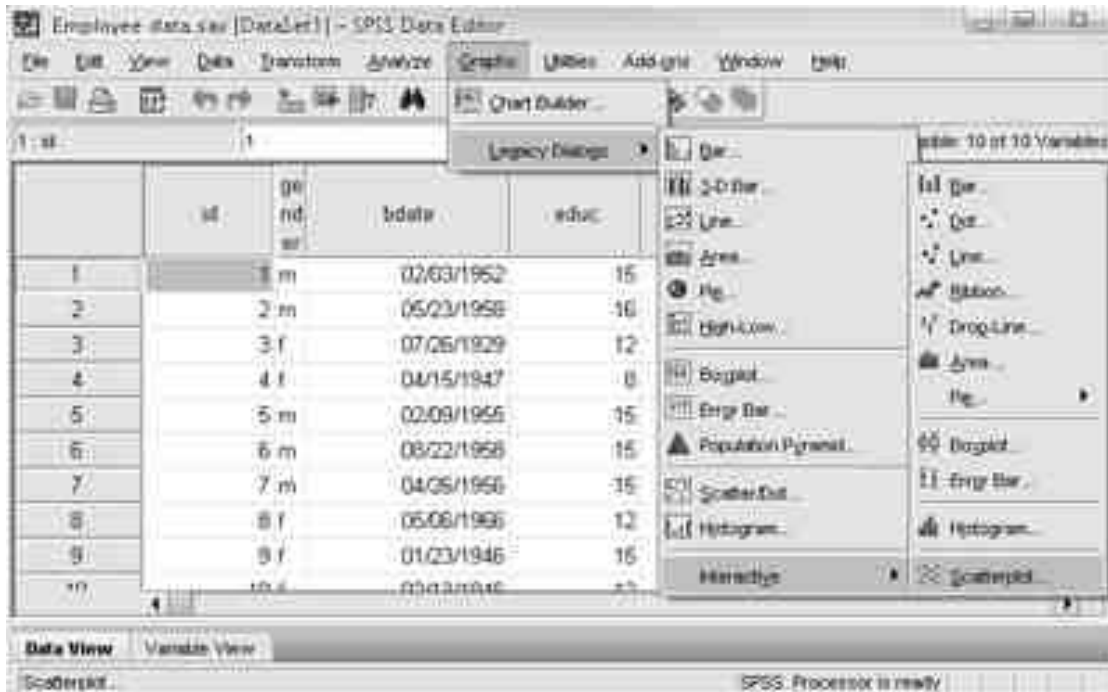


Clicking OK gives the result



Plotting the regression line

- Click 'Graphs,' 'Legacy Dialogs,' Interactive, and 'Scatterplot' from the main menu.



- Drag 'Current Salary' into the vertical axis box and 'Beginning Salary' in the horizontal axis box.
- Click 'Fit' bar. Make sure the Method is regression in the Fit box. Then click 'OK'.

