Department of Distance and Continuing Education

University of Delhi

दूरस्थ एवं सतत् शिक्षा विभाग

दिल्ली विश्वविद्यालय

B.A. (Hons.) Economics

Semester-I

Discipline Specific Course (DSC-3)

Course Credit-4

INTRODUCTORY STATISTICS FOR ECONOMICS

As per the UGCF-2022 and National Education Policy 2020

ISBN : 978-93-95774-84-0

1st edition: 2022

E-mail: ddceprinting@col.du.ac.in
economics@col.du.ac.in

Published by:

Department of Distance and Continuing Education under
the aegis of Campus of Open Learning/School of Open Learning,
University of Delhi, Delhi-110 007

Printed by:
School of Open Learning, University of Delhi

- Corrections/Modifications/Suggestions proposed by Statutory Body, DU/Stakeholder/s in the Self Learning Material (SLM) will be incorporated in the next edition. However, these corrections/modifications/suggestions will be uploaded on the website *https://sol.du.ac.in*. Any feedback or suggestions can be sent to the email- *feedbackslm@col.du.ac.in*

Discipline Specific Course – 3
INTRODUCTORY STATISTICS FOR ECONOMICS
Study Material : Lesson 1-11

TABLE OF CONTENT

About Contributors

Contributor's Name	Designation
Suramya Sharma	Guest Faculty, NCWEB, Hansraj College, University of Delhi, Delhi
Dr. Pooja Sharma	Associate Professor, Daulat Ram College, University of Delhi
Taramati	Guest Faculty, Kirori Mal College, University of Delhi
Sugandh Kumar Choudhary	Assistant Professor, Department of Economics, S.S.Khanna Girl's Degree College, University of Allahabad
Tasha Agarwal	Ph.D. Scholar, Ambedkar University, Delhi

LESSON 1

INTRODUCTION TO POPULATION AND SAMPLE

STRUCTURE

1.1. LEARNING OBJECTIVES

After reading this lesson, students will be able:

1. To gain thorough understanding of the meaning, importance, and application of statistics
2. To distinguish between quantitative and qualitative data

3. To learn about various scales of measurement viz. ratio, interval, ordinal, and nominal scales
4. To identify and comprehend the differences between statistical population and samples
5. To distinguish between various sampling techniques and
6. To demonstrate the knowledge of various branches of statistics through descriptive and inferential statistics

1.2 INTRODUCTION

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." Samuel S. Wilks (1906 - 1964).

Over the past many decades, statistics have become an indispensable part of our lives. We often come across statistics in one form or the other. If we turn our newspapers or televisions, we can definitely find some surveys that establish relationship between particular issues, say, eating fast food and the risk of having a health issue; or we may find graphs depicting growth rates or changes in some variables overtime, say growth rate of GDP or inflation level in a country. But the scope of statistics is not just limited to data collection and representation, but it establishes a basis for decision making and problem solving. We can create models to not only study the past trends but can also extend them to study the uncertain future. Statistics help us to make informed decisions in a world of uncertainty and variability. We would not have needed statistics had there been no uncertainty and variability around us.

The word **Statistics** is derived from a Latin word, "Status," which means "a group of numbers or figures that represent some information of human interest." Statistics may be formally defined as "the study of collecting, organizing, analyzing and interpreting information in the form of data."

Statistics helps us gain valuable insights not only in the Economics discipline, but is also popular amongst finance scholars, engineers, medical researchers and other science and social-science disciplines. In the financial sector, statistical analysis may be used at the micro and macro level. At micro level, it facilitates understanding of a company or business' performance like determining the revenue generating capacity, relationship between advertising and sales, etc. Whereas at macro level, statistical analysis allows a country to assess its financial condition and measure economic growth. In the field of engineering, statistics is an indispensable part for robust analysis, probability risk assessment, measurement of error etc. Statistics allow clinical researchers to compare various medical treatments, evaluate the benefits of alternate therapies, establish optimal treatment combinations, etc. Since the scope of statistics has broadened and is now used in a number of practical fields, it is also referred to as **applied statistics**.

When we talk about the **nature of statistics**, it is considered as a science as well as an art. It is a science since the statistical techniques are systematic and have broad application. In several instances, statistics are used to study cause and effect relationships and the results can be generalized in the same way as any scientific experiment or law. Statistics is also an art which refers to the "skill of handling facts so as to achieve a given objective." Managing, presenting, and drawing relevant conclusions from data is considered an art.

1.3 TYPE OF DATA

We've learnt that statistics is the study of collecting, organizing, analyzing and interpreting data. But you may wonder, what exactly is data? **Data** is nothing but pieces of raw information, facts and figures that are used for analysis. Broadly, data can be of two kinds:

1.3.1 Quantitative data

1.3.2 Qualitative data

1.3.1 Quantitative data

As the name suggests, all kinds of numerical data that can be measured comprise quantitative data. Information like age, height, distance, income, saving, GDP, imports, exports, rate of employment, etc. are quantitative data.

Now quantitative data can further be classified into two types:

i. **Discrete data** – The data which can only take specific values are termed as discrete data. For example, age of respondents- this variable can assume only whole numbers. For instance, the number of computers in a school can only be whole numbers. We will never witness 7.2 or 15.7 computers in a school. Similarly, the number of students in a class- this variable too can only take whole numbers. We never observe 39.7 or 45.2 students in a class. These are examples of discrete data. So discrete data generally takes only whole numbers, is finite and countable.

ii. **Continuous data** – The data which can take any value between an interval is referred to as continuous data. This means that such data can take up any value between two numbers. For example, daily temperature recorded in an Indian city in degree Celsius- this variable can take any value in the range of -50° to 50°. Here, 4.7° or 42.3° are acceptable observations. Similarly, the daily income of an ice-cream seller- here too Rs. 346 or Rs. 1787.5 are suitable values. So continuous data can take decimal values, is infinite and may not be countable.

1.3.2 Qualitative data

On the contrary, any information which is not directly measurable is known as qualitative data. Qualitative data represents the qualities or the characteristics of data. Information like political ideology, physical attributes of a person, problems faced by workers etc are examples of qualitative data.

Scales of measurement:

i. **Nominal scale data** – The information which cannot be sorted or put in any order is known as Nominal. Such data are individual set of information where changing the order of the information does not change any meaning. For example, occupation of respondents- the values may range from teacher, farmer, shopkeeper, unemployed etc. These data cannot be measured, nor can they be sorted in any way. Another example of nominal data may be the marital status of respondents. The variable may take values like married, unmarried, divorced, widow etc. Changing the order of the responses does not make any difference in the understanding of the sample

ii. **Ordinal scale data** – In contrast, ordinal data follows a natural order. Although these too cannot be measured explicitly, we can sort the data or order them in a way to observe basic comparison between values. For example, the education level of respondents- such a variable can take values from nursery to post graduation and the data has a specific order. Opinion of respondents towards relevance of CCTV cameras in workplaces could take values like strongly agree, somewhat agree, somewhat disagree, strongly disagree, etc. These values can also be arranged in a particular order.

iii. **Interval scale** – This scale pertains to numerical data which possesses the property that differences in values represent the real differences in the variable. With such variables, we know that not only one value is greater than the other but that the distances between the intervals on the scale are the same. For example, the temperature in Fahrenheit or Celsius, year of birth, etc. Here a temperature of 92°F is greater than 90°F and also the difference between 92°F and 90°F would be same as the difference between 90°F and 88°F.

iv. **Ratio scale** – The data belonging to ratio scale is a quantitative measurement with labels and orders the variable with evenly spaced intervals between values. These scales have a real absolute zero representing the total absence of the variable being measured. Hence ratio scale variables are exactly same as interval scale variables along with a "True zero." For example, weight of a commodity & height of a person, etc. Note that a zero value indicates that the commodity is weightless.

IN-TEXT QUESTIONS

1. Sort the following data into quantitative and qualitative data:

   A. Gender of respondents
   B. Number of lectures attended
   C. Percentage marks obtained in Economics subject
   D. Revenue in lakhs
   E. Whether interested in buying a washing machine

2. Mark whether the statements are true or false:
   A. Nominal data can be arranged in a particular order
   B. Amount of time taken to complete a class project is a continuous variable
   C. Statistics helps us to make informed decisions
   D. Qualitative data can be measured directly
   E. Discrete data are finite and countable

3. Match the following:

A.	Economic status: low, medium and high	1.	Discrete
B.	Weight of students in a class	2.	Continuous
C.	Employees in a company	3.	Nominal
D.	Religion: Hindu, Muslim, Sikh, Christian, other	4.	Ordinal

4. Fill in the blank:

   A teacher notes down the weight of each student in the class. The scale of measurement being used here is _____.

1.4 POPULATION, SAMPLE AND PROCESSES

We can also categorize data into **univariate, bivariate and multivariate data**. Uni means one and variate refers to variable. Hence univariate data consists of only a single variable. It is the simplest form of data. For instance, the number of cold drinks sold by a street vendor on weekdays. The data may look like as below:

	1	2	3	4	5
Sales (in Rs.)	2,500	2,700	2,000	3,200	3,800

In the above example, you may notice that we can only describe the data and any kind of relationship or comparison cannot be drawn. On the other hand, bivariate and multivariate data allow a researcher to establish relationships and correlations between variables. Here, bi means two and hence bivariate data involves two distinct variables. A researcher may use this data to establish a relationship between the two variables. For instance, the number of cold drinks sold by a street vendor and the daily temperature of the city the vendor resides in. They could look like:

	1	2	3	4	5
Sales (in Rs.)	2,500	2,700	2,000	3,200	3,800
Temperature (in °C)	34	37	36	37	38

A researcher may draw a conclusion that sales and temperature have a positive relationship since as temperature in the city rises, so does the number of cold drinks sell by the street vendor. Finally, multivariate data consists of more than two variables. Suppose a researcher wishes to analyze the determinants of cold drinks sales in a city. So, she gathers data on cold drinks sales, temperature of city, and price of cold drinks.

	1	2	3	4	5
Sales (in Rs.)	2,500	2,700	2,000	3,200	3,800
Temperature (in °C)	34	37	36	37	38
Price (in Rs.)	20	18	21	16	15

The researcher can identify several relationships between these variables.

Now to get reliable results from our analysis, it is crucial to understand that the data we use must be relevant and representative. To ensure the same, we need to know about populations and samples. The differences between them, and why is there a need to use a sample at all. We will finally discuss some of the commonly used sampling techniques.

1.4.1 Population

If we ask you, what do you understand by the term population, what will your answer be? Maybe you'll say that population is a group of individuals who live in a country or a state or even in a city. You may also say that population may not just be related to human beings, but the scope of population can be expanded to all living beings including animals and birds.

However, in statistics, the definition of population is slightly different. Here, **population** refers to all the individuals/entities possessing similar characteristics and belonging to a particular group under study. The population could vary from study to study. For instance, a researcher might be interested in analyzing the results of the Common Admission Test (CAT) examination in India. Here, the population would be comprised of all the candidates who appear for CAT in a particular year. Say in a year, about 2.3 lakh candidates appear for the exam. To study a population, the researcher would have to gather data for all 2.3 lakh candidates—no one from the population can be left out. The best example to understand the concept of a population is a census. In India, a census is the process of collecting and analyzing demographic, economic and social aspects of Indian residents. Since census is a study of the whole population, the data is collected from each and every Indian resident. You can imagine the scale at which the data collection is carried out that it takes about 10 years to collect, clean and publish the entire data. In statistics, the population depends on the topic and scope of research. Other examples of a population could be total number of people working under the NREGS, voter population in India, number of students enrolled in private schools in a state, total number of accidents taken place in India etc.

Studying a population helps a researcher to gain useful insights into the characteristics of all the elements under study. Since each and every unit is considered, the results are considered to be reliable and representative of the population. This also enables a researcher to study more than one aspect of the population and carry out an intensive study. Such a data can also form the basis of further investigations. However, studying a population is more suitable when we have a small scope of study.

The descriptive statistics taken from the population are termed as '**population parameter**' or simply a 'parameter'. So, a parameter describes the characteristics of a population. They are usually denoted by Greek letters such as μ (Mu) for mean and σ (Sigma) for standard deviation. Based on the data collected from the entire population of 2.3 Lakh candidates, if we want to say that the average marks a candidate scores in the CAT examination are 85, we could denote it as: $\mu = 85$.

1.4.2 Sample

As you may have identified, the major challenge of analyzing a population is that such research requires data to be collected from each and every member of the population, which is undeniably a tedious task. The possibilities of making errors in studying population is also significant since there may be missing data due to non-response by respondents or measurement complexities due to the large amount of data, etc. Gathering data from a population not only requires extra efforts, but also consumes a lot of time and is expensive. Constraints on scarce resources render a population survey unfeasible.

To overcome the drawbacks, a subset of the population, referred to as a sample, may be used instead. A **sample** is an unbiased subset of the statistical population which is representative of the entire dataset. This means that a researcher randomly draws and analyzes some observations from the population to make inferences about the whole population. The figure 1 depicts the relationship between population and sample:

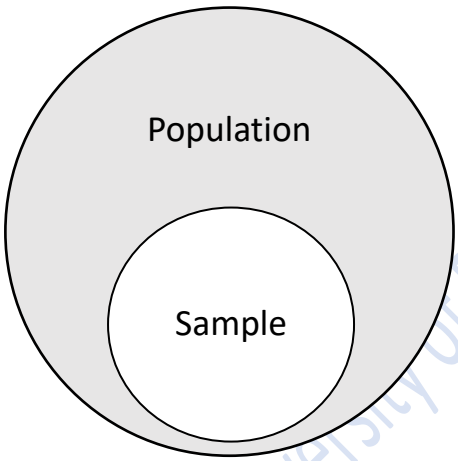


Figure 1: Sample is a subset of Population

For instance, consider the above example of a researcher who wishes to analyze the results of CAT examination, again. Now, instead of collecting data from all 2.3 lakh candidates, the researcher may select a subset from the population, say randomly selected 1000 candidates, and then perform the analysis. If the sample is well representative of the population, the researcher could generalize the results for all the 2.3 lakh candidates.

So, it is advisable to use a sample when:

a. The population is too large. For example, if a researcher wants to understand the relationship between level of inflation and unemployment in a country, then the researcher must collect data from each and every unemployed person in the country, which is not practical. Instead, the researcher may select a small sample from the population to understand the relationship at the country level.

b. The research is time sensitive. Suppose a researcher wishes to analyze the short-term sentiments of the population about the Covid-19 lockdown in a country. Clearly, it will take a lot of time to collect the information from the entire population. During that time, it is possible that the sentiments of the public change. Due to the time-consuming process of data collection, the research may not provide accurate results by the time it is completed. In such cases, using a sample is a time saving way to conduct research.

c. Data collection is expensive. Data collection from a population involves several expenses such as the cost of employing survey teams, computers/laptops,

transportation, stationery, and other miscellaneous expenses. On the other hand, samples are a cost-effective way to conduct research.

Corresponding to a parameter, the descriptive values taken from a sample are known as '**sample statistics** or just 'statistic'. So, a statistic describes the characteristics of a sample. They are denoted by Latin letters such as $\bar{x}$ ($x\ bar$) for mean and s for standard deviation. Now, based on the data collected from a sample of 1000 candidates, if we want to say that the average marks a candidate scores in the CAT examination are 86.8, we could denote it as: $\bar{x} = 86.8$.

The process of selecting a sample from the population is known as '**sampling**'. When selecting a sample from a population, deciding on the size of the sample is also important. A **sample size** is simply the number of observations we select out of a population as a sample. For instance, say we want to investigate about the work environment of a leading e-commerce company for its female employees. Assume that the total number of full-time and part-time female employees working for the company is 2 Lakh. So, this will be our study population. Since we already know that it is difficult to conduct research based on the whole population, we should draw a sample for our study. Now the sample size that we select, could either be:

- Too small. What if we select a sample of only 50 female employees as our sample and conduct our research based on their experience? There are very high chances that our results will be biased and unrepresentative of the whole population; or
- Too big. The other extreme could be if we select a sample of 1 lakh female employees. Since this sample size is quite large, we will get more accurate results. However, the process of collecting data will be as complex, time-consuming and costly as it would be if we studied the whole population.

Hence, we now understand the importance of appropriate sample size in a study. We will not study the variables and formula required to calculate the sample size since it is out of the scope of this unit.

The following table summarizes the concepts of population and sample in a tabular form:

	Population	**Sample**
Definition	Set of all items or observations that possess common characteristics	Subset or a part of population that is representative of population
Characteristics	Parameter	Statistics
Symbols	Population size = N	Population size = n

	Population mean = μ (Mu) Population standard variance = σ (Sigma)	Population mean = $\bar{x}$ (x bar) Population standard variance = s
Data collection	Census	Sampling
Advantages	1. Results are representative of population 2. Intensive study 3. Suitable for small universe	1. Convenient 2. Low cost and less time consuming 3. Better accuracy if sample is representative
Disadvantages	1. Time consuming and costly 2. Possibility of errors	1. Possibility of bias 2. Difficulty in selecting a representative sample

IN-TEXT QUESTIONS

5. Select the correct option and fill in the blank:

   Suppose a researcher wishes to compare the popularity of five advertisements on a website, and gather data for the click rates for teens, adults and elderly. The data collected in this case is _____ data. (Univariate / Bivariate / Multivariate)

6. Mark whether the following are parameter or statistic:

   A. $s = 5.7$
   B. $\mu = 120$
   C. $\sigma = 32$
   D. $\bar{x} = 18$

7. Select the correct option:

   I. Sample is used when:

   A) Data collection is inexpensive
   B) Research is time sensitive
   C) Population is small
   D) Population is unknown

   II. A mean is called a statistic if it is calculated from the:

   A) Sample
   B) Population
   C) Parameter
   D) Standard deviation

1.5 SAMPLING TECHNIQUES

To ensure that the samples drawn are representative of the population, it is crucial to understand the different ways in which we can select a sample. The two broad methods of sampling are:

1.5.1 Probability sampling techniques – It is one of the commonly used sampling techniques where each unit of population has an equal chance (or probability) of getting selected in a sample. This means that the samples selected are random and unbiased and hence are representative of the population. These techniques are also known as **random sampling** techniques. The five techniques under probability sampling are:

a. **Simple random sampling** – As the name suggests, it is the most basic and crude form of random sampling. Here each unit of population has an equal and independent chance of being chosen. Example: in a lottery system, the names of each unit of population are written on a chit and after thorough shuffling, the researcher picks the chits one by one and notes down the names. Another way of simple random sampling is through random number generation where all the units of population are assigned a number in sequential order. Then random numbers are generated using software and sample selections is carried out.

b. **Systematic random sampling** – Under systematic random sampling, the sample set is selected from the population in a fixed interval. This technique is more feasible than simple random sampling. To draw samples using systematic random sampling, the first step is to arrange the units of population in an order and assign a number to each unit from 1 to N. Then the sampling interval is calculated using the formula: $K = \frac{N}{n}$, where K is the interval, N is the population size and n is the sample size. Finally, we select one unit at random and then select following units at equal interval K. For example, suppose we have to collect information from residents of a city where the houses are numbered from 1 to 100,000. So, if the size of the population is 100,000 and we need a sample size of 1000 houses, then the interval should be 100,000/ 1000 = 100. The researcher will choose one house at random and then will select every 100th house thereafter to get the sample.

c. **Stratified random sampling** – The above two types of sampling techniques assume that the population is homogeneous. In case the population is heterogenous, we use the Stratified random sampling technique. Under this technique, we divide the population into sub-groups based on homogeneous characteristics such as gender, age group, income level, etc., called strata, and then select random samples from each sub-group or stratum. For instance, if a company has 700 male employees and 300 female employees, then simple and systematic random sampling technique may give us biased samples. To avoid the bias, we use stratified random sampling wherein we create two

groups based on gender and then select random samples from both groups. The technique has further two approaches to select a sample: Proportionate stratified sampling and disproportionate stratified sampling.

d. **Multi-stage sampling** – At times the population of interest is quite large and geographically diverse. In such cases, one sampling technique is not enough to select a sample and using multi-stage sampling is suitable. Here, a sample is selected in stages, combining different sampling techniques as described above. For instance, to study the issues faced by primary school children, a researcher may first divide the population into states, then use simple random sampling to create a sample of states. Next, the researcher could again use simple random sampling to select a few districts, and finally use systematic random sampling to identify a few schools within a district. The multi-stage sampling method is frequently used to scale down large data sets into workable sizes. Although there is no restriction on the number of stages you could use to select a sample, it is important to note that all the sampling techniques used must be probability sampling methods.

1.5.2 non-probability sampling techniques – Under non-probability sampling or **non-random sampling** techniques, each unit of population does not have an equal chance of getting selected in a sample, that is, the samples are not random—they may be biased and may not represent the population accurately. Hence, non-probability sampling techniques may not produce results that can be generalized. Yet, there are many occasions when non-probability sampling methods are preferred over probability sampling methods, which will be discussed in the following passages. The most commonly used non-probability sampling techniques are:

a. **Convenience sampling** – As is clear from the term itself, convenience sampling refers to the sampling technique in which a researcher collects the data from convenient sources. For instance, as part of the undergraduate course, a student undertakes a research project in which she tries to understand the consumer sentiments related to rooftop solar panels. Considering the time, cost and effort constraints, the student may choose to collect data from the most convenient location to her, it may be within the city she resides in or within the district. It is worth noting that such research may not be representative of the population and generalization of the results may not be appropriate. That said, such a technique is useful when a researcher carries out a pilot survey to test the questionnaire.

b. **Judgement sampling** – Under this technique, the researcher selects a sample based on her own judgement about the characteristics of the individuals. In other words, the researcher uses her expertise to identify the best fir for her sample. Take the example of a researcher who is interested in understanding the challenges faced by disabled employees in a company. In such a case, the researcher can easily identify her sample by using her sense of judgement about the characteristics of disabled persons.

c. **Snowball sampling** – At times it is difficult to locate the appropriate target group for a study. In such cases, snowball sampling techniques may be used. This is generally used in social science research. Under snowball sampling, the existing respondents are asked to identify or suggest other individuals who are well-suited for the research. So based on the references, the sample size keeps on increasing—just like a snowball. For example, a researcher may want to understand the plight of immigrants in a city. Since there may not be any official record of immigrants readily available, the researcher could identify few immigrants and then ask them to help locating other immigrants for the study. Such a technique comes in handy when we're interested in researching hard-to-find- groups. However, the risk of achieving biased results is quite high since the initial respondents may refer to their friends and family, who share common characteristics and beliefs.

The following figure summarizes all the sampling techniques we have discussed:

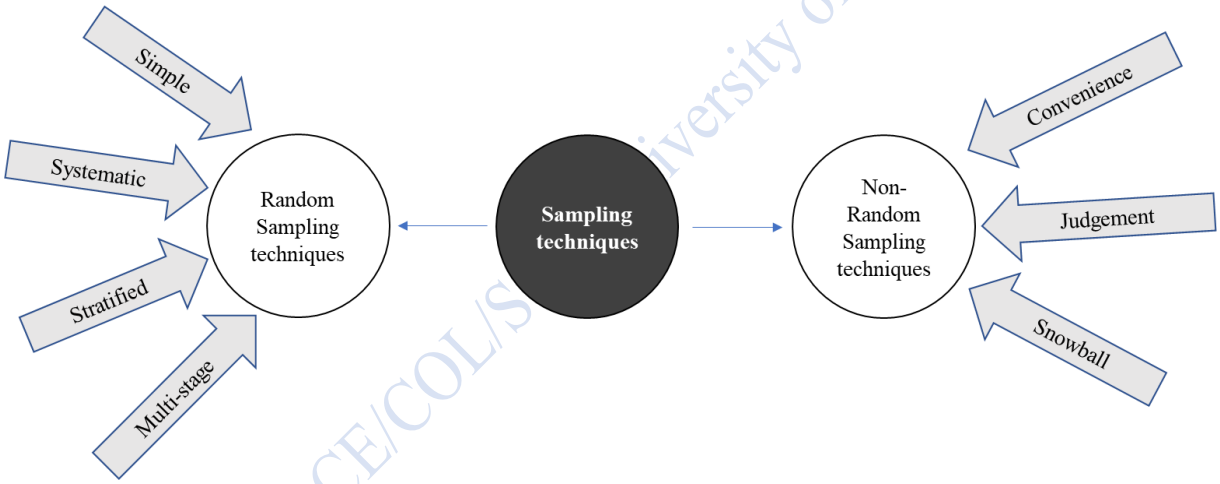


Figure 2: Sampling techniques

IN-TEXT QUESTIONS

8. What type of a sampling technique is being used in the following examples:

   A. A manager wants to select a sample of their clients to ask for donation. She arranges the list of clients in alphabetical order and randomly selects the first client. She then proceeds to select every 5^{th} client from the list.

   B. A news reporter gathers consumer sentiments regarding a government policy by interviewing people on the street.

C. A research student asks the respondents to identify other potential research participants.

D. A researcher writes the names of different states in India on separate chits and put them in a bowl. She then selects a chit without looking to get a sample of 7 states.

9. A sampling technique that does not involve probability is known as _____.

10. Which of the following is not an example of Random sampling:

A) Simple random sampling B) Stratified sampling

C) Judgement sampling D) Systematic sampling

1.6 BRANCHES OF STATISTICS

A researcher may apply statistics to simply summarize and describe the characteristics of data or employ statistics to draw some conclusions or inferences from the data. Vast number of research apply two types of Statistical analysis:

1. Descriptive statistics
2. Inferential statistics

1.6.1 Descriptive Statistics

As the name indicates, **descriptive statistics** are basically used to 'describe' the data. It refers to different techniques of summarizing and displaying data. This includes communicating the patterns in the data or conveying the summary of data through graphs, tables, numerical or simple charts. Calculating the measures of central tendency like mean, median and measures of variability such as range, standard deviation and variance, along with creation of histograms, bar chart, dot plots etc. constitute descriptive statistics.

It is important to note that we do not use descriptive statistics to draw any conclusions or generalize the results. It is used to simply state the situation as represented by the data collected.

For example, suppose a research institute gathers data from a village consisting of 100 households. The institute can use the descriptive statistic to learn the average education level of the household heads in that village. Let's say that the education level of the head of the household varies between illiterate to graduate and on an average, heads of the households have attained education till class 10. They can further study the relationship between the education level of the head of household and a household's savings. Let's assume that they find that the households in which the heads were at least graduates, saved more. In descriptive statistics, we

can only report the findings. It cannot be concluded, by merely looking at the descriptive statistics, that generally, in India, households with highly educated heads save more.

So, we see that descriptive statistics cannot be used to make general estimates or predictions.

Yet, descriptive statistics are extremely useful as they can provide a snapshot of the whole data in meaningful ways. It helps simplify large data and present it in visually attractive ways. We will learn about the various components of descriptive statistics in later chapters.

1.6.2 Inferential Statistics

On the other hand, **inferential statistics** is used to predict, estimate, and make other generalized approximations based on the data. It is usually based on sample data to draw conclusions and generalize the results to the larger population. Hypothesis testing and regression analysis are two examples of inferential statistics.

Inferential statistics is very popular among researchers since it allows them to gather limited data and extrapolate the results to a larger population. This saves them a lot of cost as well as time.

If we consider the above example again, the research institute now gathers data from 10 randomly selected villages in India with total observations equal to roughly 1000 households. Now the institute may generalize its results to state or national level through inferential statistics.

The figure 3 summarizes the branches of statistics we discussed:

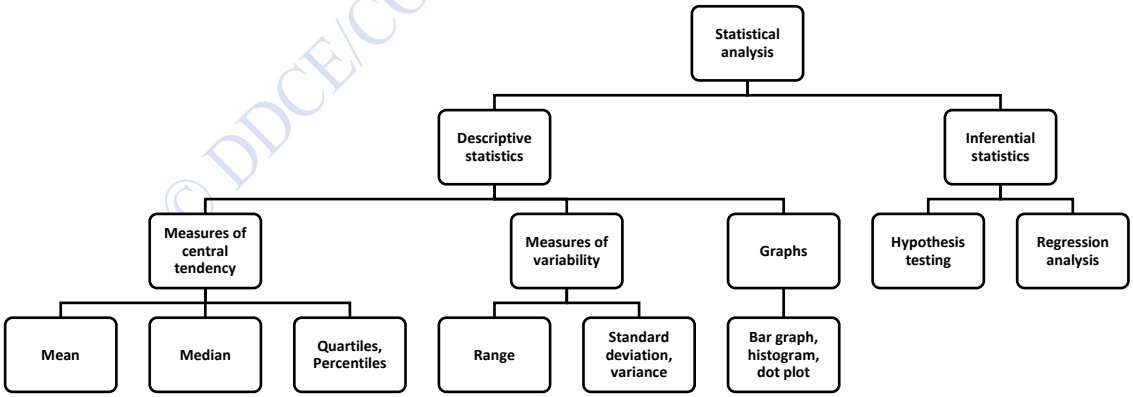


Figure 3: Branches of statistics

IN-TEXT QUESTIONS

11. Fill in the blanks:

    A. We can make predictions and estimations using _____ statistics.

    B. Histogram is an example of _____ statistics.

    C. Standard deviation is calculated as a part of _____ statistics.

12. Select the correct option:

    Inferential statistics can be used to

    A) Estimate B) Generalize

    C) Only A) D) Both A) and B)

1.7 SUMMARY

Statistics is defined as the study of collecting, organizing, analyzing, and interpreting information in the form of data. Since the scope of statistics has broadened and is now used in a number of practical fields, it is also referred to as applied statistics. It is both a science as well as an art. Data are pieces of raw information, facts and figures that are used for analysis. The two broad types of data used in statistics are- quantitative and qualitative data. Quantitative data is measurable in terms of some numbers, whereas qualitative data cannot be directly measured in numbers. quantitative data is further classified into discrete and continuous data whereas qualitative data can be categorized into nominal scale, ordinal scale, interval scale and ratio scale. Data can also be sorted into univariate, bivariate and multivariate data. Univariate data consists of only a single variable. It is the simplest form of data. While bivariate data involves two distinct variables, multivariate data includes more than two variables. Bivariate and multivariate data allow a researcher to establish relationships and correlations between variables. Population refers to all the individuals/entities possessing similar characteristics and belonging to a particular group under study, a sample is a subset of the population. When the descriptive statistics are taken from the population, they are termed as 'parameter' and descriptive values taken from a sample are known as 'statistic.' The process of selecting a sample from the population is known as 'sampling'. The sampling techniques are broadly classified into Probability sampling techniques and non-probability sampling techniques. Simple random sampling, systematic random sampling, stratified random sampling and multi-stage sampling are examples of probability sampling techniques. Whereas convenience sampling, judgment sampling and snowball sampling are examples of non-probability sampling techniques. The two branches of statistical analysis are descriptive and inferential statistics. Descriptive statistics is defined as one that describes the data through graphs, measures of central tendency and variability. On the other hand, inferential statistics utilize data to predict, estimate or make other generalized approximations.

1.8 GLOSSARY

- **Convenience Sampling:** Data is collected from convenient sources
- **Descriptive statistics:** It is used to describe the data through graphs, measures of central tendency and variability
- **Inferential statistics:** Used to predict, estimate, or make other generalized approximations.
- **Judgment Sampling:** Sample is selected based on researchers' judgement about the characteristics of the individuals
- **Multi-stage sampling:** A sample is selected in stages, combining different sampling techniques
- **Non-probability sampling:** Each unit of population does not have an equal chance of getting selected in a sample
- **Parameter:** The descriptive statistics taken from the population
- **Population:** All the individuals/entities possessing similar characteristics and belonging to a particular group under study.
- **Probability Sampling:** Where each unit of population has an equal chance (or probability) of getting selected in a sample.
- **Sample:** Unbiased subset of the statistical population which is representative of the entire dataset.
- **Sample Size:** Number of observations we select out of a population as a sample.
- **Sampling:** Process of selecting a sample from the population
- **Simple Random Sampling:** Each unit of population has an equal and independent chance of being chosen
- **Snowball Sampling:** Existing respondents identify other participants for the study
- **Statistic:** The descriptive statistics taken from the sample
- **Statistics:** Study of collecting, organizing, analyzing and interpreting information in the form of data.
- **Stratified Random Sampling:** Divide the population into sub-groups based on homogeneous characteristics and then select random samples from each sub-group
- **Systematic Random Sampling:** Sample set is selected from the population in a fixed interval

1.9 ANSWERS TO IN-TEXT QUESTIONS

1. A) Qualitative B) Quantitative C) Quantitative D) Quantitative E) Qualitative
2. A) False B) True C) True D) False E) True
3. A = 4; B = 2; C = 1; D = 3
4. Ratio
5. Multivariate
6. A. Statistic B. Parameter C. Parameter D. Statistic
7. I. B) Research is time sensitive
   II. A) Sample
8. A. Systematic random sampling
   B. Convenience sampling
   C. Snowball sampling
   D. Simple random sampling
9. Non-Probability sampling
10. C) Judgement sampling
11. A) Inferential B) Descriptive C) Descriptive
12. D) Both A) and B)

1.10 SELF-ASSESSMENT QUESTIONS

Q.1 Define Statistics. Is it science or art? Justify your answer.

Q.2 Give examples of a possible sample size of 4 from each of the following populations:

a. All news channels aired in India
b. All students in your university
c. All fast-food chains operating in India
d. Income of residents of India
e. Marks obtained out of 100 by first semester Statistics students

Q.3 What is the difference between a parameter and a statistic? Identify parameters and statistics in the following hypothetical cases:

a. A dietician wants to compute the average number of sweets consumed by children under the age of 7 within a month. From a random sample of 50 children, she discovers a mean of 59 sweets a month. Whereas when she gathered the population data, she came to know that the mean number of sweets consumed was actually 65.

b. On the occasion of National Milk Day, the government wants to estimate the number of cows a dairy farmer owns in a particular state. Using the census approach, the government finds that about 40 lakh households are engaged in dairy farming in the state with the average number of cows owned equal to 22. When a government official took a random sample of 100 households in a district engaged in dairy farming, she found out that the average number of cows owned equaled to 37.

Q.4 In a college, parking on the premises has become a major problem. To deal with the problem, the college administration wishes to compute the average parking time of the students who park their vehicles in the college parking lot. One of the college officials quietly follows 150 students and records the duration of time the students keep their vehicle in the parking lot.

a. Identify the population of interest to the college administration.

b. What is the sample size that the college administration is examining?

Q.5 Briefly describe any two methods of drawing non-probability samples.

Q.6 Why are descriptive statistics used? Can we use descriptive statistics to make generalized predictions based on the data? If not, then how can we do so?

1.11 REFERENCES

- Devore, J. L. (2016). *Probability and Statistics for Engineering and the Sciences*. Cengage learning.
- Larsen, R. J., & Marx, M. L. (2012). *An introduction to mathematical statistics and its applications*. Prentice Hall.
- McClave, J. T., Benson, P. G., & Sincich, T. (2018). *Statistics for business and economics*. Pearson Education.

1.12 SUGGESTED READING

- Gupta, S. C. (2019). *Fundamentals of statistics*. New Delhi, India: Himalaya publishing house.

LESSON 2

PICTORIAL METHODS IN DESCRIPTIVE STATISTICS

STRUCTURE

2.1 LEARNING OBJECTIVES

After reading this lesson, students will be able:

1. To understand the different types of graphical techniques
2. To comprehend the advantages and disadvantages of various graphical techniques and
3. To gain proficiency in creating various types of graphs

2.2 INTRODUCTION

We've mentioned in the earlier chapters that descriptive statistics involve the computation of the basic statistics of the data such as mean, median, standard deviation etc. These give a basic idea about the distribution of the dataset. Visual representation of the data is also an integral part of descriptive statistics. In this chapter we will take a closer look at the most common graphic methods to present the data.

We will learn about the following graphical techniques:

1. Stem and Leaf plot
2. Dot plots
3. Bar charts
4. Histograms

2.3 STEM AND LEAF PLOT

A stem and leaf plot are a convenient way to visualize continuous data. The plot can easily be constructed by hand and gives an overview of the distribution of the observations in the data at first glance. To create a plot, the data is arranged in an order and divided into equal intervals. We then create a table that presents the whole data set in two columns. The values of the data are split into two – stem and leaf. The first column—referred to as the stem—includes the tens, hundreds or thousands unit, as per the values of the data and the second column—known as the leaf—contains the rest of the digits. The concept will become clear with the following example.

Say a researcher has collected data of the weight of 10 college students, chosen at random. The weights, in kgs, are as follows:

$$71, 43, 66, 52, 59, 83, 92, 67, 74, 61$$

The first step in creating a stem and leaf plot would be to arrange the data in ascending order:

$$43, 52, 59, 61, 66, 67, 71, 74, 83, 92$$

Now, we create a table with two columns- Stem and Leaf, with the Stem column consisting of the tens digits and the leaf column containing the unit's digit. The table will look like this:

Stem	Leaf
4	3
5	2 9
6	1 6 7
7	1 4
8	8
9	2

Here, the first row with stem of 4 and leaf of 3 denotes the weight of 43kg and the last row with stem of 9 and leaf of 2 denotes the weight of 92kg. Simply looking at the table, we can work out that on average, the college students have a weight in sixties. We can observe that the shape of the display gradually rises, peaks at 6 and then steadily declines. We call this a bell-shaped curve which is symmetric in shape. We will learn about other features of a symmetrical distribution later in the unit.

Let us consider another example in which we have a data set consisting of the number of hours of daily sleep 100 students who are in college get. The stem and leaf display looks like this:

Stem	Leaf
5	4455567889
6	00223446667899
7	000112233566789999
8	00000011222234445577778889
9	0001333355777779
10	01344446899
11	12247

Note that here, the first row with stem of 5 and leaf of 4 denotes 5.4 hours of sleep and the last row with stem of 11 and leaf of 7 denotes 11.7 hours of sleep. We can clearly observe that most of the students get about 8 hours of sleep. If we ask you how many students sleep for more than 10 hours a day, then you have to simply add the number of leaves written in front of stems 10 and 11. So the answer will be 16 students.

A stem and leaf plot are helpful to get a basic understanding of the dataset, however it gets difficult to create a chart when the number of observations increase.

IN-TEXT QUESTIONS

1. The correct list of data for the following stem and leaf plot is:

Stem	Leaf
0	3
3	27
5	119
7	0

A. 03, 32, 37, 11, 51, 59,70 B. 03, 27, 37, 51, 51, 59,70

C. 03, 32, 37, 51, 51, 59,70 D. 03, 32, 37, 51, 59,70

2. The following stem-and-leaf plot depicts the number of cakes that a home-baker sells each week. If 1|7 represents 17 cakes, then,

Stem	Leaf
0	689
1	024479
2	01136889
3	122

A. How many weeks did the home-bakers sell cakes?

B. How many weeks did they sell more than 25 cakes?

2.4 DOT PLOT

A dot plot is another simplified way to visualize the data in the form of dots representing each unit of observation. The dots are stacked over one another that represent the frequency of the value in our dataset. For example, suppose a researcher would like to know the number of vaccinated children in a city. To make the analysis simpler, she divides the city into 5 localities and collects the number of vaccinated children for each locality. The data collected is tabulated in the following manner:

Locality	**No. of vaccinated children**
1	6
2	1
3	3
4	11
5	8

The dot plot for the data would look like figure 1 where the height of each column denotes the frequency of the observation.

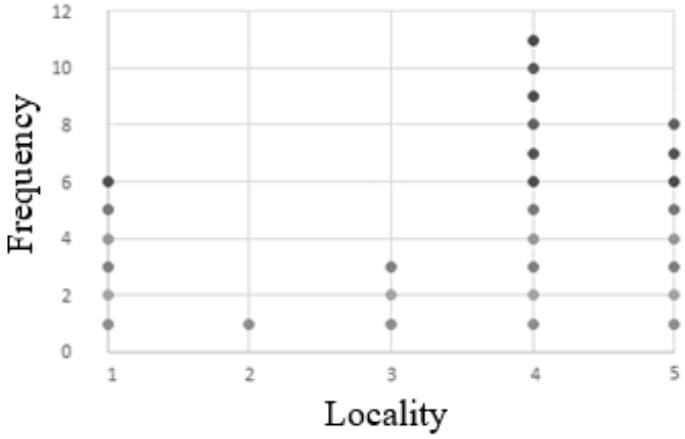


Figure 1: Dot plot of vaccinated children in a city

Dots plots can be created for continuous data as well. Consider male literacy level in 10 Indian states:

States	Male literacy Rate (%)
Andhra Pradesh	73.4
Assam	90.1
Bihar	79.7
Chhattisgarh	85.4
Goa	92.8
Gujarat	89.5
Haryana	88
Kerala	97.4
Meghalaya	77.1
Uttar Pradesh	81.8

Since the data is continuous and unique, we would have one dot for each state. Instead, to make the dot plots more informative, we create groups of data or class intervals.

Male literacy Rate (%)	No. of states
70-75	1
75-80	2
80-85	1
85-90	3
90-95	2
95-100	1

Now we can easily create the dot plot using the above table. Try making the dot plot on your own using the above table. Your dot plot should look something like figure 2:

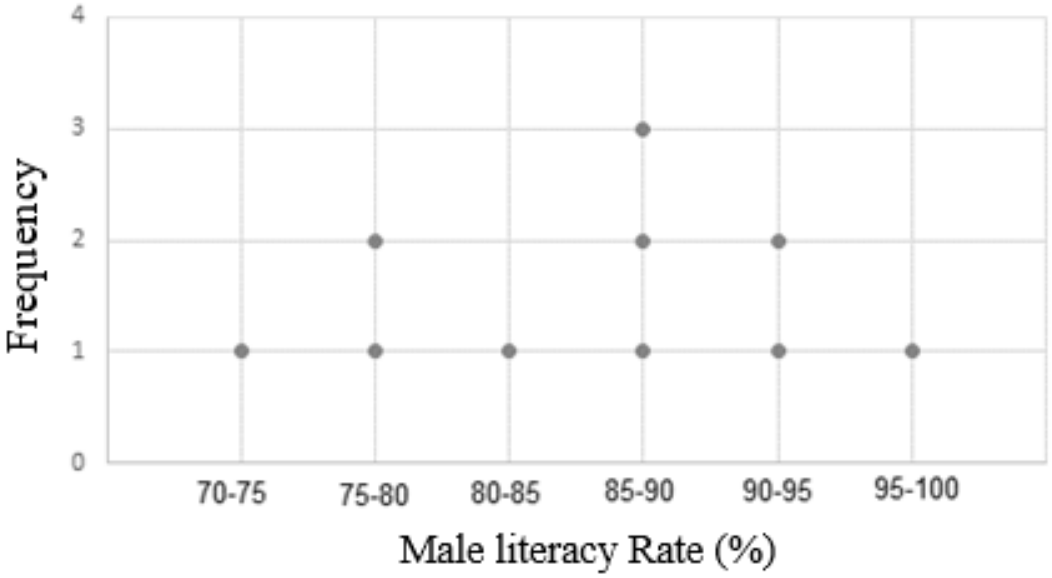
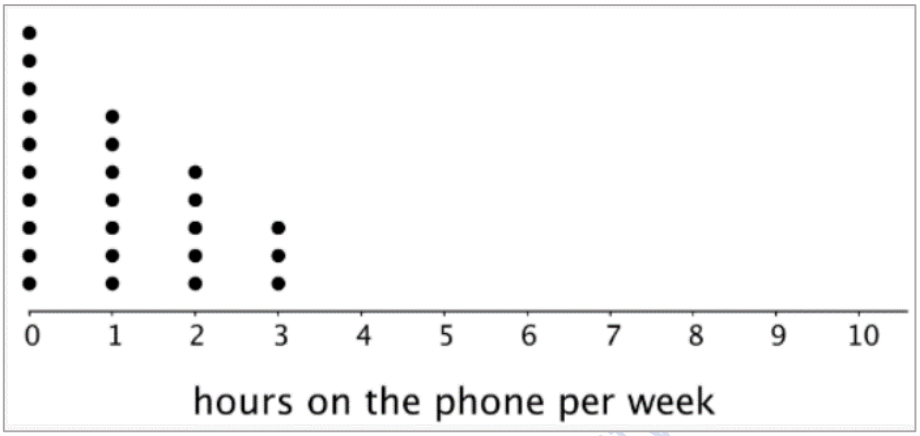


Figure 2: Dot plot of male literacy level of 10 Indian states

The dots plots are useful to highlight clusters of observation when we have continuous data. However, dot plots too, are difficult and inconvenient to create as the size of data increases.

IN-TEXT QUESTIONS

3. The following dot plot illustrates the number of hours a student spends talking on phone per week:



A. How many students report that they did not talk on the phone at all during the week?
B. How many students spend at least 2 hours on the phone?

4. True or False: We cannot create dot plots for continuous data.

2.5 BAR CHARTS

One of the most popularly used graph types is a bar chart that uses horizontal or vertical bars to depict the observations in the dataset. Bar charts can further be of two types: Stacked bar chart or grouped bar chart. Let's understand all the kinds of bar graphs using an example. Suppose we have the following data on two-wheeler sales in a city over the years:

Year	Two-wheeler Sales
2010	12,500
2015	17,000
2020	24,000

To create a vertical bar graph, we plot the years on the X-axis and the sale numbers on Y-axis. The height of the vertical rectangular bar represents the value of the data, as presented below:

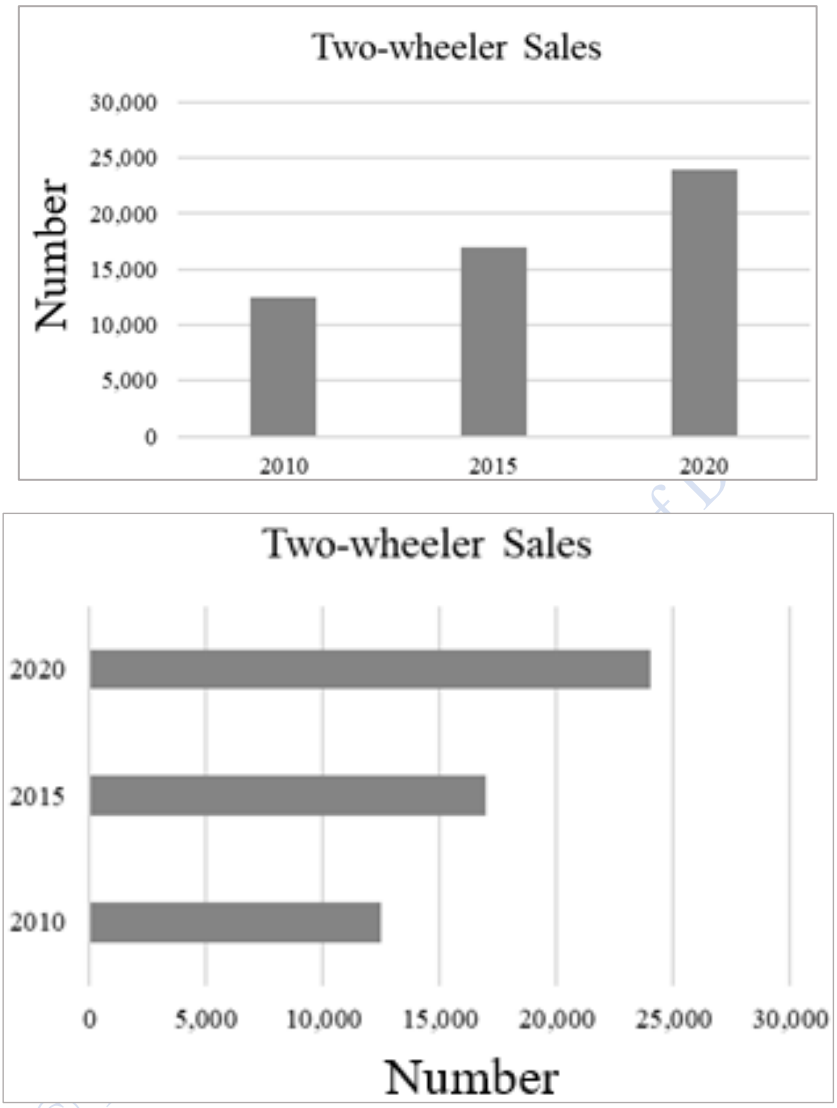


Figure 3: Bar chart of sale of two-wheelers in 2010, 2015 and 2020

Here each bar represents the number of two-wheelers sold in a given year.

Similarly, we can create a horizontal bar graph as well. Before we move on the stacked and grouped bar charts, it is important to note two points about bar charts:

a. The numerical axis must start from zero

b. The width of the bars and the distance between each bar remains constant.

Stacked bar charts are designed in a way that two or more categories of the same data are presented on the same bar. Stacked bar charts allow the reader to easily compare the value of various categories simultaneously. Let us take the above example one step further. Suppose that we have the following data for two-wheeler, three-wheeler and four-wheeler sales in a city:

Year	Two-wheeler Sales	Three-wheeler Sales	Four-wheeler Sales
2010	12,500	8,900	26,000
2015	17,000	11,000	33,500
2020	24,000	6,500	35,000

To create stacked bar charts, we draw the bars for each category on top of each other for a particular year. The final chart will look something like this:

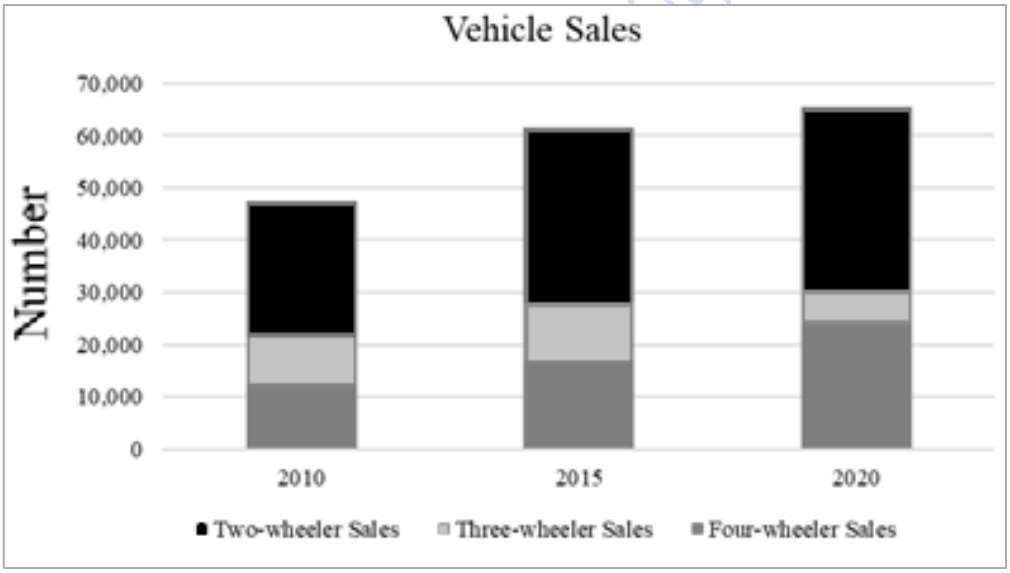


Figure 4: Stacked bar chart of sale of two-wheelers in 2010, 2015 and 2020

Here we can compare the sales of each category of vehicle for each year. We can also create **100% stacked bar charts** to present the same data in a more visually appealing way. A 100% stacked bar chart represents the share of each category in the data out of 100. The height of all the bars is equal to hundred percent and we can observe the relative changes in the values from the size of the sub-bars:

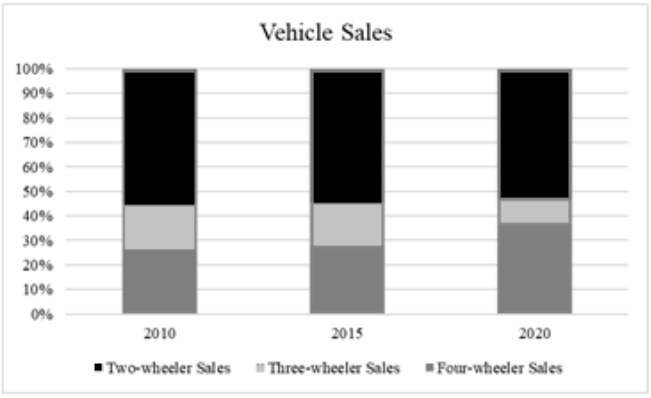


Figure 5: 100% Stacked bar chart of sale of two-wheelers in 2010, 2015 and 2020

Grouped bar charts are a convenient way to compare different categories of data. In a grouped bar chart, the bars for each category are placed adjacent to each other instead of on top of each other. In this case, we can easily observe the absolute changes in the data of each category. When interpreting stacked and grouped bar charts, it is crucial to pay extra attention to the legend that is usually displayed at the bottom of a chart.

It is important to remember that it can become complex to present too many categories of data in a stacked or grouped bar chart.

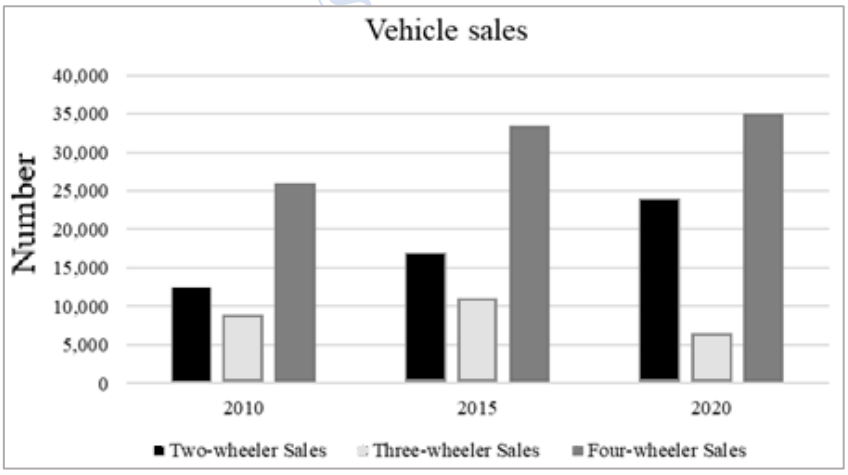
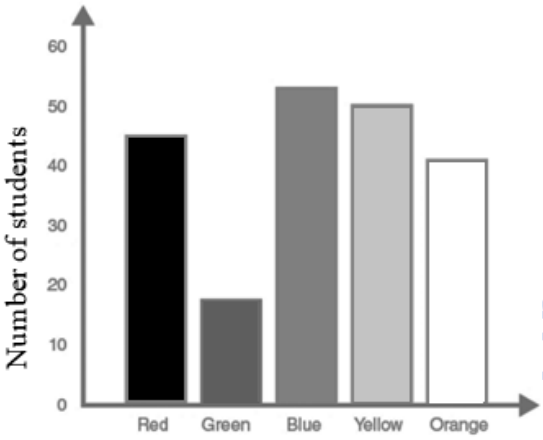


Figure 6: Grouped bar chart of sale of two-wheelers in 2010, 2015 and 2020

IN-TEXT QUESTIONS

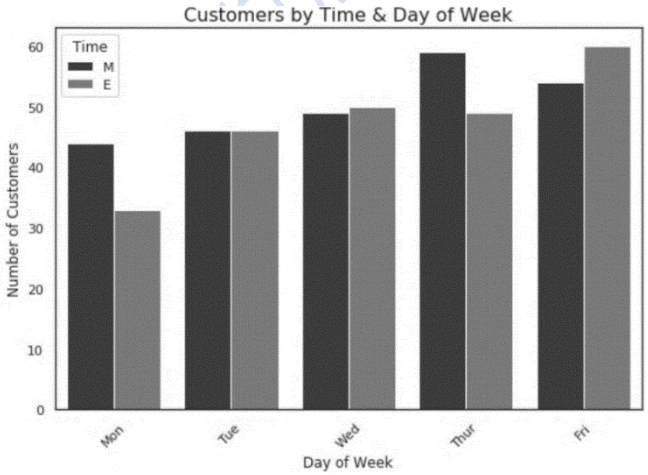
5. The following bar graph displays the favorite color of 200 kindergarten students in a school:



A. Which is the most preferred and least preferred color among the students?

B. How many students like oranges?

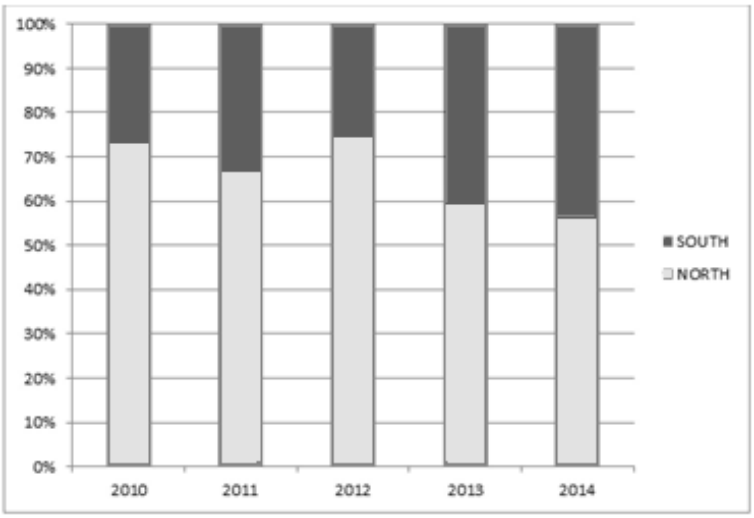
6. The following grouped bar chart shows the daily number of customers visiting a shopping center in morning (M) and evening(E).



A. On which day/days does the shopping center receive an equal number of customers in the morning as well as evening?

B. On which day/days do we see less customers shopping in the evening than in the morning?

7. A travel agent organizes trips to destinations either in South India or North India. The following 100% stacked bar chart presents the share of tourists visiting both the regions between 2010 and 2014.



A. Looking at the bar graph can you say that over the years the popularity of North Indian destinations is increasing? Why or why not?
B. In which year did maximum tourists visit South Indian destinations as compared to North Indian destinations?

2.6 HISTOGRAMS

At first, a histogram may look similar to a bar chart, but both are significantly different to each other. A histogram is also represented in the form of bars placed adjacent to each other, but each bar here represents the frequency with which an observation occurs in the dataset. The term **frequency** of any particular value is simply the number of times that value occurs in the data set. Hence, a histogram is also said to represent the 'frequency distribution of variables.' On the horizontal axis, we usually take the range/class intervals and on the vertical axis we place the frequency. We construct class intervals in such a way that each observation is contained in exactly one interval.

For instance, following are the economics marks of 20 college students:

86, 57, 69, 64, 67, 59, 81, 34, 47, 46, 38, 51, 66, 91, 42, 73, 62, 70, 77, 55

To construct a histogram, we divide the dataset into class intervals with each interval representing 10 marks. Next, we'll insert the frequency with which an observation occurs within each interval:

Marks	Frequency
30-40	2
40-50	3
50-60	4
60-70	5
70-80	3
80-90	2
90-100	1

The histogram for the above data looks like:

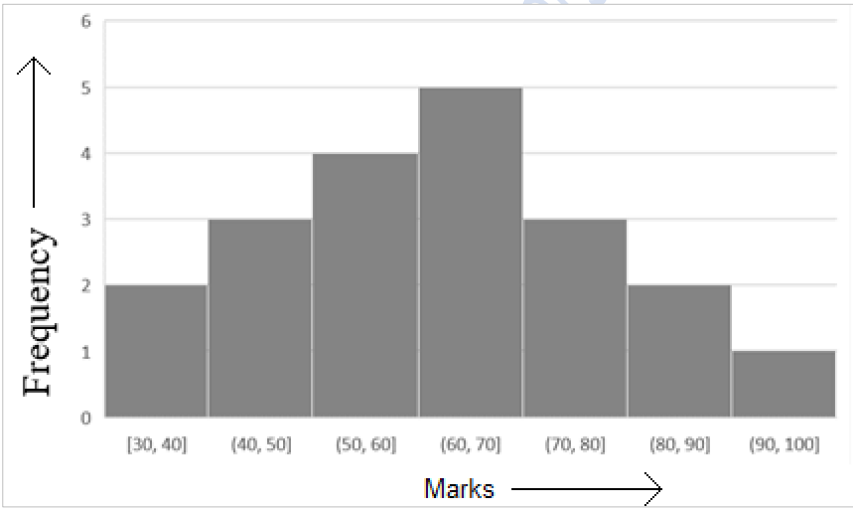


Figure 7: Histogram of economics marks of 20 college students

One clear difference between a bar chart and a histogram is that in a histogram, the bars do not have space in between them.

If there are areas of the measuring scale with a large concentration of data values and other areas with relatively sparse data, equal-width class intervals may not be the best option. The

following Dot plot displays data with a strong concentration of values in the center and a limited values on either side.

When there are only a few class intervals with equal widths, there are chances that all observations fall into just one or two of the classes. There may be some classes that have zero frequency if equal widths of class intervals are used. Using a few bigger intervals close to extreme observations and narrower intervals in the area of high concentration is a wise decision. To construct a histogram with **unequal class widths**, first determine the frequencies. Then **relative frequencies** may be calculated by the following formula:

$$Relative\ frequency\ = \frac{\text{Number of times the value occurs}}{\text{Number of observations in the data set}}$$

This signifies the proportion of times the value occurs in the data.

The height of each bar can then be computed using the formula given by:

$$Bar\ height = \frac{Relative\ frequency\ of\ class}{Class\ width}$$

The resulting rectangle or bar heights are typically called densities.

Such a histogram has an interesting property. If we multiply the bar height by the class width, we will get,

$$Relative\ frequency\ of\ class = \ Bar\ height\ \times Class\ width$$

Since the class width is nothing but the bar width, we can rewrite the above equation as:

$$Relative\ frequency\ of\ class = \ Bar\ height\ \times bar\ width$$

$$= \ Area\ of\ rectangle$$

This means that the area of each rectangle or bar represents the relative frequency of the corresponding class interval. Moreover, the sum of relative frequencies should be one and hence the total area of all rectangles in such a density histogram is one.

SHAPES OF HISTOGRAMS

The histogram in the first example follows a quite symmetric or a bell-shaped distribution. This means that if we place a mirror in the exact center of the distribution, the left and right side of the distribution will have the same shape. A symmetric or a bell-shaped distribution is also called a 'normal distribution.'

However, this is not the case always. Asymmetric distributions are called skewed. There are two types of skewed distributions – right skewed distribution and left skewed distribution. When the dataset has a greater number of observations on the left side of mean, then it is called a right or positively skewed distribution. Conversely, when the dataset has a greater number of observations on the right side of mean, then it is called a left or negatively skewed distribution.

Now consider the mathematics marks of the same 20 students:

45, 58, 51, 54, 49, 59, 88, 44, 49, 41, 48, 61, 66, 93, 40, 64, 69, 77, 72, 51

Follow the same steps to create intervals:

Marks	Frequency
40-50	7
50-60	5
60-70	4
70-80	2
80-90	2
90-100	1

The histogram looks like this:

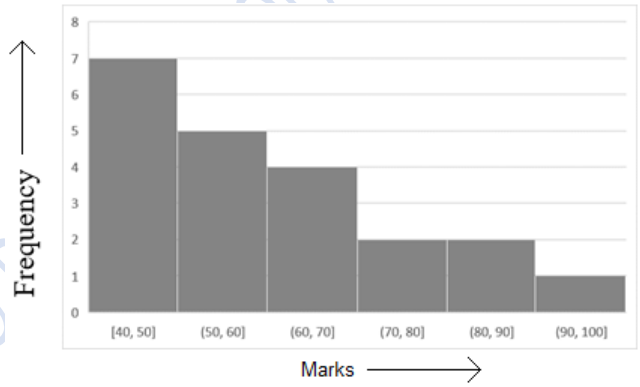


Figure 8: Histogram of mathematics marks of 20 college students

It is evident from the above histogram that it does not follow a symmetrical distribution. The average mathematics score of a student in this class is 58. Since our data is concentrated on the left side of the mean, we call such a distribution a right skewed distribution.

Lastly, consider English marks of the 20 students we have been examining:

75, 62, 93, 84, 96, 53, 81, 92, 87, 86, 91, 95, 82, 79, 64, 97, 62, 54, 71, 40

Creating equal intervals of 10 marks each:

Marks	Frequency
40-50	1
50-60	2
60-70	3
70-80	3
80-90	5
90-100	6

Using the table, we can create the following histogram:

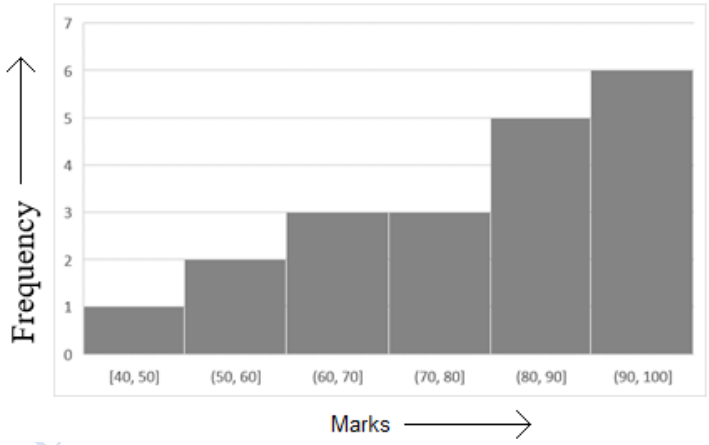


Figure 9: Histogram of English marks of 20 college students

If the average English marks in this class is 77, you may observe that we have more observations on the right side of the mean. This depicts a left-skewed distribution. We will discuss skewness in detail in the next lesson.

Histograms can take some more unusual shapes. Up till now we've seen histograms with only one peak, in the center, on the right hand or on the left-hand side of the data. Such a histogram with a single peak is known as a **unimodal histogram**. However, a histogram can have more than one peak. A histogram with two peaks is referred to as a **bimodal histogram**. Such a

histogram arises in case the data set consists of observations of two very dissimilar kinds of individuals or objects. Let us understand this with an example. A restaurant experiences high footfall during lunchtime and dinner time. Hence if a researcher collects data on the number of customers entering a restaurant in day, the histogram will display two peaks. A bimodal histogram with hypothetical numbers would look something like this:

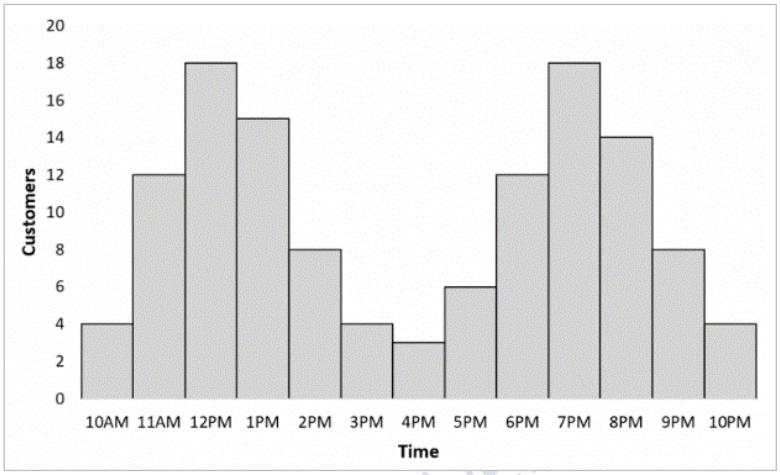


Figure 10: Bimodal histogram of number of the customers entering a restaurant in day

Similarly, **multimodal histograms** have more than two peaks. Such a graph suggests that the data may have different patterns of response. Suppose a researcher has some data on the heights of plants belonging to 3 different kinds of species. One of the species has tall plants, the other has medium plants and the third species has shorter plants. She creates a histogram to find out the pattern in her data. The following figure illustrates a histogram she created:

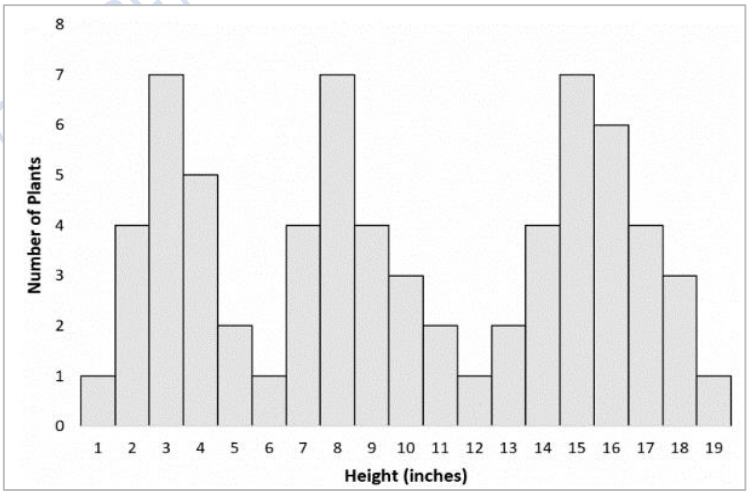


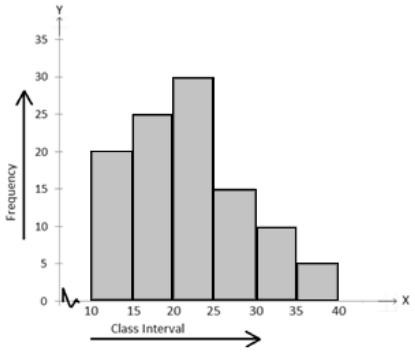
Figure 11: Multimodal histogram of height of plant species

This is clearly a multimodal histogram where each peak represents the most common height of each species of plants.

Although histograms are a useful way to present the data, one major drawback of histograms is that the reader cannot identify the individual values of the data by simply looking at the graph.

IN-TEXT QUESTIONS

8. Based on the following histogram, answer the questions:

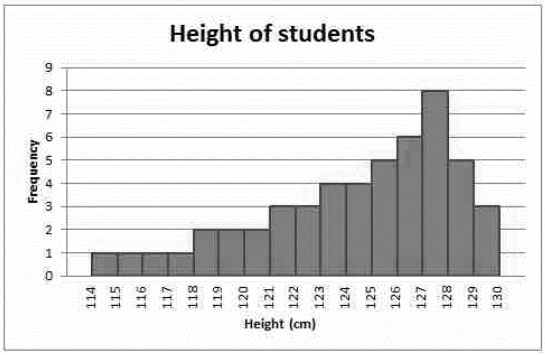


A. What is the frequency of the class interval 20-25?

B. Which class interval has the least frequency?

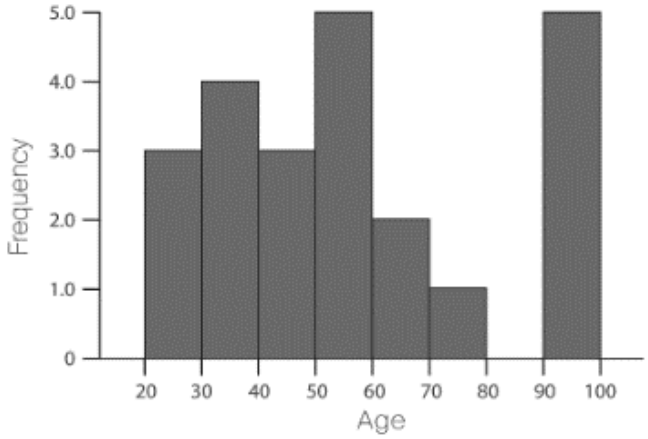
C. What is the cumulative frequency of the class interval 25-30?

9. The following histogram is an example of:



A. Symmetrical histogram
B. Left skewed histogram
C. Right skewed histogram
D. None of the above

10. The following histogram depicts a _____ histogram (unimodal / bimodal / multimodal)



11. Choose the correct option from the bracket and fill in the blank:

When a dataset has greater number of observations on the left side of mean, then it is called a _____ (right/ left) skewed distribution.

2.7 SUMMARY

In this lesson some of the most popular ways to display data, as a part of descriptive statistics, were discussed. In a stem and leaf plot, data is divided into two columns – stem and leaf. The plot gives a visual understanding of the distribution of the observations. The dot plots are illustrated in the form of dots representing each unit of observation. The dots are stacked over one another that represent the frequency of the value in our dataset. Bar charts use horizontal or vertical bars to depict the observations in the dataset. Bar charts can further be of two types: stacked bar chart and grouped bar chart. Histograms are the most widely used graphs in statistics. Vertical bars are used to depict the observations in the dataset. When the dataset has a greater number of observations on the left side of mean, it is called a right or positively skewed distribution whereas when the dataset has a greater number of observations on the right side of mean, then it is called a left or negatively skewed distribution. A histogram with a single peak is known as a unimodal histogram. A histogram with two peaks is referred to as a bimodal histogram. Histograms with more than two peaks are referred to as multimodal histograms.

2.8 GLOSSARY

- **Bar Charts :** A graph that displays data through vertical or horizontal rectangular bars with heights of each bar representing the values of the observation

- **Dot Plots :** A plot illustrating the frequency of observations through dots on a simple scale
- **Histograms :** A graph that shows the frequency of data using rectangular bars with heights of each bar representing the frequency of observations laying in that range
- **Negative Skewness :** When dataset has greater number of observations on right side of mean
- **Positive Skewness :** When dataset has greater number of observations on left side of mean
- **Stem and Leaf Plot :** A plot where each observation in data is presented in two columns- Stem (the first digit) and leaf (the other digits).

2.9 ANSWERS TO IN-TEXT QUESTIONS

1. C. 03, 32, 37, 51, 51, 59,70
2. A. 20 B. 7
3. A. 10 B. 8
4. False
5. A. Most preferred color- Blue; Least preferred color- Green
   B. 40
6. A. Tuesday B. Monday and Thursday
7. A. No, since the share of trips to North Indian destinations is declining over the years.
   B. 2014
8. A. 30 B. 35-40 C. 90
9. B. Left skewed histogram
10. Bimodal
11. Right

2.10 SELF-ASSESSMENT QUESTIONS

Q.1 Following are the weights of individuals who have taken membership of two local gyms in a city:

Gym 1: 94 90 95 93 128 95 125 91 104 116 162 102 90 110 92 113 116 90 97 103 95 120 109 91 138

Gym 2: 123 116 90 158 122 119 125 90 96 94 137 102 105 106 95 125 122 103 96 111 81 113 128 93 92

Create stem and leaf plot for both the gyms and interpret the plots.

Q.2 The following table summarizes the data collected by a researcher on the time taken to eat breakfast by 40 respondents:

Minutes:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
People:	6	2	3	5	2	5	0	0	2	3	7	4	1

A. Create a dot plot taking minutes on the X-axis.
B. Is the shape of the plot symmetrical?
C. How many people skip their breakfast in the morning?

Q.3 Honey has recently passed class 12th and the following table presents her marksheet:

Subject	**English**	**Hindi**	**Accountancy**	**Mathematics**	**Economics**	**Business studies**
Marks	87	93	96	99	97	92

A. Create a bar graph to depict her marks for each subject.

B. How can you show a comparison of her marks with her classmate Hazel who secured the following marks? Use a suitable bar graph to depict the marks of both the students on the same graph.

Subject	**English**	**Hindi**	**Accountancy**	**Mathematics**	**Economics**	**Business studies**
Marks	82	97	88	71	79	95

In which subject/subjects was the difference in the marks between Honey and Hazel highest?

Q.4 Mr. Kapoor owns a garden with 30 cherry trees. The height of the trees in inches are:

61, 63, 64, 66, 68, 69, 71, 71.5, 72, 72.5, 67.5, 73.5, 74, 74.5, 76, 76.2, 76.5, 77, 77.5, 78, 78.5, 79, 79.2, 80, 81, 82, 83, 84, 85, 87

A. Create a histogram of the above data by creating intervals of 5 inches. Comment on the shape of the graph.
B. Recently Mrs. Kapoor bought 10 more cherry plants to propagate them in their garden. Their heights in inches are: 57.5, 66, 40.5, 59, 46, 69.5, 67, 51.5, 52, 62. Incorporate this additional information and create a new histogram for all 40 cherry trees.

Comment and compare the shapes of both the histograms.

Q.5. Draw a histogram for the following dataset:

Class Intervals	50-60	60-70	70-80	80-90	90-100	100-110
Frequency	35	25	45	15	20	40

Comment on the shape of the histogram.

Q.6 The number of contaminating particles on a silicon wafer prior to a certain rinsing process was determined for each wafer in a sample of size 100, resulting in the following frequencies:

Number of particles	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	1	2	3	12	11	15	18	10	12	4	5	3	1	2	1

a. What proportion of the sampled wafers had at least one particle? At least five particles?

b. What proportion of the sampled wafers had between five and ten particles, inclusive? Strictly between five and ten particles?

c. Draw a histogram using relative frequency on the vertical axis. How would you describe the shape of the histogram?

2.11 REFERENCES

- Devore, J. L. (2016). *Probability and Statistics for Engineering and the Sciences*. Cengage learning.
- Larsen, R. J., & Marx, M. L. (2012). *An introduction to mathematical statistics and its applications*. Prentice Hall.
- McClave, J. T., Benson, P. G., & Sincich, T. (2018). *Statistics for business and economics*. Pearson Education.

2.12 SUGGESTED READING

- Gupta, S. C. (2019). *Fundamentals of statistics*. New Delhi, India: Himalaya publishing house.

LESSON 3

MEASURES OF LOCATION AND VARIABILITY

STRUCTURE

3.1 LEARNING OBJECTIVES

After reading this lesson, students will be able:

1. To recognize the different types of measures of location and variability

2. To establish the relationship between various measures of location and variability
3. To compute measures of location and variability
4. To understand the major advantages and disadvantages of measures of location and variability
5. To demonstrate the effects of change in origin and scale on the measures of location and variability
6. To identify the various shapes of distributions and
7. To develop the ability to present mean and variance through boxplots

3.2 INTRODUCTION

Continuing with our discussion about descriptive statistics, we now move on to the core elements of the descriptive statistics, also known as summary statistics. Recall that under descriptive statistics we quantitatively present an overview of the data. Descriptive statistics can be divided into two sub-groups:

a. Measures of central tendency – Measures of central tendency comprise of certain measurements that give us a typical or central value from the data around which the data is generally clustered. These central tendencies are also referred to as averages. In this course, we will concentrate on the three major measures of central tendencies- Mean, median, and mode. We will also briefly study quartiles, percentiles, deciles and trimmed mean.

b. Measures of variability – Measures of variability represent the spread of the data around the averages. It denotes the variation in the data. The most widely used measures of variability are range, standard deviation, and variance.

Let us now look at the two in detail.

3.3 MEASURES OF CENTRAL TENDENCY

3.3.1 Mean

The most common measure of central tendency is **arithmetic mean** or simply, mean of the data set. You must have studied about mean at some point in your mathematics course. Mean is simply the average value of the dataset represented which can be calculated by adding the value of all the observations and dividing the sum by the total number of observations, i.e.,

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

Where, μ denotes the mean of N observations, $\sum_{i=1}^{N} x_i = x_1 + x_2 + x_3 + \cdots + x_n$

Recall from previous lesson, population mean (parameter) is denoted by μ and sample mean (statistic) is denoted by $\bar{x}$.

Let us calculate the mean of following 5 numbers to understand the formula better:

$$12, 31, 78, 45, 29$$

Here,

$$\mu = \frac{12 + 31 + 78 + 45 + 29}{5}$$

$$= \frac{195}{5}$$

$$= 39$$

Hence, we can say that the mean of the dataset is 39.

Let's take another example. A researcher has a sample of age of 10 people.

$$45, 39, 53, 45, 43, 48, 50, 40, 40, 45$$

The researcher would like to know the average age of the group.

Again,

$$\bar{x} = \frac{45 + 39 + 53 + 45 + 43 + 48 + 50 + 40 + 40 + 45}{10}$$

$$\frac{448}{10}$$

$$= 44.8$$

So, we can conclude that the average age in the data set is 44.8 years.

Note here that we denote mean as $\bar{x}$. This is called the sample mean, that is calculated from the sample data. The mean taken from the population data is denoted by μ, as mentioned in the previous chapters.

Mean is a useful measure of central tendency with allows the reader to get an approximate idea about the typical value in a dataset. When we have very large datasets then the relevance of mean is clearer. Say a researcher gathers prices of houses in a locality. She gathers the price of about 1000 houses. The prices range between Rs. 45 lakhs to 1.9 crore. If the researcher calculates the average price of the houses as Rs. 1.2 crore, then we can easily interpret that a

typical house in that locality is worth Rs. 1.2 crore. So, mean is a convenient way to locate the average value of the dataset.

Before we move on to median, we should note some pros and cons of using mean. Arithmetic mean is the simplest measure of central tendency and is rigidly defined. The mean is not affected by the order of the data, i.e., the data may be in ascending or in descending order. Each and every observation in the dataset is used to calculate mean. This ensures that there is no loss of information. Finally, Arithmetic mean is capable of further mathematical treatment. If we have separate means of two groups of data, we can easily get the combined mean. However, mean also suffers from some limitations. First, mean is affected by extreme observations. Few extreme observations can impact the mean of the dataset which may not be the accurate representation anymore. Second, mean cannot be calculated if even one observation is missing. We cannot determine the mean by merely glancing at the dataset. It needs to be calculated each time using the formula. Finally, mean cannot be calculated for open ended class intervals.

Sometimes when we're collecting data, we may come across extremely large or extremely small values in our data that may either be incorrectly stated by the respondent or incorrectly recorded by the researcher. In such cases, we see that mean gets impacted by extreme values in our dataset. This is one of the major limitations of using mean as a central value.

The following example will make the argument clear: suppose that the salary of 10 employees in a firm, in lakhs per annum, is given below:

$$3.2\ , 4.5, 3.5\ , 3\ , 4.2, 16.5, 3.7, 3, 15, 4.4$$

Calculating mean using the formula:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

We get,

$$\mu = \frac{3.2 + 4.5 + 3.5 + 3 + 4.2 + 16.5 + 3.7 + 3 + 15 + 4.4}{10}$$

$$\frac{61}{10}$$

$$= 6.1$$

We conclude that the average salary in the firm is Rs 6.1 lakhs per annum. However, if we take a closer look at the observations, we realize that eight out of the 10 salaries lie between 3 and 4.5 lakhs and the two extreme values influenced the mean. So, we see here that mean is not representative of the above data.

When we have data that contains extreme values or outliers, in such cases, we prefer to use median over mean. We'll see why and how, in the subsequent section.

IN-TEXT QUESTIONS

1. The mean of the first 10 whole numbers is ___.

2. The mean of the following numbers is 24.

   16, 18, 19, 21, P, 23, 23, 27, 29, 29

   The value of 'P' is:

   A. 20.5 B. 30

   C. 24 D. 35

3. The mean of $6,\ 8,\ x+2,\ 10,\ 2x-1,\ and\ 2$ is 9. The value of x the value of the observations in the data are:

   A. x = 11; Observations = 13 and 21 B. x = 9; Observations = 11 and 17

   C. x = 10; Observations = 12 and 19 D. x = 12; Observations = 14 and 23

4. The average marks of 39 students of a class are 50. The marks obtained by the 40th student is 39 more than the average marks of all 40 students. Find the mean marks of all 40 students.

3.3.2 Median

The word **Median** is synonymous with 'middle'. Median is the middle observation in the data when the data is arranged in ascending or descending order of magnitude. Median is less affected by extreme values. The population median (parameter) is denoted by $\tilde{\mu}$ while sample median (statistic) is symbolized as $\tilde{x}$. There are two formulas to calculate median if we have N observations:

i. When there are odd number of observations, then median is simply the middle value, i.e., when n = odd,

$$\tilde{\mu} = \left(\frac{N+1}{2}\right)^{th} ordered\ value$$

ii. When there are even number of observations, then median is simply the average of the two middle values or, when n = even,

$$\tilde{\mu} = average\ of\ \left(\frac{N}{2}\right)^{th} and\ \left(\frac{N}{2}+1\right)^{th} ordered\ value$$

Let us consider the same example of salaries of 10 employees in a firm, in lakhs per annum:

$$3.2\ ,4.5, 3.5\ ,3\ ,4.2, 16.5, 3.7, 3, 15, 4.4$$

To calculate the median, the first step is to arrange the data in ascending order:

$$3, 3, 3.2, 3.5, 3.7, 4.2, 4.4, 4.5, 15, 16.5$$

Since we have even (10) number of observations, we will use the second formula, i.e., average of $\left(\frac{N}{2}\right)^{th}$ and $\left(\frac{N}{2}+1\right)^{th}$ observation. As we know N = 10. So,

$$\tilde{\mu} = average\ of\ \left(\frac{10}{2}\right)^{th}\ and\ \left(\frac{10}{2}+1\right)^{th}\ ordered\ value$$
$$= average\ of\ \ 5^{th} and\ 6^{th} ordered\ value$$
$$= average\ of\ \ 3.7\ and\ 4.2$$
$$= \frac{3.7+4.2}{2}$$
$$= 3.95$$

So, we can conclude that the median salary in the firm is Rs. 3.95 lakh per annum.

Note here that if we had salaries of only 9 employees, then since 9 is odd, the median would simply be the middle observation, i.e., the $\left(\frac{N+1}{2}\right)^{th}$ observation. This means that the median would've been the 5th observation, i.e., 3.7. We would then have concluded that the median salary in the firm is Rs. 3.7 lakh per annum.

It is worth noting here that even if we increase the values of the extreme observations from 15 and 16.5 to say 40 and 50, we will still get the same value of median. As median does not get affected by extreme values in a dataset, it is said to be representative of the sample. Hence, in cases when we have extreme observations, median is a better measure of central tendency than mean.

Median as a measure of central tendency is quite useful since it is easy to understand and calculate and, in some cases, median can be located by simply looking at the data. Median is better than mean since is not at all affected by extreme values in the dataset. Also, it can be calculated for open-ended distributions. However, the limitations of using mean are that the data must be in either ascending or descending order. If we have a very large dataset, arranging the data may be time-consuming. Median is not based on all the observations and so may not be representative of the dataset. Finally, it is not capable of further mathematical treatment.

IN-TEXT QUESTIONS

5. The import of electronic products in million dollars in a country for eight years was recorded as 27.4, 16.6, 1.7, 14.1, 32.9 18.7, 3.8, 22.5. The median import of the country is ____.

6. The following numbers are arranged in ascending order:

$$15, x, 22, x + 7, 32, 56, 88$$

If the median of the data is 25, then the value of x will be:

A. 17 B. 25

C. 18 C. 19

7. The runs scored in a cricket match by 11 players is as follows:

7, 16, 167, 41, 110, 57, 1, 16, 9, 0, 16

3.3.3 Mode

In simple terms, **mode** is that value in the dataset which appears most frequently. Like median, the value of mode too does not get affected by extreme observations. Mode is easy to understand and calculate and, in some cases, its value can be located by simply looking at the data. Mode can also be calculated for open-ended distributions.

For example, following are a person's daily expenditure, in Rs., on lunch in a week:

$$130, 115, 130, 130, 165, 150, 130$$

Clearly, we can observe that the person spends Rs. 130 four times a week on lunch. Hence, the mode is 130.

Data may have a single mode, two modes (known as bimodal) or more than two modes (multimodal).

Mode also suffers from some limitations. First, it is not rigidly defined. Second, mode is not based on all the observations and so may not be representative of the dataset. Finally, it is not capable of further mathematical treatment

3.3.4 Relationship between Mean, Median and Mode

Now that we have studied the three measures of central tendencies, you may believe that since all of the measures denote the central value in a dataset, then they must be the same. In this section we'll see that this is not always true.

First take the following 16 observations and try to calculate the mean, median and mode by yourself:

$$4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10$$

Mean =

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{4 + 5 + 6 + 6 + 6 + 7 + 7 + 7 + 7 + 7 + 7 + 8 + 8 + 8 + 9 + 10}{16}$$

$$= \frac{112}{16}$$

$$= 7$$

We have mean equal to 7.

Since we have even (16) observations, for median, we use the following formula:

$$\tilde{x} = average\ of\ \left(\frac{n}{2}\right)^{th}\ and\ \left(\frac{n}{2} + 1\right)^{th}\ ordered\ value$$

$$= average\ of\ \left(\frac{16}{2}\right)^{th}\ and\ \left(\frac{16}{2} + 1\right)^{th}\ ordered\ value$$

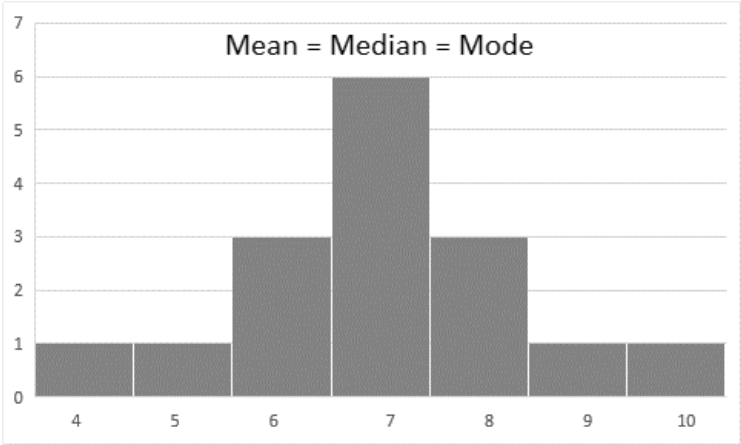
$$= average\ of\ 8^{th}\ and\ 9^{th} ordered\ value$$

$$= \frac{7 + 7}{2}$$

$$= 7$$

Hence, we get median equal to 7.

Finally, we can look at the data and can infer that the mode is 7 since it occurs the maximum number of times in the dataset.



In this example, it is evident that Mean = Median = Mode. We can look at the central tendencies using a histogram as well:

Figure 1: Histogram of symmetric data

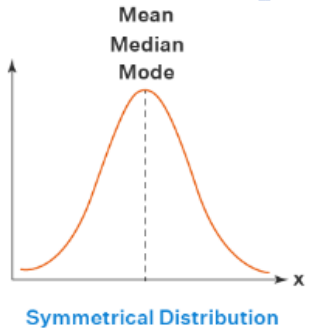


Figure 2: Histogram of symmetric data

We can see that the distribution is symmetric or is a bell-shaped curve. We can conclude from here that when the distribution is symmetric, the three measures of central tendencies converge at the same point.

However, when we have unsymmetrical or skewed data, the three measures are not equal. We will learn about skewness in detail later in the lesson.

The relationship between Mean, Median and Mode, when the data is not symmetrical, can mathematically be explained by the following formula:

$$3\,Median = 2\,Mean + Mode$$

IN-TEXT QUESTIONS

8. A researcher records the age of each participant in her study. The ages are:

   21, 59, 62, 21, 66, 28, 66, 48, 79, 59, 28, 62, 63, 63, 48, 66, 59, 66, 48, 79, 19, 79

   The mode of the above data is:

   A. 66 B. 59

   C. 79 D. 62

9. A researcher has computed the mean of her data as 22.5 and median as 20. Calculate the value of mode using these values. Is the distribution symmetrical?

10. A researcher calculates the following values of median, and mode of a distribution.

    Median = 17.5

    Mode = 20.5

    A. Calculate Mean.

    B. Do these values represent a symmetrical distribution?

3.3.5 Other measures of central tendency- Quartiles, percentiles, deciles and trimmed mean

Mean and median are not the only measures of central tendency. There are several others as well. We will briefly introduce the concepts of quartiles, percentiles and trimmed mean.

Just as a median divide the dataset into two equal halves, **quartiles** divide the data set into 4 equal parts. There are 3 quartiles, and each quartile consists of exactly 25% of observations. The first quartile, Q_1, containing the first 25% of observations is known as the lower quartile. The second quartile, Q_2 is called the median which divides the dataset into two. The third quartile, Q_3 is known as the upper quartile where 75% of observations lie below it and 25% of the observations are greater than this quartile.

To calculate quartiles, we first arrange the observations in ascending or descending order. We can then find the value of each quartile by using the following formulae:

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \ observation$$

$$Q_2 = \left(\frac{n+1}{2}\right)^{th} observation = Median$$

$$Q_3 = \left(3 \times \frac{n+1}{4}\right)^{th} observation$$

Consider the following dataset to understand the quartiles better:

0, 2, 5, 7, 8, 10, 16, 23, 35, 52, 77.

Since there are odd number of observations, you can easily identify the median in the above dataset. Median is 10. This is our second quartile, i.e., Q_2. Now as per the formula of the first quartile,

$$Q_1 = \left(\frac{11+1}{4}\right)^{th} observation$$

$$= \left(\frac{12}{4}\right)^{th} observation$$

$$= 3^{rd} observation$$

$$= 5$$

Similarly, for third quartile,

$$Q_3 = \left(3 \times \frac{n+1}{4}\right)^{th} observation$$

$$= \left(3 \times \frac{11+1}{4}\right)^{th} observation$$

$$= \left(3 \times \frac{12}{4}\right)^{th} observation$$

$$= 9^{th} observation$$

$$= 35$$

For grouped data, a quartile can be calculated using the following formula:

$$Q_j = L + \frac{\frac{j\,N}{4} - Pcf}{f} \times i \quad for\ j = 1, 2, 3.$$

Here, L = lower limit of quartile class, Pcf = Preceding cumulative frequency and i = size of quartile class.

We can easily use the above formula to obtain each quartile in the following manner:

$$Q_1 = L + \frac{\frac{N}{4} - Pcf}{f} \times i$$

$$Q_2 = L + \frac{\frac{N}{2} - Pcf}{f} \times i$$

and,

$$Q_3 = L + \frac{\frac{3N}{4} - Pcf}{f} \times i$$

Percentiles, on the other hand, simply denote that observation below which a particular percentage of observations fall. The value of percentiles varies on the scale from 1 to 100. For instance, 90th percentile would indicate that observation in the dataset below which 90% of observations fall. A percentile of an observation 'x' can be calculated by the following formula:

$$Percentile = \frac{Number\ of\ Values\ Below\ "x"}{Total\ Number\ of\ Values} \times 100$$

For grouped data, a percentile can be calculated by using the following formula:

$$P_j = L + \frac{\frac{jN}{100} - Pcf}{f} \times i \quad for\ i = 1,2,3, \dots, 99.$$

Similarly, **deciles** divide a dataset into 10 equal parts, which is in contrast to a percentile, which divides a dataset into 100 parts. As seen above, we can derive a similar formula for computing a decile:

$$D_j = L + \frac{\frac{jN}{10} - Pcf}{f} \times i \quad for\ i = 1,2,3, \dots, 9.$$

Where the symbols have the usual meaning and interpretation. You must note here that we will have ninety-nine percentiles ($P_1, P_2, \dots \dots \dots, P_{99}$) and ten deciles ($D_1, D_2, \dots \dots \dots, D_9$). For both, percentiles and deciles, the middle values, P_{50} and D_5 represent the median.

As we already know that mean is sensitive to extreme observations, we use **trimmed mean** to eliminate the extreme observations from our analysis. Such a measure is considered to be more accurate than the regular mean. For instance, to compute a 5% trimmed mean, we will eliminate the smallest 5% and the largest 5% of the sample observations and then calculate the mean of the remaining observations in the regular way.

IN-TEXT QUESTIONS

11. For the following data set, the value of upper quartile is _____.

    18, 30, 32, 39, 54, 57, 61, 62, 81, 88, 90

12. 2nd Quartile = 5th Decile = 50th Percentile =

    A. Mode B. Median

    C. Mean D. Trimmed mean

3.4 MEASURES OF VARIABILITY

Reporting the central values of a dataset provides only partial information about the entire data. It is possible that the two datasets have similar measures of central tendency, but both may differ based on the spread of the values. The logic will become clear through the following example.

Consider two restaurants selling pizzas that are located in the same town. A researcher collected data on the delivery time of both the restaurants, in minutes, for a week, as given below:

Restaurant A: $42, 50, 47, 43, 52, 55, 40$

Restaurant B: 47, 32, 70, 55, 65, 35, 25

Before you continue reading, you should try to solve the value of the central tendency. Since we do not have repeated values in the data, you should try and calculate the value of mean or median.

To calculate mean, use the following formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For restaurant A,

$$\bar{x}_A = \frac{42 + 50 + 47 + 43 + 52 + 55 + 40}{7}$$

$$= \frac{329}{7}$$
$$= 47$$

For restaurant B,

$$\bar{x}_B = \frac{47 + 32 + 70 + 55 + 65 + 35 + 25}{7}$$
$$= \frac{329}{7}$$
$$= 47$$

It is your task to check if the value of median for both the restaurants is also equal to 47 or not.

Moving on, we can conclude that the average delivery time of both the restaurants is 47 minutes. You may also try to create a dot plot of the two datasets. It will look something like this:

Now, since both the datasets have the same central value, can we claim that both the datasets convey the same information? No.

If you observe the values in each dataset carefully, you will notice that the delivery time of restaurant A ranges between 40 and 55 minutes, whereas the delivery time of restaurant B ranges between 25 and 70 minutes. Even in the Dot plots, you may observe that the data points of restaurant A are clustered together, whereas the data points of restaurant B are spread out. What can you infer from this extra piece of information? This means that restaurant A is more consistent in delivering pizzas between 40 and 55 minutes, that is, the time frame of delivery is shorter. Whereas the time taken by restaurant B to deliver a pizza is subject to more variation, that is the time frame of delivery is longer. Knowledge about such variation is important when we have to make important decisions. In this example, say you are starving, but you have just entered an Economics class that will finish in exactly 45 minutes. If you have to take the decision now to order a pizza, which restaurant would you prefer? You should prefer restaurant A since it is possible that restaurant B will deliver the pizza 20 minutes earlier and you are sure that neither would the professor finish the class, nor would you be able to leave the class that early.

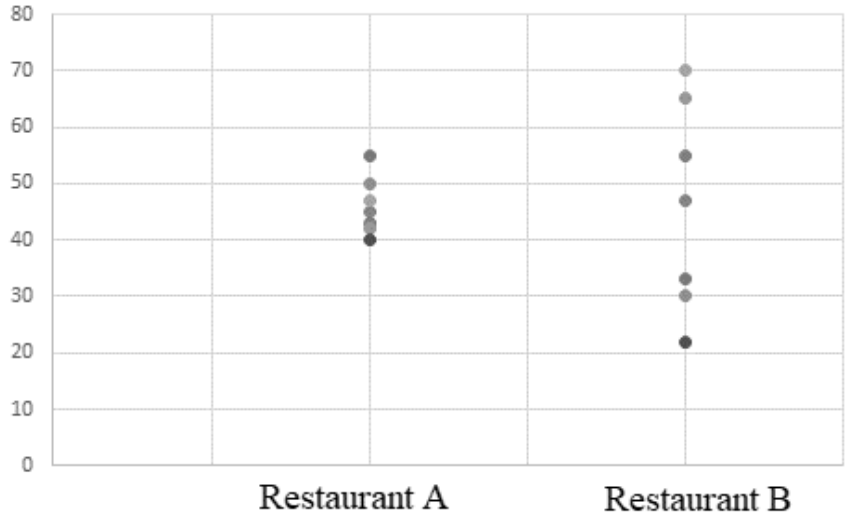


Figure 3: Dot plot of delivery time of two restaurants

This is a very minor decision to make, you may argue. However, measures of variability are a basis for making many more important distinctions and decisions. You may now appreciate the importance of studying variability.

So, measures of variability denote dispersion in a dataset. In other words, it shows how far away the data points are from the measure of central tendency. Low dispersion suggests that the data points are closer by and clustered around the average, whereas high dispersion indicates that the data points are further apart from each other. High dispersion also means that there is more variability in the data and hence, more chances of getting extreme values that may distort our estimations.

In this lesson we will study about three measures of variability – range, standard deviation, and variance.

3.4.1 Range and inter-quartile range

Range is the most straightforward measure of variability. It is simply the difference between the largest and the smallest value in the dataset.

$$Range = H - L$$

Where H is the highest value and L is the lowest value of a dataset. In the above example, the range for restaurant A $= 55 - 40 = 15$. Whereas the range for restaurant B $= 70 - 25 = 45$. This clearly indicates that there is more variability in the data points of restaurant B as compared to restaurant A.

The concept of range is extensively used in statistical quality control. Range is helpful in studying the variation in the prices of shares and debentures and other commodities that are very sensitive to price changes from one period to another. For the meteorological department too, range is a good indicator for weather forecast.

The relative measure corresponding to range, called **coefficient of range**, is obtained by the formula:

$$Coefficient\ of\ Range = \frac{H - L}{H + L}$$

Although range is easy to understand and compute, the usage of this measure is limited since it takes into consideration only the extreme data points and other observations in the data are simply ignored. So, range can be affected by extreme values. Moreover, as the size of the dataset increases, range loses its relevance as a measure of variability.

A slightly better measure of variability than range is the **inter-quartile range**. Here, instead of taking the difference between the extreme observations in the dataset, we take the difference between the upper and lower quartile. It is calculated as $Q_3 - Q_1$. This is a better measure than range since it takes into account only the middle 50% of the observations and the extreme observations do not affect the measure.

3.4.2 Standard deviation and Variance

Standard deviation and Variance are considered significantly better measures of variability as they are based on all the observations in the dataset and hence are more sensitive than range and inter quartile range. Since standard deviation is just the square root of variance, we will begin our discussion with variance.

As the name suggests, **variance** illustrates the variation in a dataset, that is, how far each data point is from the average. Formally, variance is calculated by dividing the sum of the squared deviations from the mean by (n – 1), where n denotes the sample size. We symbolize sample variance by s^2 and the formula can be written as:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

Here, the term $\sum_{i=1}^{n}(x_i - \bar{x})^2$ represents the sum of square of deviations of each data point taken from the sample mean $\bar{x}$. To make this simpler, we break this figure into three steps:

Step 1: Calculate the difference between the mean and each observation in the dataset

Step 2: Square each value of difference received from Step 1

Step 3: Add all the values you get from step 2, i.e., the squares of differences

You should note that we add up the squares of deviations (differences) instead of simply adding only the differences since by merely adding deviations, we will get zero. To understand it better, consider again the example of the two restaurants selling pizzas. The delivery time of restaurant A, in minutes was:

$$42, 50, 47, 43, 52, 55, 40$$

We have already calculated the mean delivery time, i.e., 47 minutes.

Now deviation of each data point from the mean would be:

$$(x_1 - \bar{x}) = 42 - 47 = -5$$
$$(x_2 - \bar{x}) = 50 - 47 = 3$$
$$(x_3 - \bar{x}) = 47 - 47 = 0$$
$$(x_4 - \bar{x}) = 43 - 47 = -4$$
$$(x_5 - \bar{x}) = 52 - 47 = 5$$
$$(x_6 - \bar{x}) = 55 - 47 = 8$$
$$(x_7 - \bar{x}) = 40 - 47 = -7$$

Now adding up all the deviations, we get,

$$\sum_{i}^{7}(x_1 - \bar{x}) = -5 + 3 + 0 - 4 + 5 + 8 - 7$$

$$= 0$$

Since the positive and negative deviations from the mean cancel each other out, we consider the sum of squared deviations from the mean. In this way we can eliminate all the negative values. So, in our example,

$$(x_1 - \bar{x})^2 = (-5)^2 = 25$$
$$(x_2 - \bar{x})^2 = 3^2 = 9$$
$$(x_3 - \bar{x})^2 = 0^2 = 0$$
$$(x_4 - \bar{x})^2 = (-4)^2 = 16$$
$$(x_5 - \bar{x})^2 = 5^2 = 25$$

$$(x_6 - \bar{x})^2 = 8^2 = 64$$

$$(x_7 - \bar{x})^2 = (-7)^2 = 49$$

Now adding up all the squared deviations, we get,

$$\sum_{i}^{7}(x_1 - \bar{x})^2 = 25 + 9 + 0 + 16 + 25 + 64 + 49$$

$$= 188$$

Finally, to get the variance, divide the sum of squared deviations by $n - 1$.

$$s^2 = \frac{188}{7-1}$$

$$\therefore s^2 = 31.3$$

Sample standard deviation, s, is simply the square root of the variance, that is,

$$s = \sqrt{s^2}$$

In our example, standard deviation $= \sqrt{31.3} = 5.6\ approximately.$

We can interpret the value of standard deviation as a typical deviation from sample mean. You may also understand it in this way that a typical delivery by restaurant A would be 47 minutes and ±5.6 minutes. This means that a typical delivery from restaurant A will take something between 41.4 and 52.6 minutes. In the above example, it is now your task to calculate the variance and standard deviation of the delivery time of restaurant B. Check if you get the value of $s^2 = 337.8$ and $s = 18.3$. Try and interpret the value of standard deviation you got for restaurant B on your own. What is the typical time span of deliveries from restaurant B?

We can clearly see that the variance and standard deviation of restaurant B is higher than restaurant A. This implies that there is more variation in the delivery time by restaurant B as compared to restaurant A.

Note that both s and s^2 are non-negative.

As we had differentiated the symbols of population mean and sample mean, we do the same in case of sample variance/standard deviation and population variance/ standard deviation. While sample variance is denoted by s^2 and sample standard deviation is denoted by s, population variance is denoted by the Greek alphabet σ^2 and standard deviation is denoted by σ. The basic understanding of the population parameters remains the same, just that we now say that the

population variance denotes the variability in the population and population standard deviation denotes the typical deviation of a population value from its population mean μ.

The formal representation of the population variance and standard deviation also gets modified. In terms of the population parameters, the population variance can be written as:

$$\sigma^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$$

Population standard deviation can be written as:

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

Just as we may use $\bar{x}$ to make inferences about μ, similarly, we use s^2 to make inferences about σ^2.

Note here that we divide the sum of deviations from mean by N and not N-1 since s^2 is based on *n-1* degrees of freedom. **Degree of freedom** refers to the maximum number of independent values, that have the freedom to vary, in a sample. In other words, if we fix $\bar{x}$, then we need only determine (n−1) number of the elements in the sample in order to know the nth element of the sample.

Coefficient of Variation (C.V.)

A very popular and frequently used relative measure of variation is the **coefficient of variation** denoted by C.V. This is simply the ratio of the standard deviation to arithmetic mean expressed as a Percentage.

$$Coefficient\ of\ Variation = C.V. = \frac{\sigma}{\bar{x}} \times 100$$

When C.V. is less in the data, it is said to be less variable or more consistent.

Consider the following data on the mean daily sales and standard deviation of four regions:

Region	Mean daily sales (Rs.'000)	Standard deviation (Rs.'000)
1	82	10.41
2	44	5.85
3	70	9.52
4	60	11.22

To determine which region is most consistent in terms of daily sales, we can calculate the coefficient of variation:

$$CV_1 = \frac{10.41}{82} \times 100$$

$$= 12.69$$

$$CV_2 = \frac{5.85}{44} \times 100$$

$$= 13.29$$

$$CV_3 = \frac{9.52}{70} \times 100$$

$$= 13.60$$

$$CV_1 = \frac{11.22}{60} \times 100$$

$$= 18.70$$

Since the coefficient of variation is 12.69, the minimum for region 1. Hence the most consistent region for sales is Region 1.

IN-TEXT QUESTIONS

13. If the standard deviation of a data is 0.012. The variance will be:

    A) 0.144 B) 0.00144

    C) 0.000144 D) 0.0000144

14. The variance of the first 10 whole numbers is _____.

15. In a class of 100 students, the mean marks on a particular exam was 75, and the standard deviation was 0. This implies that:

    A) All students scored 75 marks B) Variance is 0.75

    C) Standard deviation cannot be zero D) None of the above

16. If the mean of certain observations is given as 60 and the standard deviation is 12, then the coefficient of variation is 20%. (True/False)

3.5 EFFECT OF CHANGE IN ORIGIN AND SCALE

Now that we are clear with the calculation of the measures of central tendencies and variations, we will now examine the effects of change in origin and scale on both mean and standard deviation/variance. It is important to learn about these effects since often a researcher may incorrectly report the values in the dataset and to recalculate the mean and standard deviation can become a lengthy task. To avoid such an inefficiency, we will learn how does mean and variance respond if every value in the data set is changed.

3.5.1 Change in origin

Change in origin can also be understood in simpler terms as shifting of data. This suggests a situation in which we add or subtract a constant value from all the observations in our dataset. We will now see how this impacts the value of mean and standard deviation. Let us understand this concept through a simple example.

Suppose we have the following 5 observations in our sample:

3, 9, 12, 18, 23

For convenience, we will call these observations as the 'original dataset.' At this point, you can easily compute the mean and standard deviation of the original dataset.

Mean = $\frac{\sum_{i=1}^{n} x_i}{n} = \frac{3+9+12+18+23}{5} = \frac{65}{5} = 13$

Variance = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{100+16+1+25+100}{4} = \frac{242}{4} = 60.5$

Standard deviation = $\sqrt{variance} = \sqrt{60.5} = 7.7\ (approx.)$

Now, let us add a constant value 5 to each observation in our original dataset. The new values will become:

8, 14, 17, 23, 28

Again, let us calculate the mean and standard deviation of the new values. We get,

Mean = $\frac{\sum_{i=1}^{n} x_i}{n} = \frac{8+14+17+23+28}{5} = \frac{90}{5} = 18$

Variance = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{100+16+1+25+100}{4} = \frac{242}{4} = 60.5$

Standard deviation = $\sqrt{variance} = \sqrt{60.5} = 7.7$

Did you notice the pattern in the new estimates? As compared to the original dataset, we observe that the value of mean increased by 5 whereas the value of standard deviation and variance remained unchanged. You need to remember that this pattern remains the same whatever constant we add or subtract from our observations. Try repeating the exercise by subtracting 5 from each observation in the original dataset and calculating the mean and standard deviation of the new values. You will see that the value of mean reduces by 5 and standard deviation again remains the same. So, in general we can conclude that:

$If\ y_i = x_i + a, where\ a\ is\ a\ positive\ or\ negative\ constant, then,$

$$Mean: \bar{y} = \bar{x} + a,$$

$$Standard\ deviation: s_y = s_x$$

3.5.2 Change in scale

The other way we can modify data is through scaling. Scaling the data means to either multiply or divide all the observations in the data. Let us see the impact of change in scale on mean and standard deviation. Consider again the original dataset in the above example:

3, 9, 12, 18, 23

We will change the scale of the dataset by multiplying each observation by 3.

The new dataset we will get is:

9, 27, 36, 54, 69

Calculate the mean and standard deviation:

Mean = $\frac{\sum_{i=1}^{n} x_i}{n} = \frac{9+27+36+54+69}{5} = \frac{195}{5} = 39$

Variance = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{900+144+9+225+900}{4} = \frac{2178}{4} = 544.5$

Standard deviation = $\sqrt{variance} = \sqrt{544.5} = 23\ (approx.)$

Now compare the above values with the original mean and standard deviation. We find that by multiplying each observation by 3, the mean and standard deviation also get multiplied by 3. Hence, scaling the observations has also scaled the measures of central tendency and variability. Try dividing all the observations in the initial dataset by 3 and check whether the mean and standard deviation too get divided by 3. Since this pattern will follow every time we scale a dataset, we can conclude generally that,

$If\ y_i = bx_i, where\ b\ is\ a\ positive\ constant, then,$

$$Mean: \bar{y} = b\bar{x},$$

$$Standard\ deviation: s_y = bs_x$$

In conclusion, we can say that mean is affected by both- change in origin as well as change in change in scale; whereas Standard deviation is affected only by change in scale.

IN-TEXT QUESTIONS

17. Suppose the standard deviation of a dataset is 6. If each observation is divided by 3 then the standard deviation of the new dataset will be:

    A) 3 B) 2

    C) 18 D) 9

18. A researcher measures the weight of 10 students. The mean weight she calculates is 57kg. Later she realized that the weighing scale was misreporting each weight and she had to add 3 kgs to the weight of each student. The new mean weight of students will be _____.

19. If the standard deviation of 11, 21, 31…,71, 81, 91 is 'K', then the standard deviation of 15, 25, 35…,75, 85, 95 will be:

    A) K - 4 B) K +4

    C) K D) 4K

3.6 SKEWNESS

The measures of central tendencies and variation discussed above do not reveal all the characteristics of a given set of data. For example, two distributions may have the same mean, variance and standard deviation but may differ widely in terms of their shape and peakedness. The given data is either symmetrical or it is not. It may be flat, normal or peaked.

If the distribution of data is not symmetrical, it is called asymmetrical or skewed. Thus, **skewness** refers to the lack of symmetry in distribution.

A simple method of detecting the direction of skewness is to look at the tails to distribution. The rules are:

1. Data are symmetrical when there are no extreme values in a particular direction so that low and high Values balance each other. In this case, mean = median = Mode, or $\bar{x}$ = Q_2 = Mode.

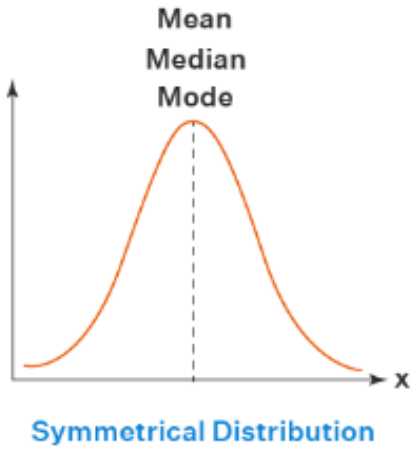


Figure 4: Histogram of symmetric distribution

2. If the longer tail is towards the lower value or left-hand side, the skewness is negative. Negative skewness arises when the mean is decreased by some very low values. Then we have, mean < median < mode.

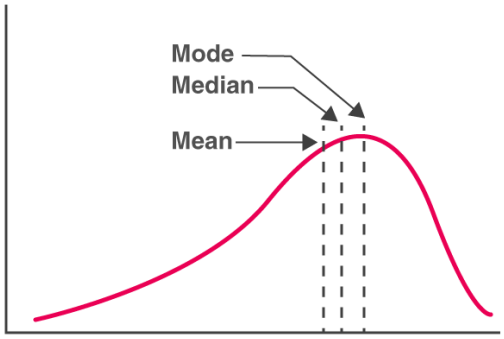


Figure 5: Histogram of negatively skewed distribution

3. If the longer tail of the distribution is towards the higher values or right-hand side, then skewness is positive. Positive skewness occurs when mean is increased by some very high valued observations. In this case, mean > median > mode.

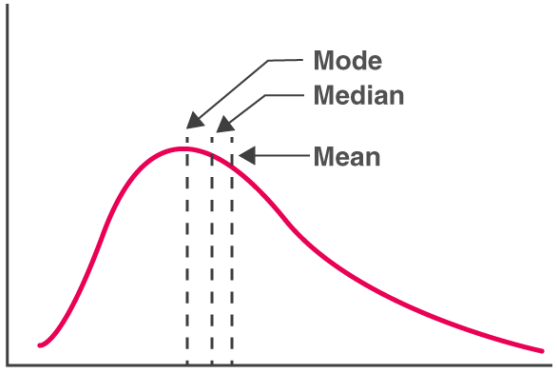


Figure 6: Histogram of positively skewed distribution

Relative Skewness

Karl Pearson's coefficient of skewness to compare between two distributions is given by:

$$SK = \frac{Mean - Mode}{Standard\ deviation}$$

$$SK = \frac{\bar{x} - Mode}{\sigma}$$

If Mode is not given, we can use the approximate relationship studied earlier in the lesson, i.e. Mode = 3 Median – 2 Mean. Hence, we can write the equation of coefficient of skewness as:

$$SK = \frac{3(Mean - Median)}{Standard\ deviation}$$

Now, if,

- SK = 0, it is a symmetrical distribution.
- SK > 0, the distribution is positively skewed.
- SK < 0, then it is a negatively skewed distribution.

But in practice, normally the value of SK lies between $+1 \leq SK \leq -1$.

For an open-ended distribution with extreme values in data with positional measures such as median and quartiles, Bowley's coefficient of skewness is used,

$$SK = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

Here, Q_2 is the median. Again, if

- SK = 0, it is a symmetrical distribution.
- SK > 0, the distribution is positively skewed.
- SK < 0, then it is a negatively skewed distribution.

IN-TEXT QUESTIONS

20. For the frequency distribution of a variable x, mean = 32, median = 30 and mode = 26. The distribution is:

    A. Positively Skewed B. Negatively skewed

    C. Symmetric D. None of the above

21. A researcher gathers data on the number of years of experience professors in a university have. The mean, median, mode and standard deviation are 25, 24, 26 and 5, respectively. Karl Pearson's coefficient of skewness is _____ (0.20 / - 0.20).

3.7 BOXPLOTS

We will conclude this lesson by discussing boxplots. Boxplots are yet another type of graphical representation that is extremely informative. On one hand, the stem and leaf plots, bar graphs and histograms depict a particular aspect of the data, on the other hand, measures of central tendency and variability also focus on separate features of the data. Is there no way we can visualize the data and also trace the mean and variability at the same time? There is, and the answer is boxplots. Boxplot is a comprehensive graphical representation of data in which we can illustrate not only the central value and the variability in the data, but it is also capable of presenting the extreme values (outliers) as well as the shape of the distribution. The following figure displays a box plot and its features:

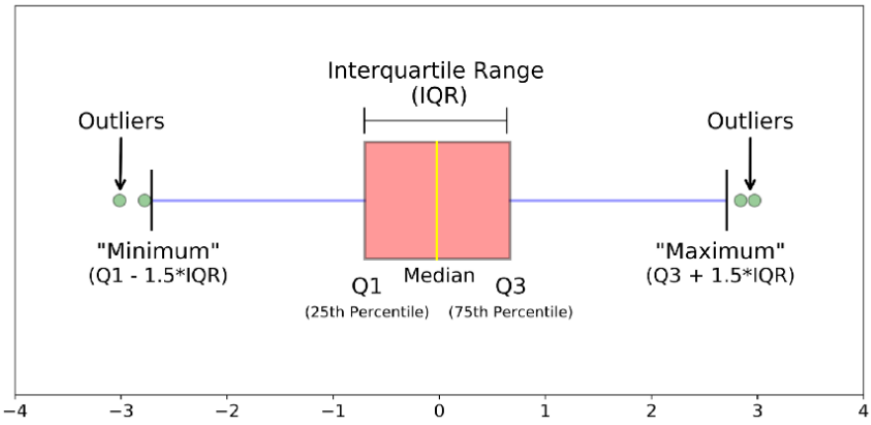


Figure 7: Boxplot

The left side of the box represents the first quartile of the dataset whereas the right side denotes the third quartile. The difference between the two, i.e., the interquartile range, or fourth spread (f_s) is the length of the box. You can see the median, or the second quartile in the middle, dividing the box into two equal halves. The whiskers (the two hands extending out of the box on the right and left side) extend towards the smallest and the largest value in the dataset, which are not outliers. The small dots beyond the minimum and maximum values are termed as outliers.

Let us create a boxplot together for better understanding. Suppose we have the following dataset containing 10 numbers:

34, 29, 25, 35, 28, 37, 30, 35, 29, 38

To create a boxplot, we first need the 5-number summary of the data, i.e., the smallest number, Q_1, Median (Q_2), Q_3 and the largest number. To get these, it is advisable to arrange the data in ascending or descending order. So, we get:

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

You should now try and identify the 5-number summary from the above data.

We will get, smallest number = 25

$Q_1 = 29$

Median (Q_2) = 32

$Q_3 = 35$

Largest number = 38

Interquartile range = 35 – 29 = 6

Now start by drawing an X-axis with appropriate labels and then draw a box around the first and third quartile. Mark the median value in the middle. The length of the box must equal the interquartile range. Next, draw two whiskers from both the ends of the box extending to the smallest value on the left and largest value on the right. You should get a graph that looks like:

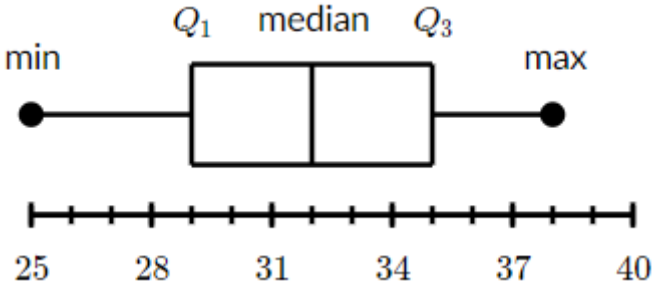


Figure 8: Boxplot

The boxplots can also be created in a vertical manner.

In a larger dataset, we can differentiate between the highest and lowest values of the dataset and the extreme values that are also known as outliers. We use the following formula to calculate the minimum and maximum values in a data set and any other value lower or greater than these, respectively are termed as outliers:

Minimum value: Q1 – 1.5 IQR

Maximum value: Q3 + 1.5IQR

Where IQR is simply the Interquartile range. So, any value below the minimum and any value above the maximum are outliers and we denote such values in the box plot as dots. Formally, any observation farther than $1.5f_s$ from the closest quartile is termed as an outlier. An outlier is extreme if it is more than $3f_s$ from the nearest quartile, and it is mild otherwise. Having an idea about outliers is important since as we have seen in earlier sections extreme values can affect our measures of central tendencies and variability. There is a possibility that the extreme value is a result of an error in any step of research. By identifying such extreme values, we can be cautious with our study ahead.

Boxplots are also useful in indicating the shape of the distribution of the data, i.e., whether we have symmetrical or skewed data. When the median sits exactly in the center of the box and has equal length of the whiskers on both the sides, we can say that the distribution is symmetrical. However, when the median lies somewhere on the right end of the box with the right whisker smaller than the left one, then the distribution is said to be left skewed and vice-

versa. The following figure represents the relationship between the shape of distribution and boxplots:

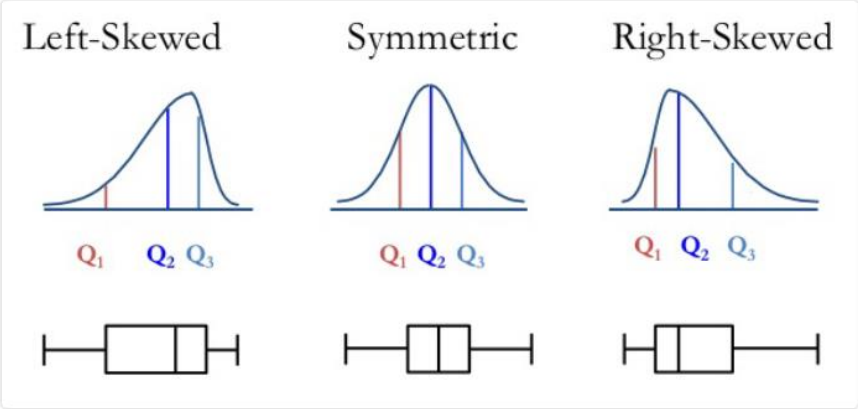


Figure 9: Boxplot and shape of distribution

We can easily compare multiple datasets using boxplots. For doing so, we draw two boxplots adjacent to each other on the same scale. For instance, suppose a researcher is studying about the duration of Indian classical songs and rap songs. She collects data from 100 classical and rap songs and draws two boxplots given below:

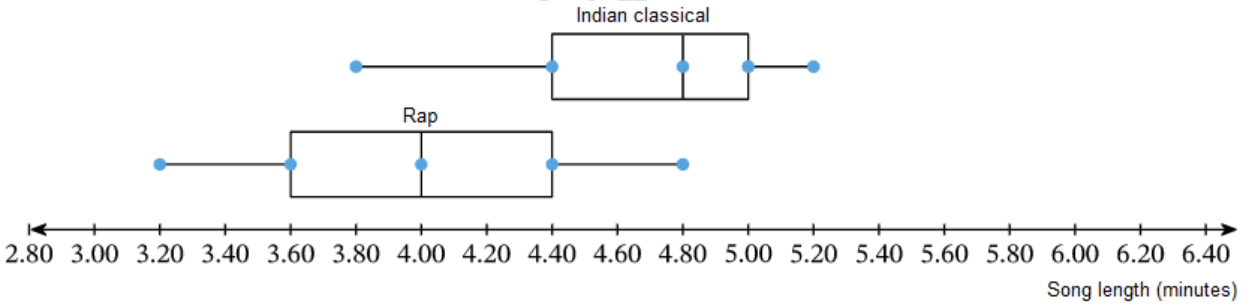


Figure 10: Comparative boxplot of duration of Indian classical songs and rap songs

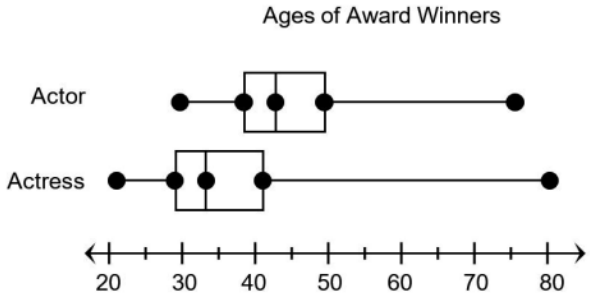
Try and attempt to interpret both the box plots yourself. What differences can you observe in both the plot? What do these differences signify?

Let us begin interpreting the median. We can clearly see that the median length of classical songs is significantly higher than that of rap songs. The average length of classical songs is 4.8 minutes, whereas the average length of rap songs is 4 minutes. Next, interpret the length of the box, i.e. the interquartile range. It shows 50% of the data lies within this range. So, we can say that half of the classical songs are 4.40 to 5 minutes long. Whereas 50% of the rap songs are 3.60 to 4.40 minutes long. We can also interpret the range of the data, i.e., the difference between the maximum and minimum value. For Indian classical songs, the range is 5.20 – 3.80

that is 1.4 minutes. Whereas for rap songs, the range is 4.80 – 3.20 = 1.6 minutes. So, we can say that the length of rap songs is more variable than classical songs. Finally, we can observe the shape of the distributions by studying the location of the median value inside the box. The median length of Indian classical songs is towards the right end of the box, indicating that the distribution is left skewed. This means that most of the observations lie towards the right of mean. In other words, we can say that most of the Indian classical songs in the data have a longer duration. In contrast, the median length of rap songs lies right in the middle of the box. This means that the distribution is quite symmetric.

IN-TEXT QUESTIONS

22. The following boxplots illustrates the data about the ages of actors and actresses who have won the National film award since 1967.



Mark all the statements that support the data shown by the boxplots:

A. The first quartile age of Best Actor winner is less that the last quartile age of Best Actress winner

B. The minimum age of Best Actor winner is equal to the minimum age of Best Actress winner

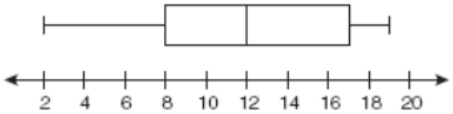
C. The range of age of Best Actor winner is higher than the range of age of Best Actress winner

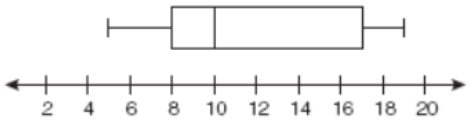
D. Both the distributions are left skewed

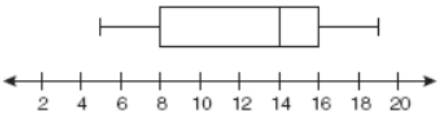
23. Inspired by his statistics class; a student started maintaining a record of the number of minutes he was late to enter the classroom every day. He recorded the time for 15 days and the following list displays the data (in minutes):

19, 12, 9, 7, 17, 10, 6, 18, 9, 14, 19, 8, 5, 17, 9

Which of the given boxplots accurately depict the data:

A. 

B. 

C. 

D. 

3.8 SUMMARY

This lesson focused on the quantitative features of data in terms of measures of central tendency and variability and their applications. Measures of central tendency refer to a typical or central value of the data around which the data is generally clustered. Arithmetic mean is the average value of the dataset which can be calculated by adding the value of all the observations and dividing the sum by the total number of observations. Since mean is affected by extreme values, median is preferred. Median is the middle observation in the data when the data is arranged in ascending or descending order of magnitude.

When the distribution is symmetric, the three measures of central tendencies converge at the same point. Quartiles divide the data set into 4 equal parts. Percentiles denote the observation below which a particular percentage of observations fall. Deciles divide a dataset into 10 equal parts. Since arithmetic mean is affected by extreme values, trimmed mean is used to eliminate the extreme observations from analysis. Under measures of variability, range is the simplest measure. It is the difference between the largest and the smallest value in the dataset. Inter-quartile range is calculated by taking the difference between the upper and lower quartile. Variance is calculated by dividing the sum of the squared deviations from the mean by (n – 1). Standard deviation is simply the square root of variance. Degree of freedom refers to the maximum number of independent values, that have the freedom to vary, in a sample. Coefficient of variation is a relative measure of variation. Mean is affected by both- change in origin as well as change in change in scale; whereas Standard deviation is affected only by change in scale. Skewness refers to the lack of symmetry in distribution. In case of a right or positively skewed distribution, the value of mean is the largest, followed by the median and mode. The opposite is true in the case of a left or negatively skewed distribution, i.e., the value of mode is the largest and mean has the smallest value. Boxplots are capable of illustrating the central values, variability in the data, the extreme values (outliers) as well as the shape of the distribution. The boxplots are extremely useful to compare two datasets and identify outliers.

3.9 GLOSSARY

- **Boxplot:** Graphical representation of measure of central tendency, variability and skewness of numerical data using quartiles

- **Change in Origin:** Addition or subtraction of a constant value from all observations in dataset
- **Change in Scale:** Multiplying or dividing all the observations in the dataset with a constant term
- **Coefficient of Variation:** Relative measure of variation
- **Deciles:** Divide dataset into 10 equal parts
- **Degree of Freedom:** Maximum number of independent values, that have the freedom to vary
- **Inter-Quartile Range:** Difference between the upper and lower quartile
- **Mean:** Average value of the dataset
- **Median:** Middle observation in the data when the data is arranged in ascending or descending order
- **Measures of Central Tendency:** Certain measurements that give a typical or central value from the data
- **Measures of Variability:** Represent spread of data around averages
- **Mode:** Most frequently occurring value in the dataset
- **Percentiles:** That observation below which a particular percentage of observations fall
- **Quartile:** Divide the data set into 4 equal parts
- **Range:** Difference between largest and smallest value in the dataset
- **Skewness:** Lack of symmetry in distribution
- **Standard Deviation:** square root of variance
- **Trimmed Mean:** Computing mean after eliminating the extreme observations
- **Variance:** Illustrates the variation in a dataset

3.10 ANSWERS TO IN-TEXT QUESTIONS

1. 4.5
2. D. 35
3. B. x = 9; Observations = 11 and 17
4. 51

5. 17.65 million dollars

6. C. 18

7. A. 40

   B. False, due to extreme values

   C. 16

   D. True

8. A. 66

9. 15. Unsymmetrical

10. A. 16. B. No.

11. 81

12. B. Median

13. C) 0.000144

14. 8.25

15. A. All students scored 75 marks

16. True

17. B) 2

18. 13

19. C) K

20. A. Positively Skewed

21. - 0.20

22. A, D

23. B

3.11 SELF-ASSESSMENT QUESTIONS

Q.1 The mean of the four numbers is 37. The mean of the smallest of three numbers is 34. If the range of the data is 15, what is the mean of the largest three numbers?

Q.2 The mean and variance of 10 observations is 4 and 2, respectively. If each observation is multiplied by 2, Calculate the mean and variance of new data.

Q.3 If V is the variance and M is the mean of first 10 natural numbers, then what is the value of $V + M^2$?

Q.4 Let 'x' be the median of the following observations:

33, 42, 28, 49, 32, 37, 52, 57, 35, 41

If 32 is replaced by 36 and 41 is replaced by 63, then the median obtained is 'y'. Calculate the value of (x + y).

Q.5 The mean of 5 observations is 3 and variance is 2. If three of the five observations are 1, 3, 5, find the other two.

Q.6 Show that for the given (n+1) numbers,

$$\bar{X}_{(n+1)} = \frac{(n * \bar{X}_n + Xn+1)}{(n+1)}$$

If the average monthly spending by 21 women in a kitty group was Rs. 5240, what is the new average spending if another member is added whose average monthly spending is 5540? Use the formula above to answer.

Q.7 The sum of deviations of a certain number of observations from 12 is 166 while the sum of deviations of these observations from 16 is (-54). Find the number of observations and their mean.

Q.8 Consider the following data that depicts the daily income of two ice cream sellers in two different regions:

Region	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
A	500	900	800	900	700	400
B	300	540	480	540	420	240

Calculate the coefficient of variation for each region and compare your result.

Q.9 If the coefficient of skewness of a distribution is 0.32, the standard deviation is 6.5 and the mean is 29.6 then find the mode of the distribution.

Q.10 There are 40 students in a class preparing for a statistics test. There are two strategies that these students can adopt to ace the test. After the test, the professor interviewed the students and noted down their strategies. The professor observed that 20 students followed strategy A and the other 20 followed strategy B. Given below are the marks of the students based on the strategies they adopted:

Strategy A: 78, 78, 79, 80, 80, 82, 82, 83, 83, 86, 86, 86, 86, 87, 87, 87, 88, 88, 88, 91

Strategy B: 66, 66, 66, 67, 68, 70, 72, 75, 75, 78, 82, 83, 86, 88, 89, 90, 93, 94, 95, 98

Write down the 5-number summary for both the strategies. Create two boxplots for each strategy using the 5-number summary and comment on the shapes of boxplots. According to you, which strategy is more likely to fetch you good marks in the test and why?

3.12 REFERENCES

- Devore, J. L. (2016). *Probability and Statistics for Engineering and the Sciences*. Cengage learning.
- Larsen, R. J., & Marx, M. L. (2012). *An introduction to mathematical statistics and its applications*. Prentice Hall.
- McClave, J. T., Benson, P. G., & Sincich, T. (2018). *Statistics for business and economics*. Pearson Education.

3.13 SUGGESTED READING

- Gupta, S. C. (2019). *Fundamentals of Statistics*. New Delhi, India: Himalaya publishing house.

LESSON 4

SAMPLE SPACE, EVENTS, AND PROBABILITY

STRUCTURE

4.1 LEARNING OBJECTIVES

After reading this lesson, students will be able:

1. To understand the concept of sample space and population and their significance.
2. To comprehend the need for the concept of sample and population in the context of probability.
3. To understand the concept of probability in the context of random experiments.
4. To visualize the applications of probability in real life and understand the definition of probability.
5. To be able to differentiate between the sample space, events, sample points, and random experiments.
6. To get familiarized with the technique of the Venn diagram, its usage in defining events, types of events and
7. To understand the properties of probability and various operations to comprehend the working of probabilities.

4.2 INTRODUCTION

This unit introduces the concept of 'probability' to the students. The phenomenon of probability indicates the presence of randomness and the existence of some element of uncertainty. Whenever we face a situation in which there is more than one possible outcome that can occur, the concept of probability renders a technique for quantifying the chances or likelihood associated with every possible outcome. There are several instances that involve chances and thus the notion of probability is applicable. For example, in political elections, based on exit polls it is plausible to predict that a certain political party could come into power. By deploying a database of the previous days and considering various parameters such as temperature, humidity, pressure, etc., the meteorologists use specific tools or techniques to predict weather forecasts and determine that there are 60 out of 100 chances that it would rain today.

Another example from day-to-day life is that 'since it is supposed to rain tomorrow, it is very likely I will use my raincoat when I go to work. Similarly, flipping a coin involves the probability of getting either a head or a tail is 0.5 and playing with dice involves one out of six chances that the required number will come. Thus, the concept of probability can be applied to several interesting events.

Probability is a mathematical term and the study of probability as a branch of mathematics is over 300 years. This chapter enables the students to understand and estimate the likelihood of

various possibilities of events and outcomes. Various elementary concepts used in comprehending the concept of probability will be discussed and explained, such as Sample, population, random experiments, Venn diagram, sample points, events, types of events etc.

4.3 SAMPLE AND POPULATION

The discipline of Statistics deals with organizing and summarizing data for drawing conclusions based on the information collected in the form of data. An investigation or experiment that results in a well-defined collection of objects, constitutes what is known as **'Population'.**

There can be several types of population. One study on a particular type of medicine will lead to a collection of particular capsules during a specified period. Another investigation might involve a population consisting of students getting enrolled in BA honors Economics. If the desired information is available for all the objects in the population, it is called a **'census'.**

A subset of the population is considered as a '**sample'.** A sample is selected in some prescribed manner. "Sample is a means to an end rather than the end itself". The technique for generalizing from a sample to a population gathered within the branch of our discipline called "Inferential Statistics".

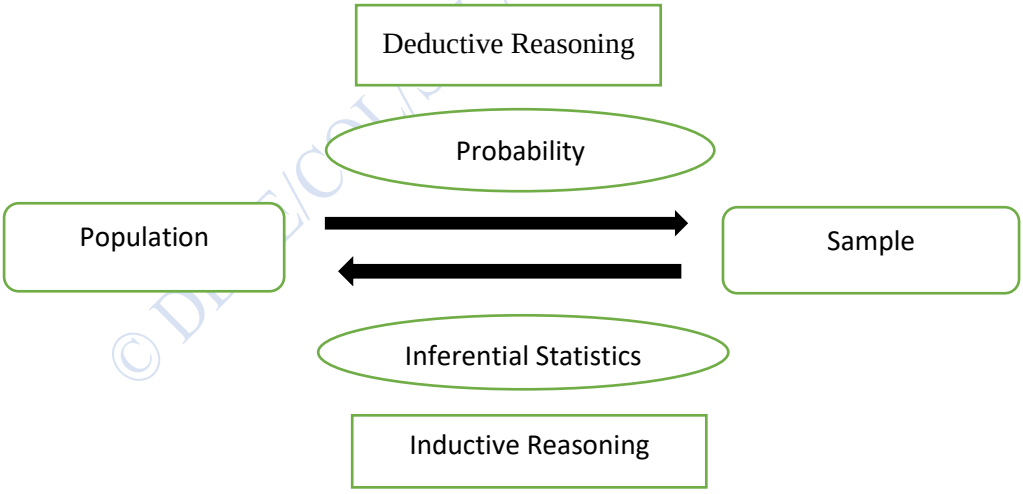


Figure 1: Relationship between Population and Sample 'a two-way process.

It can be visualized from the figure above that a sample and a population both can be deployed to examine and assess the data also called 'inference'. There are two fundamental approaches for inference, **deductive and inductive reasoning.**

When a sample is derived from the given population, then the concept of probability is used to infer anything regarding the population. This method of inference is called **deductive reasoning.** However, when the sample is used to deduct or infer the population, inferential statistics is deployed for inferring the population. The technique is referred to as **'inductive reasoning.** Thus, the role of probability is explicit and well-defined as it plays a critical role in inferring the sample derived from the population. It is crucial in the deductive method of statistical inference or research.

Having understood the difference between the sample and population and the relationship between them, also their role played in statistical inference, it is crucial to comprehend the kind of experiments or data collection.

4.3.1 Statistical or Random Experiments

Any activity or process whose outcome is subject to uncertainty is considered an experiment. Experiments generally suggest careful controlled testing of the situation or planned testing in the laboratory. However, in the disciple of statistics, experiments refer to a wider scope of trials such as tossing a coin once or several times, selecting a card from the deck, obtaining a particular blood type from a group of individuals, etc.

Any process of observation or measurement that has more than one possible outcome and for which there is uncertainty about which outcome will actually materialize is referred to as a 'random experiment'. For example, tossing a coin, throwing a pair of dice, drawing a card from the deck of cards.

4.3.2 Sample Point, Event

Each member or outcome of a sample space or population is called **Sample Point and event.** It is also called an element of sample space. Let us consider the example of the toss of the coin for which the sample space is S = {H, T}. The number of elements in the sample space or population is n(S) = 2. Each element of the sample space that is H and T are known as a Sample point. In general, n(S) is the number of sample points, a number of times the experiment is repeated.

Consider an event B which is defined as Event B: Tail appears: B={T}. The number of elements in event B is 1, denoted by n(B)=1

In a random experiment of the toss of a coin, suppose the event A denotes the event that Head appears. A = {H}. The number of elements in event A is 1, denoted by n(A)=1

A+B = S: {H}+{T} = {H,T} = S

Let us consider another example of **tossing two fair coins.** The sample space or population for this experiment is given by Sample Space: {HH, HT, TH, TT}

The number of elements in the sample space is 4, denoted by n (S) = 4.

Consider an event B that at least the head appears on one of the coins in the toss of two coins simultaneously. Event B can be represented as

B= {HH, HT, TH},

The number of elements in event B is 3, represented by n(B) = 3

Trial & Events: An experiment is repeated under essentially identical conditions but does not give unique results. It may result in several possible outcomes. The experiment is called a Trial and the outcomes are called events. For example, throwing a coin once is an experiment, and getting a Head or Tail is an event. Planting a sapling is a Trial and whether it survives, or dies is an Event. Sitting for an examination is a Trial and getting grades such as A, B, C, D, and E are events.

Exhaustive Events: All possible outcomes of an experiment constitute collectively exhaustive events. For example, tossing a coin result in two exhaustive cases which are Head and Tail. Planting a sapling leads to two exhaustive cases which are Survival and Death. Sitting for an examination where a student is awarded only 5 grades results in those many exhaustive numbers of cases.

Favourable Events: All those outcomes of an experiment that lend themselves to the objectives or favour of the experiments are favourable events. For example, a gambler betting on an Ace in a game of cards where every draw of cards decides the winner or loses has 4 favourable events, and betting on a black card has 13+13 = 26 favourable events.

Mutually Exclusive Events: Events are said to be mutually exclusive if happening of one event prevents the occurrence of other events at the same time. Such events are also referred to as disjoint events since they have no element in common. For example, in athletics meet involving 10 challengers if any one of them wins then the remaining 9 winning cannot happen and hence are mutually exclusive. Similarly, in a toss of coin, occurrence of Head or Tail are mutually exclusive.

Equally Likely Events: Two events are said to be equally likely if one of them is as likely to happen as the other. For example, in tossing a fair coin once, the outcomes Head and Tail are equally likely. In a throw of 6-faced dice, all the six numbers 1,2,3,4,5,6 are equally likely. If a person suffers a minor heart attack, the death or survival outcomes are not equally likely.

Independent Events: If the happening of one event is not affected by the happening (or not happening) of another event, such events are said to be independent. For example, successively throwing a dart on the dartboard and getting a perfect score in every throw are independent

events. However, a person throwing the dart once, practicing, and then throws it for the second time. The event of getting a perfect score in both throws is not independent.

Example: 1 **Trial:** Tossing of one fair coin

Events: Occurrence of Head, the occurrence of Tail.

Exhaustive events: Occurrence of Head

Mutually exclusive events: Head and Tail

Equally Likely Events: Head and Tail

Example 2: **Trial:** Tossing of Two fair coins

Events: Occurrence of Two Heads

Occurrence of One Head

Occurrence of Zero Head

Exhaustive Events: HH, HT, TH, TT

Favourable Events: a) HH

b) HT, TH

c) TT

Mutually exclusive event: Occurrence of Two Heads and Occurrence of Two Tails

Equally likely events: a) getting at least one Head (HH, HT, TH) is equally likely as getting at least one Tail (TT, TH, HT)

b) getting both heads is not equally likely as getting at least one head.

Independent Events: getting a Head in the second Toss is independent of getting a Head in the first Toss.

4.3.3 Population or Sample Space of an Experiment

The set of all possible outcomes of an experiment is called **Population** or simply **Sample Space**, denoted by S. Let us consider an example of **tossing one fair coin**. This is an example of a random experiment since this involves two plausible outcomes. A head or a tail can appear

in a single toss of a fair coin. For such an experiment the total number of outcomes is two, therefore the sample space is denoted by

The sample space: S = {H, T},

The number of elements in the sample space or population is n(S) = 2

Either both tosses result in a Head or both Tosses result in a Tail or the first Toss result in a Head while the second results in Tail or the first Toss results in a Tail and the second results in a Head.

Let us consider another example of **tossing two fair coins.** The sample space or population for this experiment is therefore given by

Sample Space: {HH, HT, TH, TT}

The number of elements in the sample space or population is 4, n (S) = 4.

Consider another example of **rolling a die,**

Sample space S = {1,2,3,4,5,6},

The number of elements in the sample space is n (S) = 6

And if the same dice is rolled twice, n(S)= 36 = 6^2 is the Sample space.

If rolled thrice, n(S) = 216 = 6^3 is the Sample Space.

A CASE STUDY

Consider another example of rolling a dice, The sample space for the random experiment of rolling dice is given by the Sample space S = {1,2,3,4,5,6},

The number of elements in the sample space is 6, denoted by n (S) = 6,

Let event E be an event that reflects even numbers that appear on dice, as represented by

E= {2, 4, 6},

The number of elements in event E is 3, represented by n(E)=3

There are several varieties of events as described in the next section.

IN-TEXT QUESTIONS

1. Events are said to be _____. if the occurrence of one event prevents the occurrence of another event at the same time.

2. If event A represents an event that at least a head appears, and event B represents an event that only the tail appears. Events A and B are equally likely True / False

3. In the occurrence of the event: {Head} in a single throw of the coin, the occurrence of event {Tail} is disjoint. The two events are called

   a) Mutually exhaustive b) Equally likely c) Both

4. In an experiment consisting of tossing two coins, if event A represents an Event that at least a Head occurs and event B represents that at least a Tail occurs, then

   a) Events A and B are equally likely (True/False)

   b) Events A and B are mutually exclusive (True/False)

   c) Events A and B together form an exhaustive set (True/False)

4.3.4 Events, Set theory and Venn Diagrams

An event can be considered a set, therefore the relationships and results from elementary set theory can be used to study events of any random experiment. Some of the fundamental operations of set theory can therefore be applied to events such as.

1. The complement of an event A is denoted by A'. A complement represented as A' is the set of all outcomes in the sample space S that are not contained in set A.
2. The union of the two events A and B is denoted by A ∪ B. A union B can also be read as "A or B" or in both events. In other words, the union of two events includes outcomes for which both A and B occur as well as outcomes for which exactly one occurs. It means all outcomes in at least one of the events.
3. The intersection of the two events, A and B, denoted by A ∩ B is read as "A and B". The intersection of two events indicates an event consisting of all outcomes that are in both A and B.
4. A null event is an event consisting of no outcomes whatsoever and is denoted by ∅. Suppose there are two events A and B, and it is given that A ∩ B = ∅. then A and B are said to be mutually exclusive or disjoint events.

4.3.5 De Morgan's laws

a. The complement of the union of events A and B is equal to the intersection of the complement of A and the complement of B.

$(A \cup B)' = A' \cap B'$

b. The complement of the intersection of event A and B is equal to the union of the complement of A and the complement of B.
$(A \cap B)' = A' \cup B'$

The events can be represented by using the Venn diagram as shown in the diagrams below.

Fig 1: Population or Sample Space

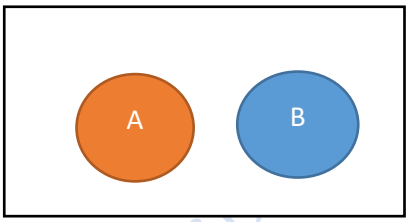


Fig 2: Event A and B are disjoint

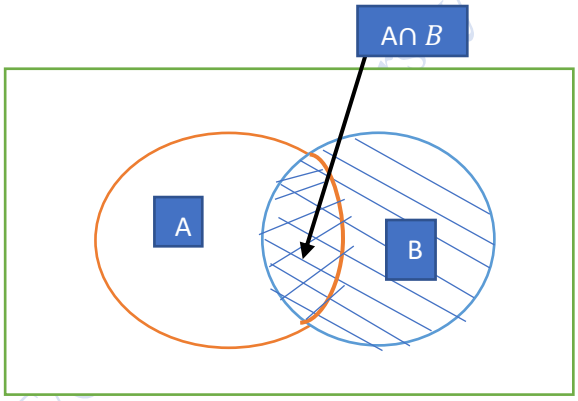


Fig 3: Events A and B are not disjoint

All elements in the sample space belong to the rectangle that represents the entire population as shown in figure 1. Event A is represented by the oval in orange colour and event B is represented by the oval shape in blue inside the rectangle, as shown in figure 2. The rectangle is the population of sample space and events A, and B are the subsets of the sample space. Events A and B have nothing in common, such events are referred to as disjoint events. These events are also referred to as mutually exclusive events. Event A is represented by the oval in orange colour and event B is represented by the oval shape in blue inside the rectangle. The rectangle is the population of sample space and events A, and B are the subsets of the sample space. In this case events, A and B have common elements therefore, events A and B are not disjoint sets.

IN-TEXT QUESTIONS

1. Consider an experiment in which each of the three vehicles taking a particular freeway exit turns left (L) or right (R) at the end of the exit ramp. Outline the sample space and events.

2. The two events E_1 and E_2 are mutually exclusive, where E_1 is the event consisting of numbers less than 3 and E_2 is the event that consists of numbers greater than 4. (True/ False)

3. If the two events have some common elements, the two events are not _____.

4.4 PROBABILITY

In the realm of random experiments, the key objective of the probability of any event A is to assign a number P(A) to event A. This value P(A) is called the probability of event A which gives a unique measure of the chances that the event will occur.

In other words, the probability is the chance of happening or occurrence of an event such as it might rain today, team X will probably win today, or I may win the lottery. Largely, probability is a measure of uncertainty.

4.4.1 Classical Definition of Probability

It is also called a priori or mathematical definition of probability. The probabilities are derived from purely deductive reasoning. This implies that one does not throw a coin to state that the probability of obtaining a head, or a tail is ½. However, there are cases where possibilities that arise cannot be regarded as equally likely. For example, the Probability of a recession next year Probability of GDP value next year. Similarly, the possibility of whether it will rain, or the outcome of an election is not equally likely.

If an experiment results in mutually exclusive and equally likely outcomes. If m outcomes are favorable to event A and n is the total number of outcomes in the sample space, then

$$P(A) = \frac{m}{n}, \quad \frac{number\ of\ outcomes\ favourable\ to\ A}{Total\ number\ of\ outcomes}\ or$$

$$= \frac{Favourable\ number\ of\ Events}{Exhaustive\ number\ of\ Events} = \frac{n(A)}{n(S)}$$

In a single throw of a die, the total occurrences or sample space is n = 6. All are mutually exclusive and equally likely.

4.4.2 Relative Definition of Probability (by Von Mises)

If a trial is repeated a large number of times under essentially homogeneous and identical conditions, then the limiting value of the relative frequency which is the ratio of absolute frequencies to the total number of occurrences is called the probability of happening of events.

$$P(A) = \lim_{n\to\infty} \frac{m}{n}$$

IN-TEXT QUESTIONS

6. In a toss of two coins simultaneously, the probability of getting exactly 2 heads P(E) no. of possible outcomes / total outcomes

7. In the toss of 3 coins simultaneously, the probability of getting exactly two heads.

8. What is the probability of getting at least 1 head when two coins are tossed simultaneously?

9. Prob of getting almost 2 tails when three coins are tossed simultaneously.

10. Probability of getting at least 2 heads when three coins are tossed simultaneously.

11. Probability of getting a greater number of tails than heads when three coins are tossed simultaneously.

4.4.3 Axiomatic Definition of Probability

The axiomatic approach to probability was provided by Russian Mathematician A.N. Kolmogorov and includes both the above definitions. In order to ensure that the probability assignments of values P(A) for a particular event in the sample space S, is consistent with the intuitive notion of probability, all assignments of values of probability P(A) must satisfy the following properties or Axioms.

1. For any event A, the probability of event A, given by P(A) is non-positive $P(A) \geq 0$. In other words, the probability that event A will occur can either be zero or some positive number. The probability of event A can never be negative.

   The Axiom 1 reflects the intuitive notion that the chance of A occurring should be non-negative and is known as the Axiom of non-negativity.

2. The probability of the entire sample space is 1, that is P(S) = 1. In other words, the probability that the entire sample space will occur is 100 percent, which means it will surely occur. This is known as the Axiom of Certainty.

The sample space by definition is the event that must occur when the experiment is performed. The sample space S contains all possible outcomes, therefore the maximum possible probability is assigned to sample space S.

3. If $A_1, A_2, A_3, \ldots\ldots\ldots$ are the infinite collection of disjoints events, then
$P(A_1 \cup A_2 \cup A_3, \ldots\ldots\ldots) = \sum_{i=1}^{\infty} P(A)$

This indicates that the probability of the union of all disjoint events belonging to the sample space sums the chances of all individual events.

The third Axiom formalizes the idea that if we wish the probability that at least one of a number of events will occur, given that no two events can occur simultaneously, then the chance of at least one occurring is the sum of the chances of the individual events. This is known as the axiom of finite additivity.

4. The probability of an event always lies between 0 and 1.
$0 <= P(A) <= 1$,

$P(A) = 0$ means event A will not occur.

$P(A) = 1$ means event A will occur certainly.

5. Let the ∅ be the null event. The event contained no outcomes whatsoever. This property mainly reflects Axiom 3 indicating the finite collection of disjoint events.

Therefore, $P(\emptyset) = 0$, the probability of a null event is zero.

6. If A, B, and C are mutually exclusive events, the probability that any one of them will occur is equal to the sum of probabilities of either individual occurrence.
P(A+B+C+.....) = P (AUBUC…….) = P(A) +P(B)+P(C) +.....

7. If A, B, C …… are a mutually exclusive and collectively exhaustive set of events the sum of the probability of their individual occurrences is 1. However, if A, B, C …… are any events, they are said to be statistically independent if the probability of their occurring together is equal to the product of their individual probabilities. P(A∩B∩C) = Probability of events A, B, and C occurring together or jointly or simultaneously, also referred to as Joint probability.

P(A), P(B), and P(C) are called unconditional marginal or individual probabilities.

8. If events A, B, and C …...are not mutually exclusive then,
P(A+B) or P(AUB) = P(A) +P(B) - P(A∩B)

Where P(AB) is the joint probability that the two events occur simultaneously, that is P (A∩ B). However, if A and B are mutually exclusive then,

P(A∩B) = P (∅) = 0

For every event A, there is an event A', called as a complement of A

$P(A + A') = P(A \cup A') = P(S) = 1$

$P(A\ A') = P(A \cap A') = P(\emptyset) = 0$

IN-TEXT QUESTIONS

12. An unbiased dice is thrown. What is the probability of getting
    (i) a multiple of 3
    (ii) a number less than 5
    (iii) an even prime number
    (iv) a prime number
    (v) a factor of 6
13. A dice is thrown once, find the probability of getting
    a) An odd number
    b) A multiple of 3
    c) A factor of 5
14. Two dice are thrown together, find the probability of getting
    a) An even number on both
    b) Sum as a perfect square
    c) Different numbers on both
    d) A total of at least 10
    e) Sum as a multiple of 3
    f) A multiple of 2 on one and a multiple of 3 on other
    g) Sum as an even number

4.5 SUMMARY

This lesson familiarized the students with the basic concepts of sample space and population along with their significance. The notion of probability was introduced with help of random experiments. Various applications of probability in real life are presented in the chapter. Certain important concepts related to probability such as space, events, sample points, and random

experiments are described in the chapter. The basic difference between the sample, population, sample points, and events have been emphasized. The types of events such as disjoint events, mutually exhaustive, and exclusive events have been explained. Further, the concept of the Venn diagram is also presented in the chapter. The notion of probability by using classical and relative definition has been introduced. Later the properties of probabilities are also discussed in the chapter.

4.6 GLOSSARY

1. **Sample:** "Sample is a means to an end rather than the end itself".
2. **Population:** An investigation or experiment that results in a well-defined collection of objects, constitutes what is known as **'Population'.**
3. **Deductive Reasoning**: When a sample is derived from the given population, then the concept of probability is used to infer anything regarding the population. This method of inference is called **deductive reasoning.**
4. **Inductive Reasoning:** When the sample is used to deduct or infer the population, inferential statistics is deployed for inferring the population. The technique is referred to as '**inductive reasoning.**
5. **Random Experiment:** Any process of observation or measurement that has more than one possible outcome and for which there is uncertainty about which outcome will actually materialize. Such an experiment is referred to as **'random experiment'.**
6. **Sample Point or Event:** Each member or outcome of sample space or population is called Sample Point. It is also called an element of sample space.
7. **Mutually Exclusive**: Events are said to be mutually exclusive if the occurrence of one event prevents the occurrence of another event at the same time. Such events are also referred to as disjoint events since they have no element in common.
8. **Equally Likely:** The events are called **equally likely** when two events are said to be equally likely if one event is as likely to occur as the other.
9. **Collectively Exhaustive:** The events are collectively exhaustive if the events exhaust all possible outcomes of an experiment.
10. **De Morgan's Law:** The complement of the union of two sets A and B is equal to the intersection of the complement of the sets A and B. This is **De Morgan's first law.**

4.7 ANSWERS TO IN-TEXT QUESTIONS

1. Mutually Exclusive

2. False
3. Both
3 The sample space S; {LLL, RLL, LRL, LLR, LRR, RLR, RRL, RRR}
The event that exactly one of the three vehicles turns right: A
The elements in event A: {RLL, LRL, LLR}
The event that at most one of the vehicles turns right: B
The elements in the event B: {LLL. RLL, LRL, LLR}
In the event that all three vehicles turn in the same direction: C
The elements in the event C: {LLL, RRR}
4. E1 = {1,2}, E2 = {5,6}. The two events are mutually exclusive. True
5. Disjoint
6. ¼
7. 3/8
8. ¾
9. 7/8
10. ½
11. ½
12. Total number of possible outcomes = 6= n(S)
    (i) a multiple of 3
    Number of favorable outcomes = 2 {3 and 6}
    Hence P (getting multiple of 3) = 2/6 = 1/3
    ii) a number less than 5
    Number of favorable outcomes = 4 {1, 2, 3, 4}
    Hence, P (getting number less than 5) = 4/6 = 2/3
    iii) an even prime number
    Number of favorable outcomes = 1 {2}
    Hence, P (getting an even prime number) = 1/6
    iv) a prime number
    Number of favorable outcomes = 3 {2,3,5}
    Hence the P (getting a prime number) = 3/6 = 1/2

v) a factor of 6

Number of favorable outcomes= 4 {1, 2, 3, 6}

Hence, P (getting a factor of 6) = 4/6 = 2/3

13. a) 1/2 b) 1/3 c) 1/3

14. a) 1/4 b) 7/36 c) 5/6 d) 1/6 e) 1/3 f) 11/36 g) ½

4.8 SELF-ASSESSMENT QUESTIONS

1. Two six-faced dice are rolled together, or dice is rolled twice. The total number of possible outcomes are 36.
2. (i) Prove that the probability of null event is zero, P (Ø) = 0.

   (ii) Prove that for any two events A and B

   P(AUB) = P(A) +P(B) - P(AB)

4.9 REFERENCES

- Devore, J. L. (2015). Probability and Statistics for Engineering and the Sciences. Cengage Learning.
- Freund, J. E., Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics: With Applications*. Pearson Education India.
- McClave, J. T., Benson, P. G., & Sincich, T. (2008). *Statistics for business and economics*. Pearson Education.

4.10 SUGGESTED READINGS

- Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.
- Larsen, R. J., & Marx, M. L. (2005). *An introduction to mathematical statistics*. Prentice Hall.

LESSON 5

CONDITIONAL PROBABILITY

STRUCTURE

5.1 LEARNING OBJECTIVES

1. To understand the concept of conditional probability and its significance in real life.
2. To comprehend the significance of the initial assignment of probability that may be followed by partial information relevant to the outcome.
3. To visualize that partial information may affect the assignment of probability assignment. This leads us to the concept of conditional probability and Bayes' Theorem.
4. To comprehend the concept of Bayes' Theorem and its applications.
5. Learning the computation of a posterior probability from the given prior probabilities and conditional probabilities plays a critical role in Bayes' Theorem and

6. To practice several cases and examples for the application of Bayes' Theorem.

5.2 INTRODUCTION

In the previous chapter, we introduced the topic of probability. In this chapter, we expose students to the deeper concepts and situations related to probability. The probabilities assigned to various events or occurrences are subject to what is known as experimental situations. The initial assignment may be followed by partial information relevant to the outcome. The partial information may affect the assignment of probability assignment. This leads us to the concept of conditional probability and Bayes' Theorem.

Conditional probability is considered a measure of the likelihood of an event occurring, assuming that another event or outcome has previously occurred. For instance, a student aims to receive an academic scholarship while applying for admission. For every 1000 applications, the college accepts 100 applications and awards an academic scholarship to 10 of every 500 students. Of the scholarship recipients, 50 % receive university stipends for books, meals, housing etc. As a result, the chance of students being accepted and then receiving a scholarship is 2% given by 0.1 multiplied 0.02. While the chance of being accepted, receiving the scholarship and then receiving the stipend for books etc. is 0.1 & given by 0.1 multiplied by 0.02 multiplied by 0.5.

In the realm of conditional probability, Bayes' Rule or Bayes' Law is used to calculate the conditional probability. Bayes' theorem is a mathematical equation that helps calculate conditional probability. The computation of a posterior probability from the given prior probabilities and conditional probabilities plays a critical role in Bayes' Theorem.

5.3 CONDITIONAL PROBABILITY

Conditional probability is defined as the likelihood of an event or an outcome occurring based on the occurrence of a previous event or outcome. The conditional probability is condition upon the occurrence of some event that has happened earlier. Therefore, the conditional probability is computed by multiplying the probability of the preceding event by the updated probability of the succeeding or conditional event.

For a particular event A, we have used P(A) to represent the probability of event A. The probability P(A) can be considered as unconditional probability, which simply implies that the probability of occurrence of event A does not depend on anything. Suppose now we introduce another event B. An occurrence of this event B affects the probability assigned to event A. In other words, the probability of event B affects the probability of event A. To represent the probability of event A such that event B has already occurred be considered as P(A/B). This represents the conditional probability of A given that event B has already occurred. Here, B is the conditioning event.

For two events A and B,

$$P(AB) = \begin{cases} P(A) * P(B \mid A), P(A) > 0 \\ P(B) * P(A \mid B), P(B) > 0 \end{cases}$$

Where P (B | A) represents the conditional probability of occurrence of B when A has already occurred, and P (A | B) is the conditional probability of occurrence of A when B has already occurred.

Let us take an example of the virus COVID-19. Event A might refer to an individual infected with the COVID-19 virus in the presence of the symptoms. However, if the blood test is performed on that individual and the result is negative. Consider this situation as event B where B reflects a negative Blood test. Thus, the probability of having COVID-19 will change, that is Probability of event A occurring will change, or P(A) will change. It is natural to think that P(A) will decrease but it may not be zero because the blood test is not fully reliable.

Another example is that suppose a student who applies for a college will be accepted is event A. There is an 80% chance that students will be accepted to college. Suppose event B is that the student will be given hostel accommodation in that college. The hostel accommodation will only be provided for 60% of all accepted students. The probability that event B occurs provided event A has occurred is given by P(B/A) given by P(AB) * P(A) which is 0.6*0.8 which is equal to 0.48.

5.3.1 Computation of Conditional Probability

If there are two events, A and B, then the probability that event A occurs knowing that event B has already occurred. This is called conditional probability of A, conditioned that event B has already occurred.

Conditional Probability is denoted by: P (A | B)

$$P(A \mid B) = \frac{P(AB)}{P(B)}, \quad P(B) > 0$$

Where P(AB) is the joint probability of events A and B occurring together. Also, read as the probability of events A and B or P(A∩B) or probability of event A intersection B.

The computation of conditional probability P (A | B) requires that the P(B) is positive or the probability of event B occurring cannot be negative or zero.

Similarly, the conditional probability of B, given A, is denoted by P (B | A)

$$P(B \mid A) = P(AB)/ P(A), \quad P(A) > 0$$

$$P(AB) = P(B \mid A) * P(B), \quad P(A) > 0$$

Where P(AB) is the joint probability of events A and B occurring together. Also, read as the probability of events A and B or P(A∩B) or probability of event A intersection B. The computation of conditional probability P (A | B) requires that the P(B) is positive or the probability of event B occurring cannot be negative or zero.

In the case of Conditional Probability is denoted by: P (A | B)

$$P(A \mid B) = P(AB)/ P(B), \qquad P(B) > 0$$

The conditional probability is expressed as a ratio of the unconditional probabilities. The numerator is simply the probability of the intersection of the two events, whereas the denominator is the probability of the conditioning event B. The conditional probability can be represented by the following Venn diagram.

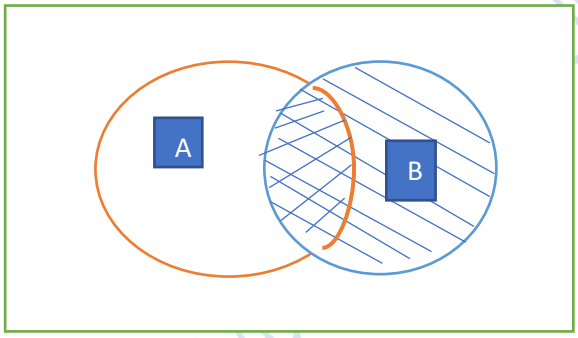


Figure 1: Conditional Probability

Suppose event B is given and it has already occurred, as a result, the sample space is no longer the entire population, but consists of outcomes in B. Event A has occurred if and only if one of the outcomes in the intersection occurred. Therefore, the conditional probability of A given B is proportional to P(A∩B).

The ratio given by 1/P(B) is considered the proportionality constant. This ratio ensures that the P(B/B) is the probability of a new sample space due to the condition that event B has already occurred and the probability P (B | B) of the new sample space is equal to 1.

For instance, A box contains 20 TVs, of which 5 are defective. If 3 of the TVs are selected at random and removed from the box in successive without replacement, what is the probability that all three fuses are defective?

Let us consider event A that the first TV is defective, B is the event that the second TV is defective, and C is the event that the third TV is defective then, given that

$$P(A) = 5/20,\ P(B \mid A) = 4/19,\ P(C \mid A \cap B) = 3/18$$

$$P(A \cap B \cap C) = P(A) * P(B \mid A) * P(C \mid A \cap B) = 5/20 * 4/19 * 3/18 = 1/114$$

Points to remember

(a) Two possible mutually disjoint events are always dependent.

Proof: Let A and B be disjoint events i.e. $A \cap B = \emptyset$

So, $P(A \cap B) = 0$

We know that $P(A \cap B) = \begin{cases} P(A)*P(B|A), & P(A) \neq 0 \\ P(B)*P(A|B), & P(B) \neq 0 \end{cases}$

Since $P(A) \neq 0$ and $P(B) \neq 0$

Implies $P(B|A) = 0$ and $P(A|B) = 0$

Implies A and B are dependent events.

(b) Two possible and independent events cannot be mutually disjoint.

Proof: Let A and B be two independent events such that both are possible i.e.

$P(A \cap B) = P(A)*P(B)$ such that $P(A) > 0$, and $P(B) > 0$

Implies $P(A \cap B) \neq 0$

Hence, A and B cannot be disjoint.

For example, in a management class, let us define two independent events namely choosing a tall person and choosing an intelligent person, for the sample Prove that the events of choosing a short person and choosing a Moron are also independent.

Let us define, set A as choosing a tall person and set B as choosing an intelligent person.

Then, P(AB) = P(A)*P(B) as A and B are independent.

Now, $P(A' \cup B') = P[(A \cup B)'] = 1 - P[A \cup B]$

$= 1 - [P(A) + P(B) - P(AB)]$

$= 1 - [P(A) + P(B) - P(A) * P(B)]$

$= [1 - P(A)][1 - P(B)]$

$= P(A') P(B')$

Here, A' (choosing a short person) and B' (choosing a Moron) are also independent.

CASE STUDY

Suppose that of all the individuals who buy the smartphone, 60% include an optional memory card in their purchase, 40% include an extra battery and 30% include both a card and a battery.

Solution:

Let us consider a randomly selected buyer, and let event A be the memory card purchased, while event B be the battery purchased. Thus, P(A) is 0.6, P(B) is 0.4. The probability that both memory card and battery are purchased, P(A∩B) or P(AB) is 0.3.

Given that the selected individual purchased an extra battery, the probability that an optional card was also purchased is P (memory card/battery) or P(A/B).

$$P(\text{memory card/battery}) = P(A|B) = P(A \cap B)/P(B), \quad P(B) > 0$$

$$= 0.3/0.4 = 0.75$$

This implies of all those purchasing an extra battery, 75% purchased an optional memory card.

Similarly, given that the memory card was purchased, the probability of buying battery is given by P (battery/memory card) or P(B/A).

$$P(\text{battery}|\text{memory card}) = P(B|A) = P(A \cap B)/P(A), \; P(A) > 0$$

$$= 0.3/0.6 = 0.5$$

It can be observed that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$

This means conditional probability is not equal to unconditional probability.

IN-TEXT QUESTIONS

1. A card is drawn from a deck of cards. What is the probability that it will be either a heart or a queen?
2. The numerator is the union of two events in the computation of conditional probability. (True/false)
3. In a class, there are 500 students of which 300 are, males and 200 are females. Of these 100 males and 60 females plan to major in accounting. A student is selected at random from this class and it is found that this student plans to be an accounting major. What is the probability that the student is a male?
4. If there are two events, A and B, then the probability that event A occurs knowing that event B has already occurred is referred to as _____.

5. If we randomly pick two TV sets in succession from a shipment of 240 T.V tubes of which 15 are defective. What is the probability that they will both be defective?

5.4 BAYES' THEOREM

Bayes' Theorem is primarily a mathematical formula for computing conditional probability and was named after 18th Century British mathematician.[1] This theorem is also known as Bayes' Rule or Bayes' Law and is also considered the foundation of Bayesian statistics. As discussed in the above section, conditional probability indicates the likelihood of a particular outcome occurring, based upon the results of an earlier or previous event that has already occurred. There is a wide range of applications of Bayes' Theorem in the field of finance such as the risk of lending money to borrowers. Furthermore, Bayes' Theorem plays an instrumental role in the implementation of machine learning.

There are many situations where the outcome of the experiment is conditional upon or depends on the outcomes associated with various intermediate stages. To comprehend such intermediate stages let us consider one example.

Suppose the completion of a construction assignment may be delayed because of some political emergency such as a curfew. Suppose there are 0.60 probabilities that there will be a political emergency, 0.85 that the construction assignment will be completed on time if there is no emergency, and 0.35 that the construction work will be completed on time if there is a political emergency. What is the probability that the construction assignment will be completed?

Solution: Let us assume that A is an event that the construction assignment will be completed on time and B is the event that there will be a political emergency. It is given that P(B) is 0.60. The probability of event A occurring such that B does not occur P(A/B') is 0.85 while the probability of event A occurring such that event B has already occurred P(A/B) is 0.35.

By using the formula $P(A) = P[(AB) \cup AB')]$

$= P(A \cap B) + P(AB')$

$= P(B).\ P(A|B) + P(B').\ P(A|B')$

$P(A) = (0.60) * (0.35) + (1 - 0.60) * (0.85) = 0.55$

Such a case can be generalized where the intermediate stage permits k different alternatives denoted by $B_1, B_2, B_3 \ldots\ldots\ldots\ldots B_k$. The following theorem connects these intermediate stages by what is known as **the rule of total probability or the rule of elimination.**

The B's constitute a partition of the sample space if they are pairwise mutually exclusive and if their union equals S as shown in figure 2. B_i's are mutually exclusive and exhaustive, if A occurs it must be in conjunction with exactly one of the B_i's. This mainly implies $A = (B_1 \cap A)$

[1] https://www.investopedia.com/terms/b/bayes-theorem.asp

∪ .....∪ ($B_k \cap A$), where all the events ($B_i \cap A$) are mutually exclusive. This "partitioning of A" is illustrated in figure 2 below.

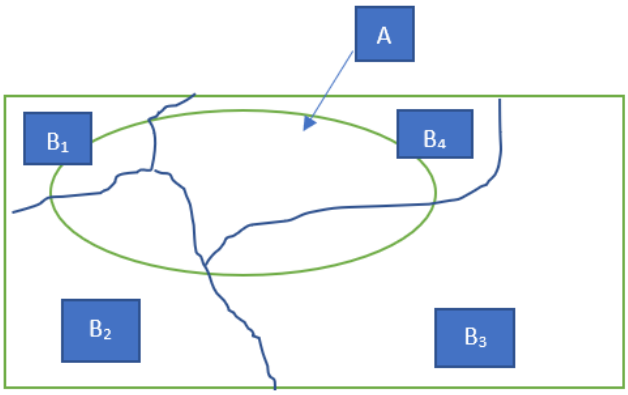


Figure 2: Partitioning of A by mutually exclusive and exhaustive A_i's.

Therefore, if events B_1, B_1, B_1…… B_k are constituting a partition of sample space S and P(Bi) = 0 for all i = 1,2,3, ...k the for any event A in S

$P(A) = \sum_{i=1}^{k} P(B_i).P(\frac{A}{B_i}))$

For any event A in S, such that $P(A) \neq 0$

$P(B_r/A) = \frac{P(Br).P(\frac{A}{B_r})}{\sum_{i=1}^{k} P(Bi).P(\frac{A}{B_i}))}$, for all r = 1, 2,3,……..k

$\sum_{i=1}^{k} P(B_i).P(\frac{A}{B_i}) = P(B_1).P\left(\frac{A}{B_1}\right) + P(B_2).P\left(\frac{A}{B_2}\right) + P(B_3).P(\frac{A}{B_3}) + \ldots\ldots P(B_k).P(\frac{A}{B_k})$

$P(B_r/A) = \frac{P(B_r \cap A)}{P(A)}$

In the figure 2, it is evident that given that event A has occurred, the probability that A had occurred from partition B_4 is given by

$P(B_4/A) = \frac{P(B_4).P(\frac{A}{B_4})}{\sum_{i=1}^{4} P(Bi).P(\frac{A}{B_i}))}$, for i = 1, 2,3,4

$$\sum_{i=1}^{4} P(B_i).P(\frac{A}{B_i}) = P(B_1).P\left(\frac{A}{B_1}\right) + P(B_2).P\left(\frac{A}{B_2}\right) + P(B_3).P(\frac{A}{B_3}) + P(B_4).P(\frac{A}{B_4})$$

$P(B_4/A)$: Probability that partition B_4 occurs given that event A has occurred.

Example: The probability of receiving a spam message given that the computer programme filter has confirmed the probability to be more than 0.6. These are related probabilities that can be calculated by using Bayes' Theorem. Using the same notations, we find two mutually exclusive and collectively exhaustive events A and B as follows.

A	:	The incoming mail is a spam message.
B	:	The incoming mail is not a spam message.
		The other events defined in the context of the same experiment are:
C	:	Filter test confirms spam
D	:	Filter test did not confirm spam.
		The data given to us are:
P(A)	:	Probability of finding spam = 0.6
P(B)	:	Probability of not finding spam = 0.4
P(C|A)	:	Probability test predicts correctly when spam is actually confirmed or found.
P(D|A)	:	Probability test predicts incorrectly when spam is actually found.
P(D|B)	:	Probability test predicts correctly when actually spam is not there.
P(C|B)	:	Probability test predicts incorrectly when actually no spam is found.
		We are interested in finding:
P(C)	:	Probability that the test says spam is there.
P(D)	:	Probability that the test says no spam is there.
P(A|C)	:	Probability of finding spam, given positive test results.
P(A|D)	:	Probability of finding spam, given negative test results.
P(B|C)	:	Probability of not finding spam, given positive test results.
P(B|D)'	:	Probability of not finding spam, given negative.
		Applying Bayes' Theorem

$$P(A|C) = \frac{P(C|A)*P(A)}{P(\bar{C}|A).P(A)+P(\bar{C}|\bar{B}).P(B)}$$

$$= \frac{0.9*0.6}{0.9 .* 0.6 +0.3*0.4} = 0.818$$

$$P(B|C) = \frac{P(C|B)*P(B)}{P(\bar{C}|\bar{B}).P(B)+P(\bar{C}|A).P(A)}$$

$$= \frac{0.3*0.4}{0.3 .* 0.4 +0.9*0.6} = 0.182$$

$$P(A|D) = \frac{P(D|A)*P(A)}{P(\bar{D}|A).P(A)+P(\bar{D}|\bar{B}).P(B)}$$

$$= \frac{0.1*0.6}{0.1 .* 0.6 +0.7*0.4} = 0.176$$

$$P(B|D) = \frac{P(D|B)*P(B)}{P(\bar{D}|\bar{B}).P(B)+P(\bar{D}|A).P(A)}$$

$$= \frac{0.7*0.4}{0.1*0.6+0.7*0.4} = 0.824$$

We also know that

P(C) = P (C | A). P(A) + P(C|B). P(B) = 0.9 *0.6 + 0.3*0.4 = 0.66

P(D) = P (D | A). P(A) + P(D|B). P(B) = 0.1 *0.6 + 0.7*0.4 = 0.34

We also find that

P(C) + P (D) = 0.66 + 0.34 = 1

P(A|C) + P(B|C) = 0.818 + 0.182 = 1

P(A|D) + P(B|D) = 0.176 + 0.824 = 1

CASE STUDY

Suppose a consulting firm rents motorbikes from three rental agencies, 60 percent from agency 1, 30 percent from agency 2, and 10 percent from agency 3. Suppose 9 percent of bikes from agency 1, need a tune-up and 6 percent of the cars from agency 3 need a tune-up, what is the probability that a rental bike delivered to the firm will need a tune-up?

Let A be the event that the bike needs a tune and B1, B2, and B3 are the events that the bike comes from rental agencies 1,2, or 3. P(B1) = 0.60, P(B2) = 0.30, P (B3) = 0.10, P(A/B2) = 0.20, and P(A/B3) = 0.06

According to the Bayes' Theorem,

P(A) = (0.60) * (0.09) + (0.30) * (0.20) + (0.10) * (0.06) = 0.12

If a rental bike delivered to the consulting firm needs a tune-up, then what is the probability that it came from rental agency 2?

P(B2/A) = (0.30) * (0.20) / (0.60) * (0.09) + (0.30) * (0.20) + (0.10) * (0.06)

= 0.060 / 0.120 = 0.5

It is observed that although only 30 percent of the bike delivered to the firm come from agency 2, 50 percent of those who require a tune-up come from the agency.

IN-TEXT QUESTIONS

6. A balanced die is tossed twice. If A is the event that an even number comes up on the first toss, B is the event that an even number comes up on the second toss, and C is the event that both tosses result in the same number, are the events A, B and C
   a. Pairwise independent
   b. Independent?
7. The probability of simultaneous occurrences of two events can never exceed the sum of probabilities of these events. T
8. The conditional probability of an event given another event can never be less than the probability of the joint occurrence of their events. T

5.5 INDEPENDENCE OF EVENTS

The concept of conditional probability suggests that the probability of an event A, P(A) must be modified in context of another event B has occurred whose outcome affects the occurrence of event A. The new probability now assigned to A can be expressed as P(A|B). This is considered as conditional probability of event A occurring given that event B has already occurred. Therefore, the conditional probability of A such that B has occurred, given by P(A|B)

differs from unconditional probability P(A). This mainly indicates that the information that B has occurred results in change in the chance of A occurring.

The chances are that the occurrence of A is not affected by the fact that B has occurred, implying that P(A|B) = P(A). In other words, the occurrence or non-occurrence of one event has no consequence on the chances that the other will occur. Such events are referred to as independent events.

The two events A and B are independent if P(A|B) = P(A), while if they are dependent P(A/B) ≠ P(A). There exists a strong connection between the concept of independence and conditional probability.

The conditional probability formula for P(A|B) and P(B|A) as given below,

$$P(A|B) = P(AB)/ P(B), \quad P(B) > 0 \qquad \text{eq (1)}$$

For P(B/A),

$$P(B|A) = P(AB)/ P(A), \quad P(A) > 0 \qquad \text{eq (2)}$$

From equation 1, $P(AB) = P(A|B) * P(B)$ eq (3)

Substituting equation 3 in equation 2 we get

$P(B|A) = P(A|B) * P(B) / P(A)$

If the conditional and unconditional probability of A are same implying that

$$P(A|B) = P(A)$$

In other words, events A and B are independent,

As a result, $P(B|A) = P(A) * P(B) / P(A)$

$$P(B|A) = P(B)$$

CASE STUDY

If a coin is tossed three times and each of the outcomes is equally likely to occur. Suppose A is the event that a Head occurs on each of the two tosses, B is the event that a tail occurs on the third toss and C is the event that exactly two tails occur in the three tosses. Show that

a. Events A and B are independent

b. Events B and C are dependent.

Sample space: S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

A = {HHH, HHT} B = {HHT, HTT, THT, TTT} C = {HTT, THT, TTH}

A ∩ B = {HHT}, B ∩ C = {HTT, THT} P(A) = ¼, P (B) = ½, P (C) = ⅜, P (A ∩ B) = ⅛, and P (B ∩ C) = ¼

Since, P (A) * P (B) = ¼ * ½ = ⅛ = P (A ∩ B) = ⅛, implies that events A and B are independent.

Since P (B) * P (C) = ½ * ⅜ = 3/16 ≠ P (B ∩ C) = ⅛, implies that events A and B are not independent.

IN-TEXT QUESTIONS

9. Prove that if A and B are independent the A' and B' are also independent
10. The two mutually exclusive events must be independent. (True / False)
11. A bag contains 7 red and 4 blue balls. Two balls are drawn at random with replacement. The probability of getting the balls of different colors is:
    a. 28/121
    b. 56/121
    c. ½
    d. None of these
12. The conditional and unconditional probability of random variables are always equal.

5.6 SUMMARY

The lesson presents a measure of the likelihood of an event occurring, assuming that another event or outcome has previously occurred, which is called conditional probability. The conditional probability has wide range of application. The computation of conditional probability has been explained systematically in the lesson. The role of conditional probability to define independence and dependence of events has been described. In case of independent events conditional and unconditional probabilities are same. The notion and relevance of Bayes theorem which is primarily the application of conditional probability has been explained.

5.7 GLOSSARY

1. **Conditional Probability:** Conditional probability is defined as the likelihood of an event or an outcome occurring based on the occurrence of a previous event or outcome. It is condition upon the occurrence of some event that has happened earlier.
2. **Unconditional Probability:** It is a chance that one outcome occurs out of many outcomes. It refers to the likelihood that one outcome occurs irrespective of other outcomes.

3. **Bayes' Theorem:** It states that based on the occurrence of another event; the conditional probability indicates the likelihood of the second event given the first event multiplied by the probability of the first event.

4. **Independence of Events:** The occurrence or non-occurrence of one event has no consequence on the chances that the other will occur. Such events are referred to as independent events. In other words, the two events are said to be independent only if the conditional probability is equal to the unconditional probability.

5.8 ANSWERS TO THE QUESTIONS

1. 4/13
   Hint: P(S), n(S) = 52

   A: A card drawn from 52 cards

   H: Heart appearing

   Q: Queen appearing

   P(HUQ) = P(H) +P(Q) - P(HQ)

   P(H) = n(H)/n(S), = 13/52

   P(Q) = n(Q) / n(S) = 4/52

   P(HQ)= n(HQ)/n(S) = 1/52

   P(HUQ) = 13/52 + 4/52 - 1/52 = 4/13

2. False

3. 5/8
   Hint: Accounting major: A: n(A) = 160

   n(S) = 500

   M: n(M) = 300, n(MA) = 100, P(MA) = n(MA)/n(S) = 100/500

   F: n(F) = 200

   P(A) = n(A)/n(S) = 160/500

   P (M/A) = P(MA) or P(AM)/ P(A) = 100/160 = ⅝

4. Conditional Probability

5. 7/1,912

   Hint: Let event A be the event when the first randomly picked-up TV set out of the two is defective, A: 1st TV defective. The number of elements in the sample space is n(S) is 240.

The probability of event A, P(A) = n(A)/n(S) = 15/240.

Let event B be the event when the second randomly picked-up TV set out of the two is defective, B: 2nd TV defective, P(B/A) = 14/239

P(AB) = P(A)*P(B|A) =15/240*14/239 = 7/1,912

6. a) the events are pairwise independent
   b) the events are not independent
7. True
8. True
9. Hint: Event A can be expressed as (A ∩ B) ∪ (A ∩ B')

   Also note that (A ∩ B) and (A ∩ B') are mutually exclusive. It is given that A and B are independent.

10. False

11. (b)

12. False

5.9 SELF-ASSESSMENT QUESTIONS

1. Two six-faced dice are rolled together, or dice is rolled twice. The total number of possible outcomes are 36.
2. (i) Prove that the probability of null event is zero, P (∅) = 0.
   (ii) Prove that for any two events A and B
   P(AUB) = P(A) +P(B) - P(AB)

5.10 REFERENCES

- Devore, J. L. (2015). Probability and Statistics for Engineering and the Sciences. Cengage Learning.
- Freund, J. E., Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics: With Applications*. Pearson Education India.
- McClave, J. T., Benson, P. G., & Sincich, T. (2008). *Statistics for business and economics*. Pearson Education.

5.11 SUGGESTED READINGS

- Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.
- Larsen, R. J., & Marx, M. L. (2005). *An introduction to mathematical statistics*. Prentice Hall.

LESSON 6

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

STRUCTURE

6.1 LEARNING OBJECTIVES

After reading this lesson, students will be able :

1. To understand the concept of random variables, and their significance in statistical analysis.
2. The students will be able to distinguish between the two fundamental types of random variables, namely the discrete random variable and continuous random variable.
3. To familiarize the students with some commonly used discrete and continuous distributions of random variables.
4. To understand the concept of probability distribution function or probability mass function and
5. To comprehend the derivation of the probability distribution function

6.2 INTRODUCTION

We have seen in earlier units how the concept of probability enables us to compute the extent of uncertainty associated with random experiments. An experiment may yield both qualitative and quantitative outcomes. Statistical analysis focuses on the numerical aspect of the data or experiment. Thus, the term random variable is introduced to represent any event or outcome

that can take different values. Such a variable takes the values that are the plausible outcomes of any event or experiment.

Since the values are random and associated with random experiment such variables are termed as random variables. The concept of random variable allows us to pass from the experimental outcomes themselves to a numerical function of the outcomes.

There are two types of random variables discrete random variables and continuous random variables. This chapter will define the concept of random variables along with the two fundamental types of random variables. The chapter will help us to understand the derivation of probability distribution function both for discrete and continuous random variables.

6.3 RANDOM VARIABLE

Each outcome of an experiment can be associated with a number by specifying a rule of association example: total weight of baggage for a sample of 25 airline passengers.

Such a rule of association is called random variable

A random variable is a variable because the observed value depends on which of the possible experimental outcomes results.

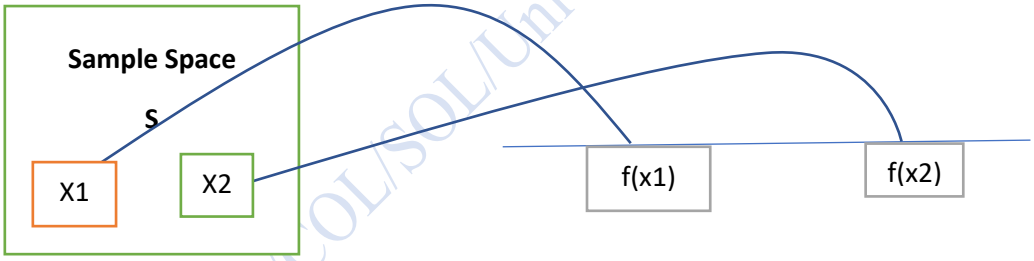


Fig 1: Rule that connects outcome of sample space to the value

For a given sample space S of some experiment, a random variable (rv) is any rule that associates a number with each outcome in S. In other words, a random variable is a function whose domain is the sample space and whose range is the set of real numbers as shown in figure 1. Various outcomes of random experiments denoted by X_1, X_2 belonging to the sample space of a certain experiment are the random variables that act as a dependent variable for the rule or a function that defines probability associated with each plausible outcome x_1.

A random variable is denoted by a capital letter such as X, Y, and the values picked up are referred as x,y,z …….

In a toss of two coins, S = {HH, HT, TH, TT}

If Event A is defined as the number of heads appearing in the toss of two coins.

A: {no. of heads in the toss}, then a variable X can be assigned as a random variable to A, where the values that X picks up are all random based on the outcome of the experiment i.e. Toss of two coins.

Similarly, if an event B is defined as B as the number of tails in the toss of two coins,

B: {no. of tails in the toss}, then a variable Y can be assigned as a random variable to B, where the values that Y picks up are all random based on the outcome of the experiment i.e. Toss of two coins.

6.3.1 Types of random variables

A random variable is a variable that takes values that are nothing, but the outcomes associated with the random experiment. Here, on the basis of values or data taken by the random variable, the random variables can be distinguished on the basis of the observed data and its countability. Thus, a random variable can be distinguished as a discrete rv or a continuous rv.

a) A discrete random variable: A discrete random variable is a rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on …….. (countable finite).

b) A continuous random variable: The continuous random variable consists either of all the numbers in a single interval on the number line (infinite from - infinity to infinity) or all numbers in a disjoint union of such intervals. No possible value of the variable has a positive probability, P(X=c) = 0 for any possible value c.

On the basis of the data taken by the random variable, we define the functions that yield the corresponding value of probability for each specific value of a random variable. Such functions are called probability distribution which gives probabilities of occurrences of different possible outcomes of an experiment. Further, depending on whether the random variable takes discrete or continuous values, these functions are referred to as Probability mass function (Pmf) or Probability density function (pdf).

6.4 PROBABILITY MASS FUNCTION

For a discrete random variable X that can take at most a countably infinite number of values $x_1, x_2, \ldots\ldots\ldots\ldots$, we associate a probability

$p_i = P[X = x] = P[\text{ all } s \in S, X(s) = x]$

that must satisfy the following conditions,

1. $p(x) \geq 0$ for all x which implies that the value of probability distribution is positive at all the values taken by the random variable x.
2. $\sum_x p(x) = 1$ which implies that all the values taken by the random variable complete the sample space, therefore the sum of all probabilities of every value of the random variable is 1.

Example 1:

Let us consider an example of a lab in the Department of Economics, where six computers are reserved for an Economics major. Let the random variable X denote the number of these computers that are in use at a particular time in a day. Suppose the probability distribution of X corresponding to each value of X is given below.

X	0	1	2	3	4	5	6
P (X =x)	0.05	0.10	0.15	0.25	0.20	0.15	0.10

It is easy to verify that the above values satisfy both the properties of a Probability Mass Function as each p(x) is positive and all of them sum to unity.

By using the above-given probabilities, various probabilities could be computed such as

The probability that at most 2 computers are in use is given by P ($X \leq 2$), the probability that the random variable takes at most the value 2.

P ($X \leq 2$) = P (X =0 or 1 or 2) = p (0) + p (1) + p(2) = 0.05 + 0.10 +0.15 = 0.30

In case of the event that at least 3 computers are in use is given by P ($X \geq 3$). Since the event that at least 3 computers are in use is complementary to at most 2 computers are in use, therefore, the probability can be computed as follows.

P ($X \geq 3$) = 1 - P ($X \leq 2$)

= 1 - 0.30

= 0.70

Another way to compute the probability of the event that at least 3 computers are in use is by adding values of probability when X takes the values 3, 4, 5 and 6.

The probability that between 2 and 5 computers are in use is given by

P ($2 \leq X \leq 5$) = P (X = 2, 3, 4 or 5) = 0.15 + 0.25 + 0.20 + 0.15 = 0.75

The probability that the number of computers in use is strictly between 2 and 5 is

P ($2 < X < 5$) = P (X = 3 or 4) = 0.25 + 0.20 = 0.45

In the above example of the number of computers in use in the computer lab of the Department of Economics, let us verify if the probability distribution function satisfies the properties.

Firstly, the value of each probability distribution function is positive. Therefore, the first property is satisfied. The sum of all probabilities is 1. The second property is also satisfied. Thus, the given distribution function can serve as a probability distribution function.

Let us create a probability distribution function for the toss of two fair coins. In this random experiment, consider the event X which is the number of heads that appear.

S: Sample space	X: Random variable (r.v.) (no. of heads)	P(X=x)	p(x)
TT	0	P(X=0) = ½ * ½	¼
HT	1	P(X=1) = ½ * ½	¼
TH	1	P(X=1) = ½ * ½	¼
HH	2	P(X=2) = ½ * ½	¼

Now, one can create the probability distribution function defined for the specific values x was taken by the random variable X that represents the event, the number of heads that appear in the toss of two fair coins. Thus, X can take values 0,1 and 2 because in two tosses of fair coins we can have no heads, only one head or both outcomes as head.

The probability distribution function for the discrete random variable can be represented by the following function p(x) which defines below.

$$p(x) = \begin{cases} ¼ \text{ if } x = 0 \\ ½ \text{ if } x = 1 \\ ¼, \text{ if } x = 2 \end{cases}$$

1. The value of the probability distribution function is always positive. This implies $p(x) \geq 0$
2. The sum of all values of probability distribution function at given values of x is one.

   $\sum_{x=0}^{2} p(x) = 1$

   The above function satisfies both conditions; therefore, it serves as pdf or pmf, which can be depicted the form of a graph as below.

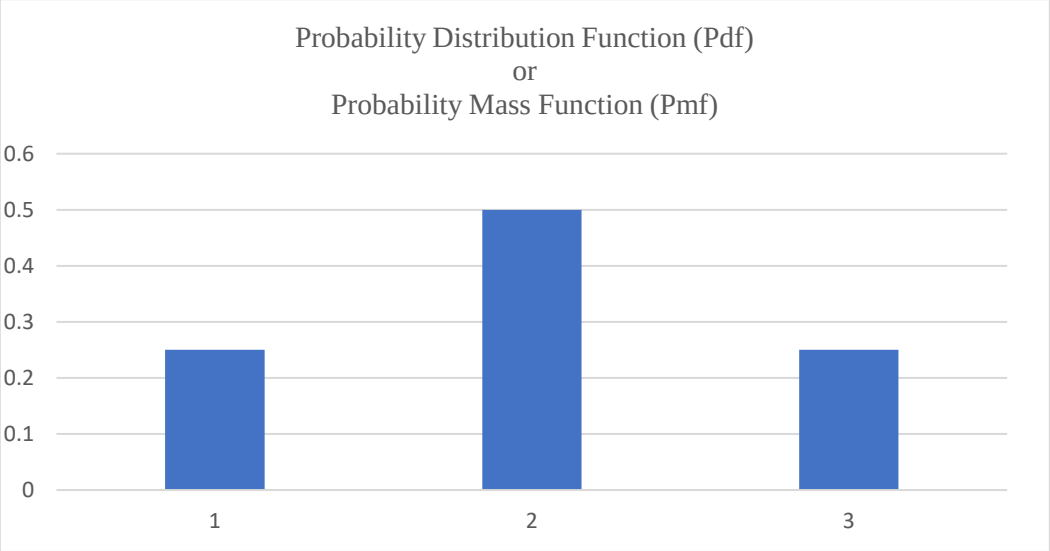


Figure 2: Graph of Probability Mass Function (Pmf)

The function depicted in figure 2 is the graph plotted for the probability distribution function or probability mass function defined in the example. The graph of the probability mass function (Pmf) is a discrete function, and it can be depicted as a histogram as reflected in the figure 2.

CASE STUDY

Find the probability distribution or the pmf of the sum of numbers obtained on throwing a pair of dice.

Solution: The Sample Space of the experiment is:

(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 sum of the numbers are as follows:

2 3 4 5 6 7

3 4 5 6 7 8

4 5 6 7 8 9

5 6 7 8 9 10

6 7 8 9 10 11

7 8 9 10 11 12

The probability distribution or pmf is therefore given by:

x	p(x) = P(X=x)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

IN-TEXT QUESTIONS

1. A random variable that can assume any possible value between two points is called a _____ (discrete/continuous) random variable.
2. If C is a constant in a continuous probability distribution, then p(X=c) is always equal to Zero. This statement is True/False.
3. A listing of all the outcomes of an experiment and the probability associated with each outcome is called:
   a) Probability density function
   b) Cumulative distribution function
   c) Probability distribution
   d) Probability tabulation

CASE STUDY

Check if the following function given by

$$p(x) = x + 2 / \quad 25, \text{ for } x = 1,2,3,4,5$$

can serve as the probability distribution of a discrete random variable.

Solution: For all the values of x, p(x) is computed as follows,

For, x=1, f (1) = 3/25

x=2, f (2) = 4/25

x= 3, f (3) = 5/25

x=4, f (4) = 6/25

x=5, f (5) = 7/25

For every x, p(x) $>$ 0, and also the sum of all p(x) is 1. Thus, all two properties of the probability distribution function have been satisfied.

IN-TEXT QUESTIONS

4. Find the probability distribution of the total number of heads obtained in four tosses of a balanced coin.
5. The probability distribution of a random variable is defined as

x	-1	-2	0	1	2
p(x)	c	2c	3c	4c	6c

Then, c is equal to

a) 0

b) ¼

c) 1

d) 1/16

6. The suitable graph for the probability distribution of a discrete random variable is

   a) Probability Histogram

   b) Stepwise Function

   c) Both (a) and (b)

6.5 PROBABILITY DENSITY FUNCTION

For continuous random variable X that can take all possible values between certain limits, we consider the small interval (x, x+Δx) of length Δx. If f(x) is any continuous function of x such that f(x) Δx represents the probability that X lies in infinite small interval (x, x+Δx)

that is, P [x$\leq X \leq x + \Delta x] = f(x)\ \Delta x$

The function f(x) so defined is known as probability density function and is denoted by

$$f(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X \leq x+\Delta x]}{\Delta x}$$

Let X be a continuous rv the probability distribution or probability density function pdf of X is a function p(x) such that for any two numbers a and b with a ≤ b,

P (a ≤ X ≤ b) = $\int_a^b f(x)dx$

Which is the probability that X takes value in the interval [a, b] is the area above this interval and under the graph of the density function. The value of function f (c) at a point c is irrelevant in this case and does not provide the value P (X = c) as in the case of a discrete case. In the case of a continuous random variable, probabilities are always associated with intervals and therefore, probability value when the random variable takes a value c then, P (X = c) = 0 for any real constant c.

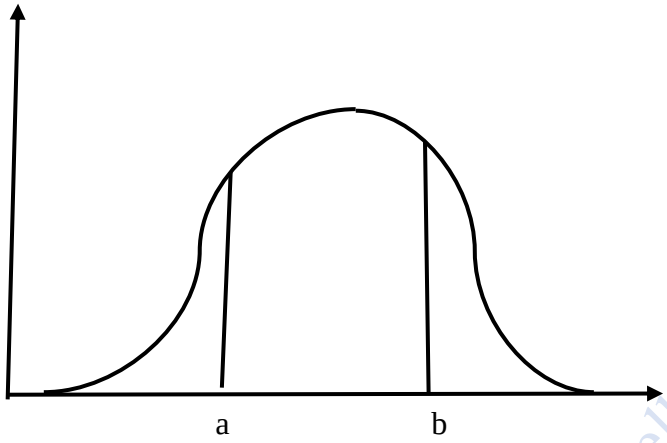


Figure 3: The density curve between a and b given by P ($a \leq X \leq b$)

The probability between intervals a and b can be computed by integrating the density function as depicted in figure 3. This implies that probability between a and b P ($a \leq X \leq b$) can be obtained by integrating the density function $\int_a^b f(x)dx.$

6.5.1 Properties of Probability Density Function

Every Probability density function qualifies certain properties. The first and foremost property is the two conditions that should be satisfied for any function of a continuous random variable to be addressed as the probability density function.

1. A function can serve as a probability density of a continuous random variable X if its values, p(x), satisfy the following two conditions.

(a) $p(x) \geq 0$ for $-\infty < x < \infty$, for all x

(b) $\int_{-\infty}^{\infty} p(x)dx = 1$

2. If X is a continuous random variable and a and b are real constants with $a \leq b$ then,

$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$

CASE STUDY

For example, If X random variable has the probability density given by

$$f(x) = \begin{cases} k.e^{-3x} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find k and P ($0.5 \leq X \leq 1$)

The given function satisfies the two necessary conditions for the probability density function.

(a) $p(x) \geq 0$ for $-\infty < x < \infty$

(b) $\int_{-\infty}^{\infty} f(x)dx = 1 = \int_{-\infty}^{\infty} k.e^{-3x}\, dx = k/3 = 1$

For k = 3, P $(0.5 \leq X \leq 1) = \int_{0.5}^{1} 3.e^{-3x}\, dx$ = **0.173**

IN-TEXT QUESTIONS

The pdf of a continuous random variable X is given by:

$$f(x) = \begin{cases} 0.075x + 0.2, & 3 \leq x \leq 5 \\ 0, otherwise \end{cases}$$

7. The total area under the density curve of X is
   (a) 1
   (b) 0
   (c) ½
   (d) ¼
8. Find P(X≤ 4).
9. P(X≤ 4) will be the same as P(X<4). This statement is True/False.

6.6 SUMMARY

The lesson describes the concept of a random variable which takes values that are outcomes of random experiments. The two types of random variables namely the discrete and continuous random variables have been described in the lesson. The discrete random variables are the variables whose values are either finite or infinite in character. While the continuous random variable consists either of the intervals on the number line or a disjoint union of intervals. The probability distribution functions, also known as probability mass functions, depicting probabilities corresponding to each outcome or value of the discrete random variable have been derived. The corresponding graph of probability mass function has been plotted. Similarly, the probability density function, depicting the probability associated with each of the continuous random variable has been derived. Finally, the properties of the probability distribution and probability density functions have been presented in the lesson.

6.7 GLOSSARY

1. **Random Variable:** A random variable is a rule that assigns a numerical value to each outcome in a sample space. It is a variable because the observed value depends on which of the possible experimental outcomes results.
2. **Discrete random variable:** A discrete random variable is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on
3. **Continuous random variable:** The continuous random variable consists either of all the numbers in a single interval on the number line (infinite from - infinity to infinity) or all numbers in a disjoint union of such intervals.
4. **Probability distribution function:** The probabilities assigned to various outcomes in S in turn determine probabilities associated with the values of any particular random variable rv. X. It is a function that provides relative likelihood of occurrence of all possible outcomes of an experiment.
5. **Probability density function:** The function defines the probability function representing a continuous random variable belonging to some specified range of values. The function provides the likelihood of values of continuous random variables.

6.8 ANSWERS TO IN-TEXT QUESTIONS

1. Discrete
2. True
3. (c) Probability Distribution
4. $p(x) = 4Cx /16$, for $x = 0, 1,2, 3, 4$
5. (d) 1/16
6. (a) Probability Histogram
7. (a) 1
8. 0.4625
9. True

6.9 SELF-ASSESSMENT QUESTIONS

1. What is the difference between discrete and continuous random variables?
2. An event management company has overbooked the tickets for an upcoming music concert. The available seating capacity is 40 while the company has sold 45 tickets.

Suppose X denotes the number of ticketed people who actually show up for the concert. The probability mass function of X is given by:

What is the probability that the event management company will be able to accommodate all ticketed people who show up?

3. Prove that the following function defined by

$$f(x) = \begin{cases} \frac{1}{5}, & 2 < x < 7 \\ 0, otherwise \end{cases}$$

can serve as a valid pdf for a random variable X.

4. The pdf of a continuous random variable Y is given by

$$f(y) = \begin{cases} \frac{k}{\sqrt{y}}, & 0 < y < 4 \\ 0, otherwise \end{cases}$$

Find the value of k. Also, find P (Y≥ 1)

6.10 REFERENCES

- Devore, J. L. (2015). Probability and Statistics for Engineering and the Sciences. Cengage Learning.
- Freund, J. E., Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics: With Applications*. Pearson Education India.
- McClave, J. T., Benson, P. G., & Sincich, T. (2008). *Statistics for business and economics*. Pearson Education.

6.11 SUGGESTED READINGS

- Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.
- Larsen, R. J., & Marx, M. L. (2005). *An introduction to mathematical statistics*. Prentice Hall.

LESSON 7

CUMULATIVE DISTRIBUTION FUNCTION, DENSITY FUNCTION, EXPECTED VALUE, AND VARIANCE

STRUCTURE

7.1 LEARNING OBJECTIVES

After reading this lesson, students will be able :

1. To understand the need to evaluate the descriptive statistics of the probability distribution function.
2. To learn the formula and method to derive the expected value of the probability distribution function.
3. To compute and apply the expected value in different random experiments by deriving the probability distribution functions.
4. To learn the formula and method to derive the variance of the probability distribution function.

5. To compute and apply the variance in different random experiments by deriving the probability distribution functions and
6. To provide exposure to various useful properties of expected value and variance to the students and their applications.

7.2 INTRODUCTION

In the earlier lesson 6, we discussed random variables, types of random variables, and probability distributions. Once we have the probability distribution function of a random variable, it is essential to evaluate and assess its mean and other descriptive statistics such as expected value and variance. The population mean for a random variable is a measure of centre for the distribution of a random variable. The expected value is essentially a formula that enables us to evaluate the mean value as more and more values of the random variables are collected either by trials or random experiments or any kind of experiment involving probability, the sample mean becomes closer and closer to the expected value. It is obtained by summing the product of the value of the random variable and its associated probabilities over all the values of the random variable.

Another important measure of dispersion is variance. It determines the measure of spread for the distribution of a random variable. It reflects the degree of variability of values of a random variable from the expected value. The variance for a given probability distribution function is obtained by summing the product of the square of the difference between the value of the random variable and the expected value, and the associated probability of the value of the random variable taken across all the values of the random variable.

7.3 CUMULATIVE DISTRIBUTION FUNCTIONS (CDF)

The cumulative distribution function (cdf) of a discrete r.v. variable X with pmf f(x) is defined for every number x by

$F(x) = P(X \leq x) = \sum_{y=-\infty}^{x} p(y)$

For any number x, F(x) is the probability that the observed value of X will be at most x.

Let us compute the Cumulative Distribution Function for the toss of two coins.

S: Sample space	X: Random variable (r.v.) (no. of heads)	f(x): pmf	F(x): CDF (cdf)
TT	0	¼	¼
HT	1	½	¾ = ¼ + ½
TH	1		
HH	2	¼	1 = ¼ + ¾

The first column represents the elements in the sample space. The next column presents the values of a random variable which is defined as the number of heads in the toss of two coins, denoted by X random variable. The values that random variable X takes are denoted by x and these values are 0,1 or 2. The third column represents the corresponding probabilities associated with each value of random variable X. The fourth column is obtained by adding or cumulating all the earlier probabilities till each random variable.

The Cumulative Distribution Function for the toss of two coins can be presented in the following way.

$$F(x) = \begin{cases} ¼ \text{ if } x < 1 \\ ¾, \text{ if } 1 \leq x < 2 \\ 1, \text{ if } x \geq 2 \end{cases}$$

1. $f(x) \geq 0$

The probability distribution function is positive for each value of random variable X.

2. $F(x) = \sum_D f(x) = 1$

This property signifies that the value of the cumulative distribution function for the last value of the random variable for which it is defined is equal to 1. This is due to the fact that while cumulating all the probabilities till the last value of random variable X for which the pdf is defined, we tend to exhaust all plausible values or outcomes of sample space and therefore the value of CDF function is 1 as the probability of whole sample space is 1.

The graph of the above cumulative distribution function can be presented in figure 1 below.

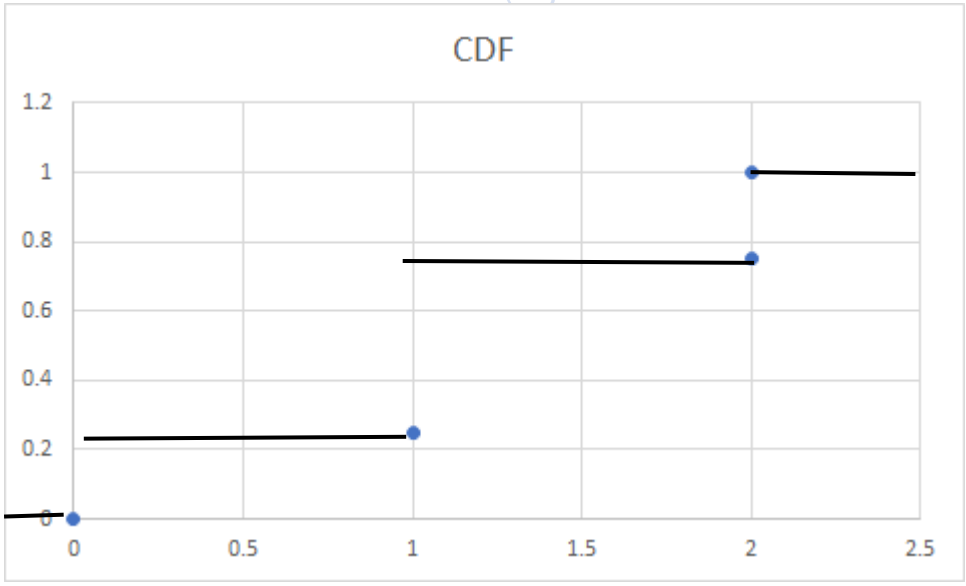


Figure 1: Graph of the cumulative distribution function cdf.

The graph of the cumulative distribution function is a step function. For all the values of random variables less than zero, the value of the cumulative distribution function CDF is zero since the probability or the value of the probability distribution function is zero. The cumulative distribution function is defined for intervals where the extreme points of each interval are not included in the interval concerned but included in the next slab or interval. At each interval the values of the probability function keep getting added or cumulated, thus the cumulative distribution function appears like a ladder or steps moving upwards. At the last value of the random variable, all the values of probability distributions get added and the value of the cumulative distribution function attains the value one, where it reaches the maximum value and the value of cdf remains at one for all the infinite values of random variable for which it is defined.

Consider whether the next person buying a computer at a university bookstore buys a laptop or a desktop model.

X = 1, if the customer purchases a laptop computer

0, if the customer purchases a desktop computer

If 20% of all purchasers during a week select a laptop computer

For X = 0, p (0) = P(X=0) = P (next customer purchases a desktop model) = 0.8

For X = 1, p (1) = P (X =1) = P (next customer purchases a laptop model) = 0.2

p(x) = P (X = x) = 0 for $x \neq 0, 1$

In the above activity, the example mentioned related to the next person buying a computer at a university bookstore buying a laptop or a desktop model. The cumulative distribution function can be derived in the following manner.

X	f(x)	Prob	F(x)
0	0.8	P ($X \leq 0$)	0.8
1	0.2	P ($X \leq 1$)	1

$$F(x) = \begin{cases} 0, \text{ if } x < 0 \\ 0.8, \text{ if } 0 \leq x < 1 \\ 1, \text{ if } x \geq 1 \end{cases}$$

The value of CDF remains zero for all negative values of x for which the probability distribution function is zero, while the CDF takes a value of 0.8 between 0 and 1 and finally 1 for all values 1 and greater than 1.

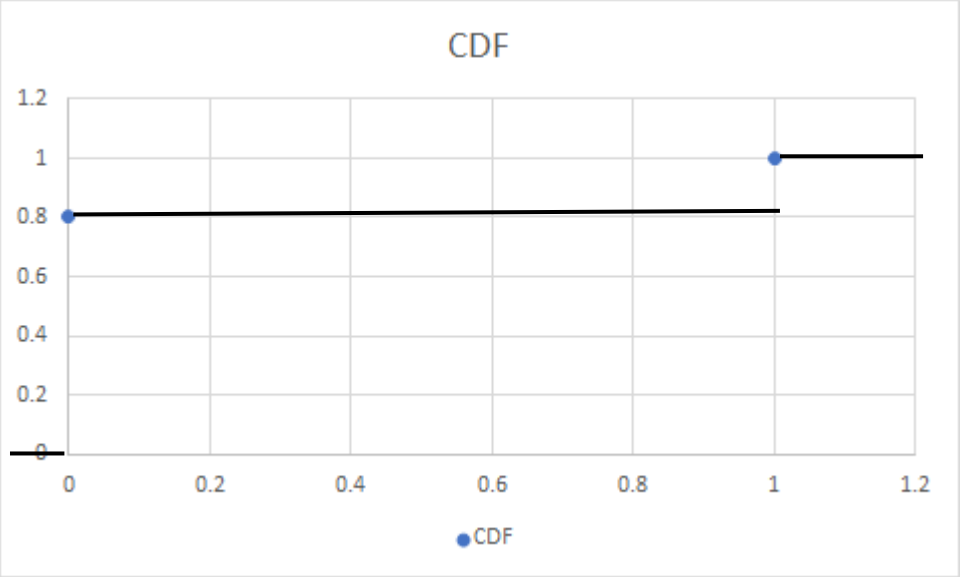


Figure 1: Graph of the cumulative distribution function cdf.

The probability distribution function is defined for two values of random variables 0 and 1. Therefore, while computing the cumulative distribution function the value of CDF remains zero for all negative values of x for which the probability distribution function is zero, while the CDF takes a value of 0.8 till the points are strictly less than 1. Finally, the CDF value becomes 1 when the probability distribution function value at 1 is added. The CDF continues to attain the value 1 for all values of random variable X.

7.3.1 Properties of Cumulative Distribution Function

The value of F(x) of the distribution function of a discrete random variable X satisfies the conditions:

1. $F(-\infty) = 0$ and $F(\infty) = 1$

   The value of Cumulative Distribution Function CDF is 0 for all negative values of random variable X for which the probability distribution function remains zero.

2. If a <b, then $F(a) \leq F(b)$ for any real numbers a and b.

The CDF is a non-decreasing function in the random variable X. This implies for each greater value of X; the value of the cumulative distribution function is also greater.

If the probability distribution of a discrete random variable is given, the corresponding distribution function can be derived.

The distribution function of the total number of heads obtained in four tosses of a balanced coin can be obtained as below. For x = 0,1,2,3,4.

‘

f (0) = 1/16, f(1) = 4/16 , f(2) = 6/16, f(3) = 4/16 , and f(4) = 1/16

It follows that

F (0) = f (0) = 1/16

F (1) = f (0) + f(1) = 5/16

F (2) = f (0) + f(1) + f(2) = 11/16

F (3) = f(0) + f(1) + f(2) + f(3) = 15/16

F (4) = f (0) + f(1) + f(2) + f(3) + f (4) = 1

The properties of Distribution Function are satisfied by the above CDF.

1. $F(-\infty) = 0$ and $F(\infty) = 1$
2. If $a < b$, then $F(a) \leq F(b)$ for any real numbers a and b.

Therefore, the distribution function is given by

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1/16 & \text{for } 0 \leq x < 1 \\ 5/16 & \text{for } 1 \leq x < 2 \\ 11/16 & \text{for } 2 \leq x < 3 \\ 15/16 & \text{for } 3 \leq x < 4 \\ 1 & \text{for } x \geq 4 \end{cases}$$

The distribution function is defined not only for the values taken on by the given random variable but for all real numbers.

F (1.7) = 5/16 and F (100) =1, although the probabilities of getting “at most 1.7 heads” or “at most heads” in four tosses of a balanced coin may not be of any real significance.

If the range of a random variable X consists of the values $x_1 < x_2 < x_3 < x_4 \ldots\ldots < x_n$, then

$f(x_1) = F(x_i) - F(x_{i-1})$, for $i = 2,3, \ldots, n$

The above equation reveals that the probability distribution function can be computed from the given cumulative distribution function by taking the difference of CDF for each consecutive random variable.

IN-TEXT QUESTIONS

1. A bulb manufacturing company testing the number of defective bulbs. Let X denotes the number of defective bulbs

$$f(x) = \begin{cases} 0.25 & x=0 \\ 0.10 & x=1 \\ 0.30 & x=2 \\ 0.35 & x=3 \end{cases}$$

(A) Calculate the Probability that almost 1 bulb is defective.

(B) find CDF.

2. Let X be the number of a group that attended the festival of the college.

X	0	1	2	3	4	5
P(X)	0.20	0.10	0.05	0.15	0.30	0.20

(a) Find the probability that at most two groups attended the festival.

(b) Find the probability that at least four groups attended the festival.

3. Cumulative distribution function of a random variable Y is the probability that Y takes the value _____

(a) Equal to Y

(b) Greater than Y

(c) Less than or equal to Y

(d) Zero

7.4 CUMULATIVE DENSITY FUNCTION

A cumulative density function is the cumulative distribution function for a continuous random variable. If X is a continuous random variable and the value of its probability density at x is f(x), then the function given by

F (x) = P $(X \leq x) = \int_{-\infty}^{x} f(y)dy$ for $-\infty < x < \infty$, for all x.

Is called the distribution function or the cumulative distribution of X

For each x, F(x) is defined as the area under the density curve to the left of x as shown in figure 2. The density function increases as x increase

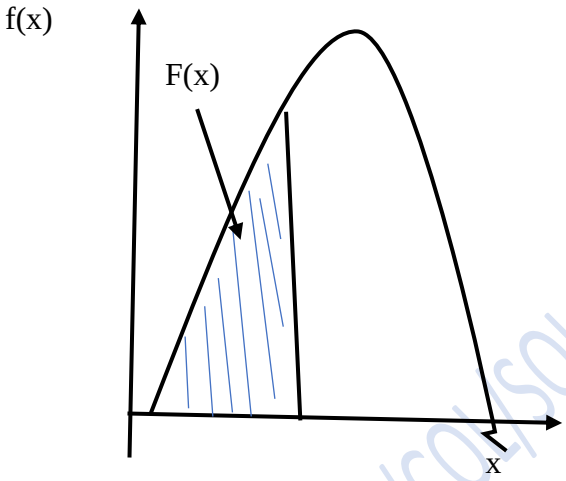


Figure 2 (a): Probability density function

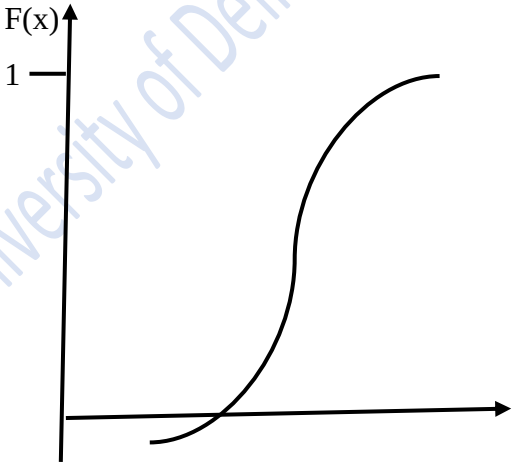


Figure 2(b) : Cumulative density function

Figure 2 (a) depicts the probability density function while figure 2(b) depicts the cumulative density function.

7.4.1 Properties of Cumulative Density Function

There are certain properties of cumulative density function:

1. This further indicates, F $(-\infty) = 0$, F $(\infty) = 1$
2. For any two numbers a and b with $a \leq b$,
   P $(a \leq X \leq b)$ = F (b) - F (a -)

   a - represents the largest possible X value that is strictly less than a.

3. If a and b are integers, then
   $P(a \leq X \leq b) = P(X = a \text{ or } a+1 \text{ or } \ldots\ldots \text{or } b)$

   $= F(b) - F(a - 1)$

   Taking a = b yields $P(X = a) = F(a) - F(a - 1)$ in this case

4. For any real constants a and b with $a \leq b$, and

   $f(x) = dF(x) / dx$

   where the derivative exists.

The third property indicates that one can obtain the probability function from the given distribution or cumulative density function by simply differentiating the density function.

CASE STUDY

In the case study depicted in lesson 7, section 7.5.1, If X random variable has the probability density given by

$$f(x) = \begin{cases} k.e^{-3x} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

The distribution function derived in lesson 7 section 7.5.1 is given by

$$F(x) = \int_{-\infty}^{x} 3e^{-3t}\, dt$$

$$= 1 - e^{-3t}$$

Since $F(x) = 0$ for $x \leq 0$,

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - e^{-3x} & \text{for } x > 0 \end{cases}$$

Now, to determine the probability $P(0.5 \leq X \leq 1)$, Cumulative density function F (x) can be used

$$P(0.5 \leq X \leq 1) = F(1) - F(0.5)$$

$$= (1 - e^{-3}) - (1 - e^{-1.5})$$

$$= 0.173$$

IN -TEXT QUESTIONS

4. Given the following probability density function pdf

$$f(x) = \begin{cases} px^2 & 0 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

(A) Find the value of p such that pdf is defined.
(B) P(X>2)

7.5 EXPECTED VALUE

Let x be a discrete rv with a set of possible values D and pmf f(x). The expected value or mean value of X, denoted by E (X) or μ_x is

$E(X) = \mu_x = \sum_D x.f(x)$

The expected value of any random variable X is obtained by summing the product of each random variable x and the corresponding probability distribution function, where D is the domain to which all values of x random variable belong.

Let us consider the computation of the expected value of the toss of two coins as shown below. In the first column, the random variable X that denotes the number of heads is defined. The random variable X takes values as small x, which could be 0, 1, or 2. The corresponding probability for each value of X is depicted as a probability distribution function or probability mass function in the second column.

For computing the expected value, the third column which is the product of random variable and corresponding probability is presented. The final column computes the expected value

X: Random variable (r.v.) (no. of heads)	**f(x) : pmf**	**x. f(x)**	**E (X)**
0	¼	0* 1/4	0
1	½	1 * 1/2	½
2	¼	2 * 1/4	½
$E(X) = \sum_D x.f(x)$			$E(x) = \sum_D x.f(x) = 1$

If the r.v. X has a set of possible values D and pmf f(x), then the expected value of any function h(x), denoted by E [h(x)] or $\mu_{h(x)}$ is computed by

E [h(x)] or $\mu_{h(x)} = \sum_D h(x).f(x)$

The expected value of h(x) is given by the sum of the product of h(x) and the corresponding probability distribution function f(x) as shown in the above expression, where D is the domain.

Example 1: Suppose there is a collection of 12 audio sets that include 2 with white cords. If three of the sets are chosen at random for shipment to a hotel, how many sets with white cords can the shipper expect to send to the hotel?

First, we need to construct the probability distribution of X, the number of sets with white cords shipped to the hotel, given by

$$f(x) = \frac{{}^{2}C_{x} * {}^{10}C_{3-x}}{{}^{10}C_{3}} \quad \text{for} \quad x = 0,1,2$$

X	0	1	2
f(x)	6/11	9/22	1/22

E(X) = 0* 6/11 + 1 * 9/22 + 2 * 1/22 = ½

If X is the number of points rolled with a balanced die, find the expected value of

$$g(X) = 2X^2 + 1$$

The probability of each outcome is ⅙., we get

$$E[g(X)] = \sum (2X^2 + 1) * \frac{1}{6}$$

$$= 94/3$$

IN-TEXT QUESTION

5. Suppose the probability distribution function is given as below

x	-3	-2	-1	0	1	2	3
f(x)	0.20	0.05	0.15	0.05	p	0.15	0.25

(A) Find the CDF F(x)

(B) Find the expected value E(X)

6. If 'Y' is a continuous random variable then the expected value of Y is

(a) P(y)

(B) $\int_{-\infty}^{\infty} y.f(y)dy$

(c) $\sum_D y.f(y)$

(D) None of these

7.5.1 Properties of Expected Value

One needs to find the average number of heads obtained when tossing a coin several times, referred to as the expected value.

Some of the properties of expected value E(X) are

1. The expected value of a constant is a constant

   E(b) = b

   If b =2, then E (2) = 2

2. Expectation of the sum of the two random variables X and Y is equal to the sum of the expectations of those random variables.

   E (X+Y) = E (X) + E (Y)

3. The expected value of ratio of two random variables is not equal to the ratio of the expected values of those random variables.

   E (X/ Y) ≠ E(X) / E(Y)

4. The expected value of the product of two random variables that are dependent is not equal to the product of expectations of those random variables.

   E (XY) ≠ E(X) * E(Y)

   However, if X and Y are independent random variables then

   E (XY) = E(X) * E(Y)

Hint: Since the joint probability mass function for all values of X and Y is equal to the product of individual probability distribution function pdf of two random variables for all values of variables.

$f(X, Y) = f_x(x) * f_y(y)$

5. The expected value of the square of X is not equal to the square of the expected value of X

$$E(X^2) \neq [E(X)]^2$$

6. If a is a constant, then

$$E(aX) = a\,E(X)$$

7. If a and b are constants,

$$E(aX + b) = a\,E(X) + E(b)$$

$$= a\,E(X) + b$$

$$E(4X + 7) = 4\,E(X) + 7$$

8. The expected value of multivariate Probability distribution is

$$E(X) = \sum_x \sum_y x.y\, f(xy)$$

IN-TEXT QUESTION

7. If E(XY)= E(X). E(Y) then X and Y are

   (A) Dependent

   (B) Independent

   (C) Correlated

   (D) None of these

8. E(X) = 4

   E(Y) = 1

   E(X-Y) is _____

   (A) 2
   (B) 4
   (C) 3
   (D) None of these

7.6 VARIANCE V(X)

The variance is the square of the mean of deviation between the values of a random variable from the expected value or population mean. The variance of the random variable X is denoted by $\boldsymbol{\sigma}_x^2$, Var (X). The symbol represents standard deviation under the root of variance.

$V(X) = E(X^2) - [E(X)]^2$ or $E(X^2) - [\mu_x]^2$

Proof:

$V(X) = E[X - E(X)]^2 = E[X - \mu_x]^2$

$= E[X^2 - 2\mu_x X + \mu_x^2]$

$= E(X^2) - 2\mu_x E(X) + E(\mu_x^2)$

$= E(X^2) - 2\mu_x^2 + \mu_x^2$ Since, Expected value of a constant is constant itself

$= E(X^2) - \mu_x^2$

$= E(X^2) - [E(X)]^2$

Let X have pmf f(x) and expected value E(X) as μ_x. Then the variance of X denoted by V(X) or $\boldsymbol{\sigma}_x^2 X$ is

$V(X) = \boldsymbol{\sigma}_x^2 X = \sum (X - \mu_x)^2 . f(x) = E(X - \mu_x)^2$

The standard deviation sd of X is $\boldsymbol{\sigma}_x = \sqrt{\boldsymbol{\sigma}^2_x}$

Let us consider the computation of variance of the two coins

X: Random variable (r.v.) (no. of heads)	f(x) :pmf	x.f(x)	$x^2 f(x)$
0	¼	0* 1/4	0 * ¼ = 0
1	½	1 * 1/2	1* ½ = ½
2	¼	2 * 1/4	4* ¼ = 1
$E(X) = \sum_D x.f(x)$ or E (X) = 1			$E(X^2) = \sum_D x^2.f(x) = 1.5$

$E(X^2) - [E(X)]^2 = 1.5 - 1 = 0.5$

A CASE STUDY

Calculate the variance of a random variable X that represents the number of points rolled with a balanced die.

X	0	1	2	3	4	5	6
f(x)	1/6	1/6	1/6	1/6	1/6	1/6	1/6

For this problem, we need to first compute the expected value of X, E (X)

$E(X) = 1 * ⅙ + 2 * ⅙ + 3 * ⅙ + 4 * ⅙ + 5 * ⅙ + 6 * ⅙ = 7/2$

$E(X^{2)} = 1^2 * ⅙ + 2^2 * ⅙ + 3^2 * ⅙ + 42 * ⅙ + 5^2 * ⅙ + 6^2 * ⅙ = 91/6$

$V(X) = E(X^{2)} - [E(X)]^2$

$= 91/6 - (7/2)^2$

$= 35/12$

IN-TEXT QUESTION

9. Variance of first 5 natural number is _____

(A) 4

(B) 2

(C) 3

(D) 1

7.6.1 Properties of Variance

Variance thus defined shows how the individual values are spread or distributed around its expected or mean value. Some of the useful properties of variance are mentioned below.

1. If all X values are equal to E (X) the variance is 0 value whereas if they are widely spread around the expected value, it will be relatively large.
2. The variance of a constant is zero. It implies V (a) = 0
3. If X and Y are independent random variables then,

$$V(X + Y) = V(X) + V(Y)$$
$$V(X - Y) = V(X) - V(Y)$$

4. If b is a constant then,

$$V(X + b) = V(X) + V(b)$$
$$= V(X) + 0$$
$$= V(X)$$

5. If a is a constant then,

$$V(aX) = a^2 V(X)$$

$$V(5X) = 25\, V(X)$$

6. If a and b are constants then,

$$V(aX + b) = a^2 V(X) + 0$$

$$V(5X + 9) = 25\, V(X)$$

7. If X and Y are independent random variables and a & b are constants then,

$$V(aX + bY) = a^2 V(X) + b^2 V(Y)$$

$$V(3X + 5Y) = 9\, V(X) + 25\, V(Y)$$

8. The variance can be computed as

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \sum_D x^2 . f(x)$$

$$E(X) = \sum_D x . f(x)$$

IN-TEXT QUESTION

10. $V(X) = 9$, Find $V(2X)$.

(A) 18

(B) 9

(C) 36

(D) 72

11. If the standard deviation of a set of observations is 5 and if each observation is divided by 5, then the new standard deviation is.

(A) 1

(B) 2

(C) 4

(D) 5

7.7 SUMMARY

The lesson presents the cumulative distribution function both for the discrete and continuous random variables. The properties of the cumulative distribution function have been discussed with respect to both the discrete random variable and continuous random variable. The method to derive the probability density function from the cumulative density function by differentiating the cumulative density function is explained. Further the descriptive statistics such as expected value and variance is computed for probability distribution function. There are several useful and interesting properties of expected value and variance comprehensively explained in the lesson.

7.8 GLOSSARY

1. **Cumulative Distribution Function:** The cumulative distribution function is another way of defining the distribution of the discrete random variable. The cumulative distribution function (cdf) is a discrete r.v. variable X with pmf p(x) is defined for every number x
2. **Cumulative Density Function:** The cumulative density function, also referred to as Density Function is the cumulative function of probability density distribution for continuous random variables.
3. **Expected value of Distribution Function:** Let x be a discrete rv with a set of possible values D and pmf p(x). The expected value or mean value of X, denoted by E (X) or μ_x. It is the sum of the product of each random variable and the associated probability function
4. **Variance of Distribution Function:** The variance is the square of the mean of deviation between the values of random variables from the expected value or population means. The variance of the random variable X is denoted by $\boldsymbol{\sigma}^2$, Var (X). The symbol represents standard deviation under root of variance.

7.9 ANSWERS TO IN-TEXT QUESTIONS

1. (A) 0.35

(B) $F(x) = \begin{cases} 0 & x<0 \\ 0.25 & 0 \le x<1 \\ 0.35 & 1 \le x<2 \\ 0.65 & 2 \le x<3 \\ 1 & x \ge 3 \end{cases}$

2. (A) 0.35

   (B) 0.50

3. (C) less than or equal to x

4. (A) 1/9

   (B) 19/27

5. (A) 0.15

   (B) 0.35

6. (B) $\int_{-\infty}^{\infty} y.f(y)dy$

7. (b) independent

8. (C) 3

9. (b) 2

10. (c) 36

11. (a) 1

7.10 SELF-ASSESSMENT QUESTIONS

Q. 1) If distribution function of X is given by

$$F(x) = \begin{cases} 0 & x<2 \\ 2/10 & 2 \le x<4 \\ 6/10 & 4 \le x<6 \\ 7/10 & 6 \le x<8 \\ 1 & x \ge 8 \end{cases}$$

(a) Find the probability distribution of the random variable f(x)

(b) P(X≥6). Probability of X taking the value at least 6.

Q 2) Suppose the probability distribution

X	0	1	2	3	4
P(x)	1/6	2/6	0	p	1/3

Find the value of the following

(a) F(x) CDF

(b) E(X), Expected value

(c) E(X+2)

(d) V(X), Variance

(e) V(X+2)

Q 3) The probability density function of random variable Y is given by

$$f(y) = \begin{cases} c/\sqrt{y} & 0<y<9 \\ 0 & \text{Otherwise} \end{cases}$$

Find the value of

(a) c

(b) P(Y>4)

(c) E(Y), Expected value of Y

Q.4 A coin is tossed thrice. Let X denotes the number of tails. Find its expectation and variance

Q.5 The probability density function of a random variable Y is given as below:

$$f(y) = \begin{cases} 1/16y & 0 \le y \le 9 \\ 0 & \text{Otherwise} \end{cases}$$

Find the value of c such that $P(Y \le c) = 1/2$

7.11 REFERENCES

- Devore, J. L. (2015). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

- Freund, J. E., Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics: With Applications*. Pearson Education India.
- McClave, J. T., Benson, P. G., & Sincich, T. (2008). *Statistics for business and economics*. Pearson Education.

7.12 SUGGESTED READINGS

- Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.
- Larsen, R. J., & Marx, M. L. (2005). *An introduction to mathematical statistics*. Prentice Hall.

LESSON-8

DISCRETE DISTRIBUTION

STRUCTURE

8.1 LEARNING OBJECTIVES

After reading this lesson, students will be able to learn about :

1. Different kinds of discrete distributions.
2. Uniform distributions, Bernoulli distributions, Bernoulli trials,
3. Binomial distributions, Poisson distributions with some important numerous and
4. Waiting distributions i.e., geometric distributions, negative binomial and hypergeometric distributions have been discussed.

8.2 INTRODUCTION

In this unit we will study the different types of discrete distributions. We have studied the discrete random variable in unit 3. The discrete random variables form the discrete probability

distributions. Possible values of discrete random variables along with the probabilities forms the discrete probability distribution.

1. Uniform Distribution
2. Bernoulli Distribution
3. Binomial Distribution
4. Poisson Distribution
5. Limiting case of Binomial distribution
6. Hyper-geometric distribution

8.3 PROBABILITY DISTRIBUTION FOR DISCRETE RANDOM VARIABLE

8.3.1 Uniform Distribution

Under this distribution, random variable can take 'n' different values with equal probability. For example, rolling a fair dice we get 1, 2, 3, 4, 5, 6 as outcome each with the probability of $\frac{1}{6}$. So the probability of occurrence of each event is equally likely.

$$f(x=x)=f(x)=\frac{1}{n} \qquad x=1,2,3....n$$

Mean=E(x) and Variance=V(x)

$$E(x)=\frac{k+1}{2} \qquad V(x)=\frac{k^2-(1)^2}{12}$$

Note that x has started from 1, 2, 3, ... n not from 0.

If I consider trials from 0 then

$$f(x)=\frac{1}{n+1} \qquad x=0,1,2,....n$$

$$0 \qquad \text{otherwise}$$

Since there will be n + 1 terms

Mean=E(x) and Variance=V(x)

$$E(x)=\frac{n}{2}, \qquad V(x)=\frac{n^2}{12}$$

8.3.2 BERNOULLI DISTRIBUTION

A random experiment which gives rise to only two outcomes say 'pass' and 'fail' is known as Bernoulli experiment. A random variable X is said to have Bernoulli distribution if its probability mass function

$$f(x ; p) = \begin{cases} p^x(1-p)^{1-x} ; x = 0,1 ; 0 \leq p \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

Here the probability of pass is p and probability of fail is (1 – p).

So, when x = 0; f (0; p) = 1 – p and x = 1 f (1; p) = p

It is to be noted that under Bernoulli distribution, the number of trial, is only 1 i.e. If we have to toss a coin to get 'head' as outcome, then the trial is just one toss to decide the outcome. So, the probability of getting head is p and probability of getting a tail is 1 – p.

Mean = E(x) = p

Variable = V(x) = p (1-p)

8.3.3 Binomial Distribution

Binomial distribution was introduced by James Bernoulli in the year 1700. Binomial distribution considers a sequence of Bernoulli trials i.e., having only two outcomes i.e. 'pass' and 'fail'. The n trials are conducted under identical conditions and are independent with constant probability.

Let X denotes success or pass in n independent trials with probability p as success and (1 – p) = q as failure.

So, f (X = r) = ${}^{n}C_{r}p^{r}q^{n-r}$;r=0,1,2,3…..n ; 0≤p≤1

0≤q≤1

Where n, p are the parameters of Binomial distribution

Mean = np and variance = npq or np (1–p)

Example: The probability that a student will pass an examination is 0.4. Determine the probability that out of 5 students: (i) at least one, (ii) at least two will pass the exam.

(i) p = P (passing the exam) = 0.4

q = P (failing the exam) = 1 – 0.4 = 0.6

P (at least one will pass the exam) = 1 – P (none will pass)

P (none will pass) = P (X = 0) = $(0.6)^5$ = 0.07776

P (at least one will pass the exam) = 1 – 0.07776

= 0.92224

(ii) P (at least two will pass) = P(X≥2) = 1- P(X<2)

$$P(X\geq 2) = 1-[P(X=0)+P(X=1)]$$

$$= 1-[0.0776+{}^{n}C_1 p^1 q^4$$

$$= 1-[0.07776+{}^{5}C_1 (0.4)^1 (0.6)^4]$$

$$= 1-[0.07776+0.2592]$$

$$= 0.66304$$

8.3.4 Poisson Distribution

Poisson distribution was developed by Simeon Denis Poisson in 1837.Here the random variable X represents successes in a specific interval of space and time. Its probability mass function is given by

$f(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$; x=0,1,2,3….. ; $\lambda > 0$

0 ; otherwise

where λ represents average number of successes.

E(X) = λ V (X) = λ i.e., mean and variance = λ

For eg: Noting the number of deaths in an area during the month.

The number of cars arriving at parking during a given period of time.

The number of errors made by typist per page.

The number of defective bulbs in a manufacturing unit etc.

The average number of customers per hour at a shop.

Visits to a particular website, e mail messages sent to a particular address.

Accidents in an industrial facility.

Cosmic ray showers observed by astronomers at a particular observatory.

These are some of the examples where Poisson distribution can be used.

For example, the number of customers at a shop is 4 in an hour. Find the probability that during an hour (i) no customer arrived (ii) 2 or more customer arrived at shop.

Let X be the number of customers at shop is an hour.

So, X follows Poisson distribution with $\lambda = 4$

As λ represents average number of successes.

So, every hour on average 10 customers arrives at the shop.

So, $f(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$; x=0,1, 2,3…...infinity

0 ; otherwise

(i) $f(X = 0) = \frac{e^{-4} \times 4^{0}}{0!} = 0.01831$

(ii) $f(X \geq 2) = 1 - f(X < 2)$

$= 1 - [f(x = 0) + f(x = 1)]$

$= 1 - [0.1831 + \frac{e^{-4}4^{1}}{1!}]$

$= 1 - [0.01831 + 0.07326]$

$= 0.90843$

8.3.5 Limiting Case of Binomial Distribution

If the probability of success in the binomial distribution is too small and number of trials are large, then binomial distribution can be approximated to Poisson distribution. In such an approximation the average number of successes is the mean of binomial distribution which is np i.e.

$\lambda = np$

Mathematically, if $n \to \infty$ and $p \to 0$ then binomial distribution is approximated to Poisson distribution.

For e.g., the probability of ineffective covid vaccine is 0.002, determine that out of 1000 individuals:

(i) exactly 2 will suffer Covid infection after being vaccinated.

(ii) more than 2 will suffer from infection.

Answer. since the p is 0.002 i.e., p → 0 and n is large i.e., 1000. So, the Binomial distribution will approximate to Poisson distribution with λ= np.

λ = np

$$\lambda = \frac{0.002}{\cancel{1000}} \times \cancel{1000}$$

$$\lambda = 2$$

(i) $$f(X=2) = \frac{e^{-2}2^2}{2!} = 0.2706$$

(ii) f(X>2)=1-f(X≤2)

$$= 1 - f(X=0) + f(X=1) + f(X=2)$$

$$= 1 - [\frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!} + \frac{e^{-2}2^2}{2!}]$$

$$= 1 - [0.1353 + 0.2706 + 0.2706]$$

$$= 0.3235$$

IN-TEXT QUESTIONS

1. The Probability of a car having flat tire while crossing a Bridge is 0.00006. Use the Poisson distribution to approximate the binomial probabilities that, among 10,000 cars crossing this Bridge. Exactly one will have a flat tire.
2. _____Distribution has only one trial.
3. If the Probability of binomial distribution is _____and Number of trials are_____ then Binomial distribution can be approximated to Poisson distribution.
4. Fit Binomial distribution to the following data.

X	0	1	2	3	4
F	25	68	45	12	5

8.3.6 Hyper Geometric Distribution

Hyper-geometric distribution is obtained if the population is finite, and sampling is done without replacement and events though random are statistically dependent. Consider an urn

with N Bals of which K are Red and remaining N-K are white. Let us draw a random sample of n balls without replacement. Then the probability of getting x red balls out of 'n' is given by

$$f(x) = \frac{{}^{K}C_{x} \times {}^{N-K}C_{n-x}}{N_{C_n}} \qquad ;x = 0, 1, 2, ...n$$

$$0 \qquad \text{otherwise}$$

For example: Let there be 5 economics and 10 commerce graduates. The organization requires 5 analysts that are to be chosen from the economics and commerce students. Find the probability of selecting 3 economics students for this job.

K = 5 = number of economic students

N – K = 10 = number of commerce student

N = 15 total student

x = selecting economics student as analyst

n = total student selected as analyst

$$f(x) = \frac{{}^{K}C_{x} \times {}^{N-K}C_{n-x}}{N_{C_n}}$$

$$f(x=3) = \frac{{}^{5}C_{3} \times {}^{10}C_{2}}{{}^{15}C_{5}}$$

$$= 0.1498$$

8.4 SUMMARY

A thorough knowledge of all the discrete distribution will help to assess the level of uncertainty and plan accordingly. All the distribution along with their probability mass function, mean and variance are shown in the following table.

Distribution	**Probability Mass Function PMF**	**Mean**	**Variance**
Uniform Distribution	$f(x=x) = f(x) = \frac{1}{n} \qquad x = 1, 2, 3.....n$	$\frac{n}{2}$	$\frac{n^2}{12}$
Bernoulli	$p^x(1-p)^{1-x}$	p	(1-p) p

Binomial	${}^{n}C_{r}p^{r}q^{n-r}$	np	np(1-p)
Poisson	$\frac{e^{\lambda}\lambda^{x}}{x!}$	λ	λ

8.5 ANSWER TO IN-TEXT QUESTIONS

1. Since p is 0.00006 le p tends to 0 and n is large i.e., 10,000
So, the Binomial distribution will approximate to Poisson distribution with λ = np.

λ= np

λ = 0.00006×100000

100000

λ =6

f(X=x)=$\frac{e^{-6}6^{1}}{1!}$

= 0.0148

So, Answer is 0.0148

2. Bernoulli

3. small; large

4. We have n=4 and N = ∑f =155

The mean of the given distribution

Mean= $\frac{\sum fX}{\sum f}$=$\frac{0X25+1X68+2X45+3X12+4X5}{155}$

1.380=$\frac{214}{155}$

Mean of Binomial Distribution is п * p = mean

$4p=\frac{214}{155}$

$p=\frac{107}{310}$; $q=\frac{203}{310}$

Expected binomial frequencies

$f(x)= N\ P(X=x)=155X C_0^4 X \left(\frac{107}{310}\right)^x X\left(\frac{203}{310}\right)^{4-x}$

X	P(X=x)	Expected Binomial Frequency N P(x)=155P(x)
0	$C_0^4 X \left(\frac{107}{310}\right)^0 X\left(\frac{203}{310}\right)^4$ =0.1838	28.50≅28
1	$C_1^4 X \left(\frac{107}{310}\right)^1 X\left(\frac{203}{310}\right)^3$ =0.3876	60
2	$C_2^4 X \left(\frac{107}{310}\right)^2 X\left(\frac{203}{310}\right)^2$ =0.3065	47.51≅48
3	$C_3^4 X \left(\frac{107}{310}\right)^3 X\left(\frac{203}{310}\right)^1$ =0.1077	16.69≅17
4	$C_4^4 X \left(\frac{107}{310}\right)^4 X\left(\frac{203}{310}\right)^0$	2.19≅2

X	0	1	2	3	4
F	28	60	48	17	2

8.6 SELF-ASSESSMENT QUESTIONS

(1) Probability of a car having a flat tyre is 0.00005 while crossing a bridge. Use the Poisson distribution to approximate the binomial probabilities of 10,000 cars crossing the bridge.

(a) exactly two have a flat tyre

(b) at most one has flat tyre

(2) Suppose car accidents in Delhi follow Poisson distribution with an average of one accident per day. What is the probability more than 5 accidents will occur in a week?

(3) Let X = B (n, p) with n = 25 & p = 0.3. find P(X< μ–2σ).

(4) In an air pollution survey, an inspector decides to examine 3 of a company's 30 factories. If 6 of the company's factories omit excessive pollutants and 24 factories follow all the standards. What is the probability the one of the factories will be included in the examination.

8.7 REFERENCES

- Devore, J. (2012). *Probability and Statistics for Engineers*, 8th ed. Cengage Learning
- John A. Rice (2007). *Mathematical Statistics and Data Analysts*, 3rd ed. Thomson Brooks/Cole.
- Miller, 1, Miller, M. (2017). J. Freund's *Mathematical Statistics with Applications*, 8th ed. Pearson.
- Hogg, R., Tanis, E., Zimmerman, D. (2021) *Probability and Statistical inference* 10th Edition, Pearson
- Larsen, R, Marx, M. (2011) *An Introduction to Mathematical Statistics and its Applications*, Prentice Hall.
- James McClave, P. George Benson, Terry Sincich (2017), *Statistics for Business* and *Economics*, Pearsons Publication.

LESSON 9

CONTINOUS DISTRIBUTION

STRUCTURE

9.1 LEARNING OBJECTIVES

After reading this lesson, students will be able to learn :

1. Uniform and exponential distributions.
2. Normal distribution with its properties.
3. Standard Normal distribution with its applications and
4. Central limit theorem and in applications.

9.2 INTRODUCTION

In previous lessons, you have learned about continuous random variables and associated functions. In this chapter you will read about different kinds of continuous distribution. Normal distribution is used extensively in economics and statistics. Several important applications of normal distribution with cognizable examples have been discussed. The central limit theorem is one of the most celebrated theorems in statistics and is used extensively.

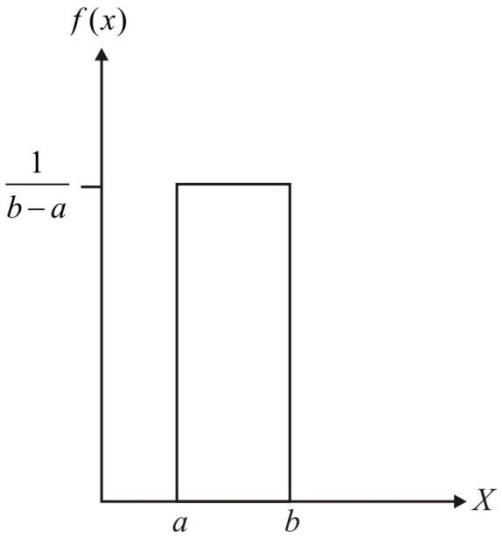
9.2.1 Uniform Distribution

A random variable X is said to have continuous uniform distribution over an interval (a, b) if its probability density function is constant over the entire range of X. Random variable x takes value between a and b.

So, the total probability is evenly distributed between the entire interval such that subintervals with same length have same probability.

$$f(X) = \frac{1}{b-a}, \quad a < X < b$$

$$0 \quad ; \text{otherwise}$$



Mean $\frac{b+a}{2}$ Variance $= \frac{(b-a)^2}{12}$

CDF $f(X \le x) = \int_a^x \frac{1}{b-a}\,dt = \frac{x-a}{b-a} = CDF$

For eg: If $y \sim U(50,140)$, what are $P(Y > 70)$ and $P(50 < Y < 130)$

$$f(x) = \frac{1}{140-50}, \text{ Since } a = 50,\ b = 140$$

$$f(x) = \frac{1}{90}$$

$$P(Y > 70) = \int_{70}^{140} \frac{1}{90}\,dx = \left[\frac{x}{90}\right]_{70}^{140}$$

$$= \frac{140}{90} - \frac{70}{90}$$

$$= \frac{70}{90} = 0.78$$

$$P(50 < Y < 130) = \int_{50}^{130} \frac{1}{90}\,dx$$

$$= \frac{130}{90} - \frac{50}{90} = 0.89$$

9.2.2 Exponential Distribution

A continuous non-negative random variable X is said to have an exponential distribution with parameters λ if its pdf is given by

$$f(X) = \lambda e^{-\lambda x} \text{ ; } x \geq 0$$

$$0 \text{ ; otherwise}$$

Exponential distribution is also known as waiting distribution where waiting parameter is λ.

So, $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$

Cumulative density function

$$P(X \leq x) = \int_0^x \lambda e^{-\lambda t} dt$$

$$P(X \leq x) = \frac{\lambda}{-\lambda} e^{-\lambda t} \Big]_0^x$$

$$P(X \leq x) = 1 - e^{-\lambda x}$$

Le us consider that probability of getting a success x in a time interval t and $t + \Delta t$ is λ.

$$P(X > x) = 1 - P(X \leq x)$$

$$= 1 - [1 - e^{-\lambda x}]$$

$$= e^{-\lambda x}$$

Example: A call center receives 4 calls per hour. What is the probability that next call arrives after ½ hours?

Solution: So $\lambda = 4$

$$P\left(X > \frac{1}{2}\right) = \int_{1/2}^{\infty} (4e^{-4x}) dx$$

$$= 4\left[\frac{e^{-4x}}{-4}\right]_{1/2}^{\infty}$$

$$= -e^{-4\infty} + e^{-4/2}$$

$$= -e^{-\infty} + e^{-2}$$

$$= 0.13533$$

9.2.3 Normal Distribution

This is the most important distribution developed in 1733 by French Mathematician De Moivre. The normal distribution is also called Gaussian distribution as German Mathematician Friedrich Gauss (1777-1855) derived in equation. It is a symmetric bell-shaped curve

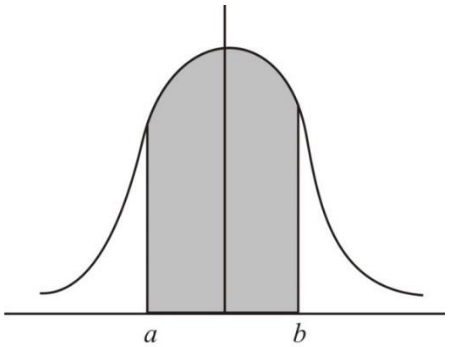
A random variable X is called normal random variable

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

Where constant π=3.14 and e=2.718. μ and σ are the two parameters of the distribution and X is a real number denoting the continous random variable of interest.

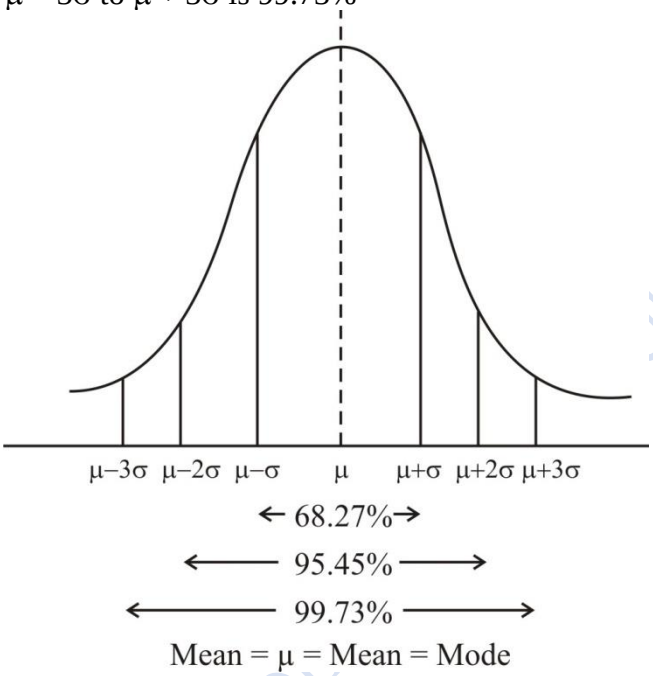
Properties

- It is symmetric through the mean.
- Because of the symmetry at the points of inflexion at ±σ distance, the normal curve has a bell shape
- The right and left tails of the curve extend infinitely without touching the horizontal x axis.
- In normal distribution Mean = Median = Mode.
- The area between two points a and b is represented by the shaded region.



Mean = μ = Mean = Mode

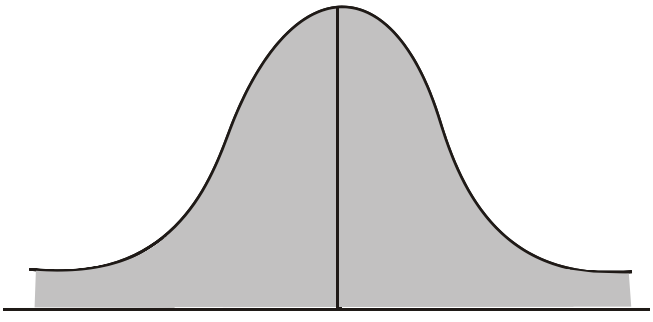
- Area between μ – σ to μ + σ is 68.27%.
- Area between μ – 2σ to μ + 2σ is 95.44%
- Area between μ – 3σ to μ + 3σ is 99.73%



9.2.4 Standard Normal Distribution

Making the transformation, $Z=\frac{X-\mu}{\sigma}$, we obtain the standard normal variable Z that has mean μ = 0 and standard deviation σ = 1.

Area under the standard normal curve = 1.



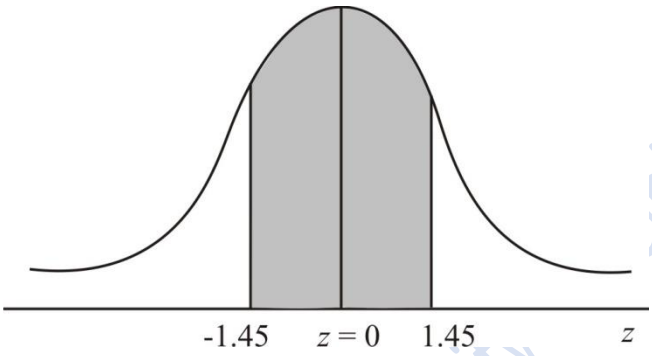
Property O

(1) $p(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$

(2) In standard Normal curve areas at the right and left of 0 is 0.5
Normal curve is symmetric So, any area between 0 and a particular point c and area between 0 and point –c will be same.

$$P(-c < z < 0) = P(0 < z < c)$$

For e.g., Area between 0 and 1.45 will be equal to area between O and –1.45.

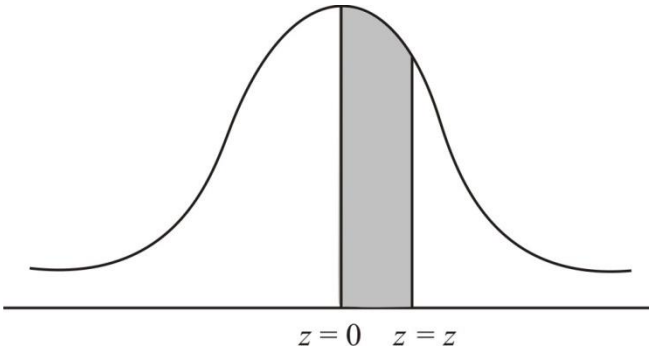


$$P(0 < z < 1.45) = P(-1.45) < z < 0) = 0.4265$$

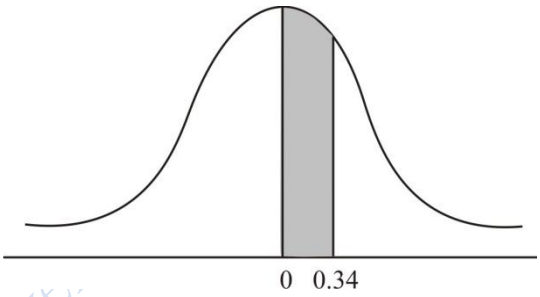
Now you all must be wondering how I got the value 0.4264. For this you have to learn to look at the standard normal table.

The probability values of the different values of z are given in the table, which represents the area under the standard normal curve.

z / Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517



So, to get the value of $P(0 < Z < 0.34)$, we have to search 0.3 in column and 0.04 is first row. So, we get the value 0.1331.



Similarly, if we want $P(-0.34 < Z < 0.34)$ then as we know that standard normal curve is symmetrical and

$$P(0 < Z < 0.34) = P(0.34 < Z < 0)$$

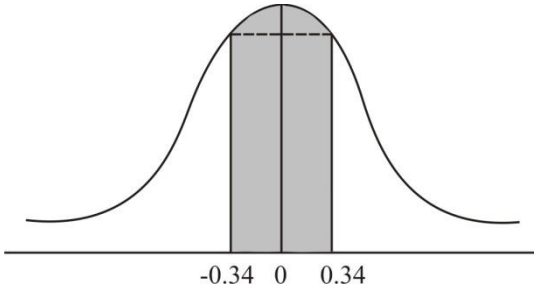
So, $$P(-0.34 < Z < 0.34) = P(-0.34 < Z < 0) + P(0 < Z < 0.34)$$

$$P(-0.34 < Z < 0.34) = P(0 < Z < 0.34) + P(0 < Z < 0.34)$$

$$P(-0.34 < Z < 0.34) = 2P(0 < Z < 0.34)$$

$$P(-0.34 < Z < 0.34) = 2 \times 0.1331$$

$$P(-0.34 < Z < 0.34) = 0.2662$$



Suppose X be a normally distributed random variable with mean μ and variance σ Now, normally distributed random variable can be transformed to standard normal distribution i.e., $z = \frac{X - \mu}{\sigma}$

Now the new random variable z will have mean = 0 and standard deviation 1.

Example: If X is a normally distributed with mean μ = 30 and variance $\sigma^2 = 16$. Find

1. $P(X > 15)$
2. P(10< X<25)
3. $P(12 < X < 36)$

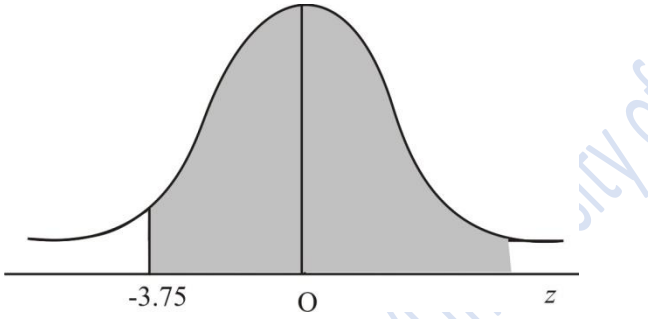
Sol. First, we have to convert X to z by transformation

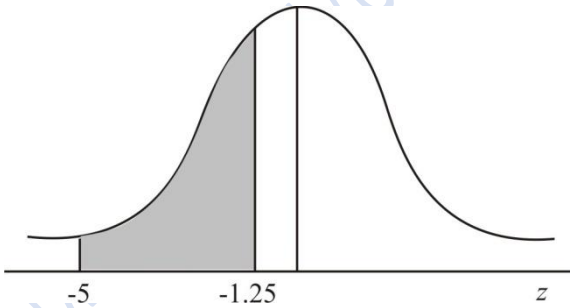
$$Z = \frac{X-\mu}{\sigma} = \frac{X-30}{4}$$

(I) For $X = 15, Z = \frac{15-30}{4} = \frac{-15}{4} = -3.75$

$$P(Z > -3.75) = P(X > 15) = P(-3.75 < Z < 0) + 0.5$$
$$= 0.4999 + 0.5$$
$$= 0.9999$$

(II)



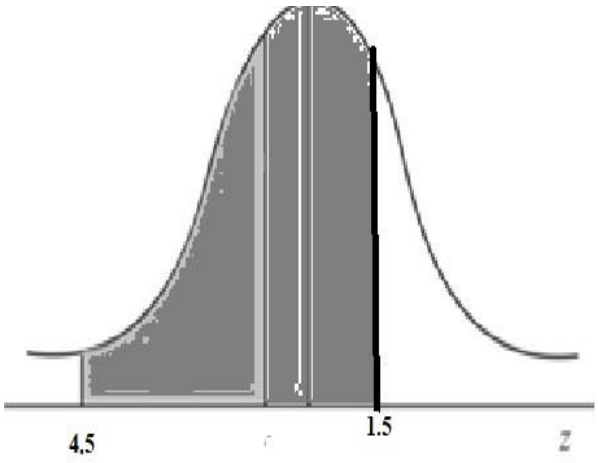


(III)
$$P(10 < X < 25) = P\left(\frac{10-30}{4} < \frac{X-\mu}{\sigma} < \frac{25-30}{4}\right)$$
$$= P(-5 < Z < -1.25)$$
$$= P(-5 < Z < 0) = P(0 < Z < -1.25)$$

$$= 0.49999 - 0.3944$$
$$= 0.1055$$

(IV)
$$P(12 < X < 36) = P\left(\frac{12-30}{4} < \frac{X-\mu}{\sigma} < \frac{36-30}{4}\right)$$
$$= P(-4.5 < \frac{X-\mu}{\sigma} < 1.5)$$

$= P(-4.5 < Z < 1.5)$
$= (-4.5 < Z < 0) + P(0 < Z < 1.5)$
= 0.4999 + 0.4332
= 0.9331



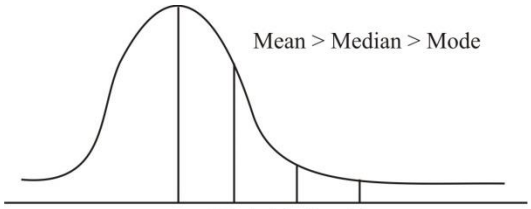
IN-TEXT QUESTIONS

1. In normal distribution, 34% of the items are under 50 and 5% of items are over 70. Find the mean and variance of the distribution.
2. Standard normal distribution has mean ……... and variance……

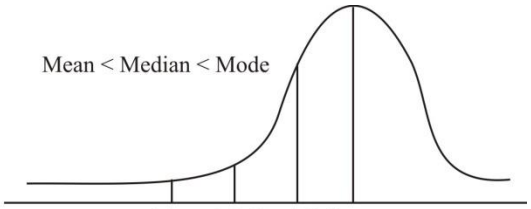
SKEWNESS

When a distribution aperture from its symmetry then it said to have a skewed distribution. There are two types of skewed distribution, i.e., positive and negatively skewed distribution. A positive skewed distribution is skewed to right or have a longer tail at the right side.

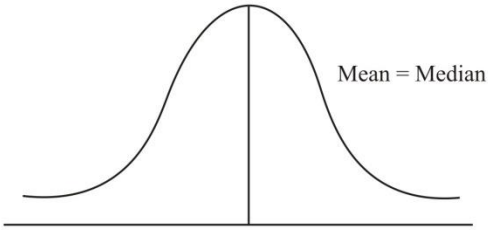
Similarly, negatively skewed distribution is skewed towards left or have longer tail at the left side.



Positively Skewed



Negatively Skewed



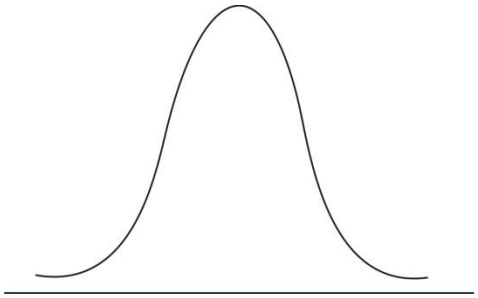
Mean = Median = Mode

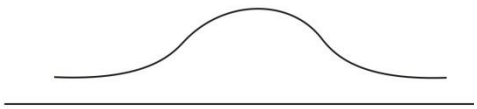
KURTOSIS

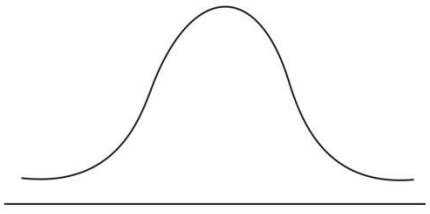
The degree of peakedness is determined by the kurtosis. A high peak distribution is characterized as leptokurtic

A low peak distribution is characterized as Platykurtic

The normal distribution is Mesokurtic neither too peaked nor too low.


Leptokurtic


Platykurtic


Mesokurtic

So, normal distribution is symmetric i.e., neither positive nor negatively skewed and they are mesokurtic i.e., neither too high peaked nor too low peaked.

9.2.5 Central Limit Theorem

According to the central limit theorem, if $x_1, x_2, x_3, ..., x_n$ are independently identically distributed (IID) random variable following normal distribution with Mena μ and Variance σ^2 then central limit theorem will help to find the distribution of $\bar{X}$ and ΣXi random variable.

So, to find the distribution of Random variable ΣX_i

Rule $E(X+Y) = E(X) + E(Y)$

Since $X_1, X_2, X_3, ..., X_n \sim N(\mu, \sigma^2)$, So $E(X_1) = \mu, E(X_2) = \mu$ and so on.

$$\Sigma X_i = X_1 + X_2 + ... + X_n$$

$$E(\Sigma X_i) = E(X_1 + X_2 + ... + X_n)$$

$$E(\Sigma X_i) = E(X_1) + E(X_2) + ... + E(X_n)$$

$$E(\Sigma X_i) = \mu + \mu + ... + \mu$$

So, adding μ n times will give the result as $E(\Sigma X_i) = \mu n$ Eq. (1)

To find variance of $V(\Sigma X_i) = V(X_1 + X_2 + X_3 + ... X_n)$.

Since $X_1 + X_2 + X_3 + ... X_n \sim N(\mu, \sigma^2)$, So, $V(X_1) = \sigma^2,\ V(X_2) = \sigma^2 ..... V(X_n) = \sigma^2$

$$V(\Sigma X_i) = V(X_1) + V(X_2) + V(X_3) + ... V(X_n)$$

$$V(\Sigma X_i) = \sigma^2 + \sigma^2 + ... + \sigma^2$$

So, adding σ^2 n times will give the result as

$$V(\Sigma X_i) = n\sigma^2$$ eqn. (2)

So, $\Sigma X_i \sim N(n\mu, n\sigma^2)$ and

$$z = \frac{\Sigma X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$$

Now, let us find the distribution of random variable $\bar{X}$

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

$$E(\bar{X}) = E\left(\frac{\Sigma X_i}{n}\right) \qquad E(X_i) = n\mu$$ From (1)

$$= \frac{1}{n} E(\Sigma X_i)$$

$$= \frac{1}{n} n\mu$$

$$= \mu$$

$$V(\bar{X}) = V\left(\frac{\Sigma X_i}{n}\right)$$ # Rule $V(aX) = a^2 V(X)$

$$V(\bar{X}) = \frac{V(\Sigma X_i)}{n^2}$$ $V(\Sigma X_i) = n\sigma^2$ from (2)

$$V(\bar{X}) = \frac{n\sigma^2}{n^2}$$

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \quad \text{or} \quad \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Ques. In a large population the distribution of a variable has a mean of 165 and standard deviation 25 units. If a random sample of size 35 is chosen, find the approximate probability that the sample mean lies between 162 and 170.

Solution: $X \sim N(165, 25^2)$

Where, sample size (n) = 35

We have to find the distribution of sample mean

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, Since $\mu = 165$, $\sigma^2 = 25^2$

$\bar{X} \sim N\left(165, \frac{25^2}{35}\right)$

Note that $\frac{25^2}{35}$ i.e., $\frac{\sigma^2}{n}$ is variance

Standard deviation is $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

To find $P(162 < \bar{X} < 170) = P\left(\frac{162-165}{\frac{25}{\sqrt{35}}} < z < \frac{170-165}{\frac{25}{\sqrt{35}}}\right)$

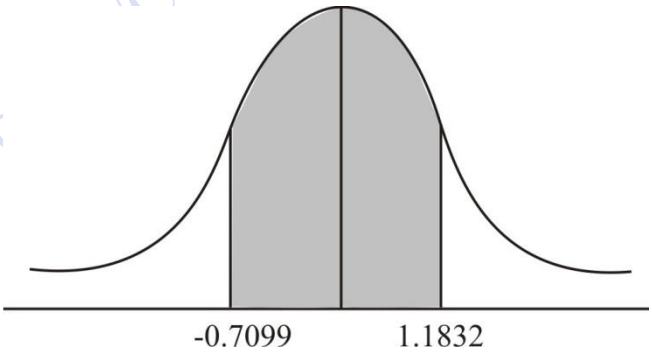
$= P(-0.70 < z < 1.18)$

$= P(z < 1.18) + P(z < 0.70)$

as $P(z < 0.70) = P(z < -0.70)$

$= 0.3810 + 0.2580$

$= 0.639$



As the sample size increases then distribution of $\bar{X}$ will tend to normal distribution. For a distribution to be approximated to normal distribution, sample size must be at least 30 or in other words $n \geq 30$. As the sample size increases, even discrete distribution approximates normal distribution.

IN-TEXT QUESTIONS

3. Consider a random sample of size 30 taken from a Normal distribution with Mean 60 and variance 25. Let the sample mean be denoted by $\bar{X}$. So, calculate the probability that $\bar{X}$ assumes a value greater than 62.

4. Mean and variance of distribution of random variable $\bar{X}$ is …….... and ……. respectively.

9.3 SUMMARY

Continuous distributions are most extensively used in economics. Thorough knowledge of these distributions will help you to solve the self-assessment questions properly. The central limit theorem is one of the most celebrated theorems of statistics. It states that all discrete and continuous distribution will approximate to normal distribution with increase in sample size i.e., $n \geq 30$. We can also represent all the continuous distributions with their probability density function, mean and variance.

Distribution	Probability Density function	Mean	Variance
Uniform distribution	$\frac{1}{b-a}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Exponential distribution	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal distribution	$-\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$	μ	σ^2
Standard normal distribution	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	0	1

9.4 ANSWER TO IN-TEXT QUESTIONS

1. As per the given information
$f(X < 50) = 0.34$ and $f(X > 70) = 0.05$
Let μ and σ the mean and variance of X. Converting X into standard normal variable z.
For $X = 50,\ z_1 = \frac{50-\mu}{\sigma}$ and for $X = 70,\ z_2 = \frac{70-\mu}{\sigma}$

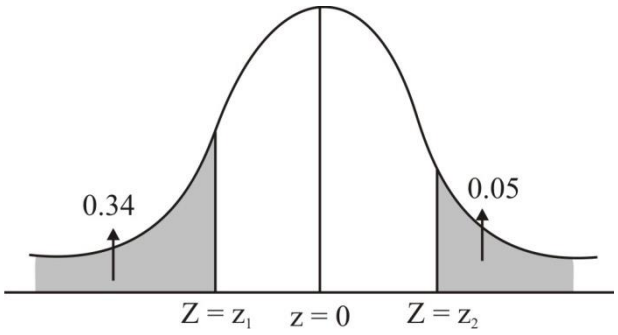
The respective areas are represented on the graph

$P(Z < z_1) = 0.34$

$P(Z_1 < Z < 0) = 0.5 - 0.34$

i.e., it represents area between 0 and $-z_1$

$P(0 < Z < -z_1) = 0.16$



So, value of $-z_1$ can be found through the standard normal table. The area 0.16 is represented through 0.42

So, $-z_1 = 0.42$

$$-\left(\frac{50-\mu}{\sigma}\right) = 0.42$$

$$50 - \mu = -0.42\sigma \quad (1)$$

$P(X > 70) = 0.05$

$P(Z > z_2) = 0.05$

$P(0 < Z > z_2) = 0.5 - 0.05 = 0.45$

So, the value of z_2 from the standard normal table

$z_2 = 1.66$

$$\frac{70-\mu}{\sigma} = 1.66$$

$$70 - \mu = 1.66 \quad (2)$$

Solving eqn. (1) and (2) we get

$\sigma = 9.61$ and $\mu = 54$

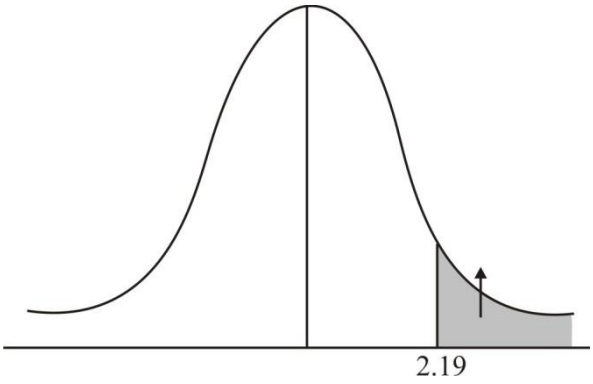
2. The standard normal distribution has a mean of 0 and variance 1.

3. $n = 30,\ \mu = 60,\ \sigma^2 = 25$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X} \sim N\left(60, \frac{25}{30}\right)$$

$$P(\bar{X} > 62) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > \frac{62-60}{\frac{5}{\sqrt{30}}}\right)$$



4. $P(\bar{X} > 2.19) = 0.5 - P(0 < z < 2.19)$

$P(Z > 2.19) = 0.5 - 0.4857$

$P(Z > 2.19) = 0.0143$

$\mu;\ \frac{\sigma^2}{n}$

9.5 SELF-ASSESSMENT QUESTIONS

1. A magazine claims that 30% of its readers are Students A random sample of 100 readers is taken & is found to Contain 25 students. Calculate the probability of obtaining 25 or fewer students' readers assuming that the magazine's claim is correct.
2. A fair die is tossed 150 times Determine that the face 6 will appears.
    1. between 20 and 30 times inclusive
    2. less than 20 time
3. A call center Receives 6 calls per hour. What is the probability that next call arrives after ¼ hour.
4. Only 4 Students came to attend the class today, find the portability for exactly 6 students to attend the class tomorrow.
5. A random variable X has the distribution B (10, p), Given that p= 0.30 find
    1. P (X < 4)
    2. P(X≥8)

9.6 REFERENCES

- Devore, J. (2012). *Probability and Statistics for Engineers*, 8th ed. Cengage Learning
- John A. Rice (2007). *Mathematical Statistics and Data Analysts*, 3rd ed. Thomson Brooks/Cole.
- Miller, 1, Miller, M. (2017). J. Freund's *Mathematical Statistics with Applications*, 8th ed. Pearson.
- Hogg, R., Tanis, E., Zimmerman, D. (2021) *Probability and Statistical Inference* 10th Edition, Pearson
- Larsen, R, Marx, M. (2011) *An Introduction to Mathematical Statistics and its Applications*, Prentice Hall

LESSON 10

JOINT PROBABILITY DISTRIBUTION AND MATHEMATICAL EXPECTATIONS

STRUCTURE

10.1 LEARNING OBJECTIVE

After reading this lesson, student will be able to :

1. Identify the probability distribution of two or more events occurring together
2. Calculate marginal distributions of more than one variable of discrete and continuous distributions

3. Calculate conditional probability and verify independence of probability distributions and
4. Calculate mathematical expectations of joint probability mass function and joint probability density functions

10.2 INTRODUCTION

This chapter deals with the probability distribution of two or more random variables called joint probability distribution. There are two types of joint probability distribution. One is probability mass function (PMF) and the other is probability density function (PDF). In the case of joint probability distribution of two discrete variables, the probability distribution function is called probability mass function. In the case of continuous variables, it is called probability density function. The chapter first deals with the joint probability mass function and then probability density function in the second half of the chapter.

10.3 JOINT PROBABILITY MASS FUNCTION

Joint probability mass function is related to the probability distribution of two discrete variables. It is characterized by the following features.

Let X and Y be the two discrete random variables on the sample space S. The joint probability mass function (PMF) is given by

$$P(x, y) = P(X - x \text{ and } Y = y)$$ where

(x, y) is a pair of possible values for the pair of random variables (X, Y) and $P(x, y)$ must satisfy the following conditions –

(a) $0 \le P(x, y) \le 1$

(b) $\sum_x \sum_y P(x, y) = 1$

The probability $P[(X, Y) \in A]$ is obtained by summing the joint PMF.

(c) $$P[(X, Y) \in A] = \sum_{(x,y) \in A} \sum P(x, y)$$

It must be noted that conditions (a) and (b) are required for $P(x, y)$ to be a valid joint PMF.

Example-1

Consider two random variables X and Y with joint PMF as shown in the table below:

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	1/6	1/4	1/8

$X = 1$	1/8	1/6	1/6

Find the following

(i) $P(X = 0, Y \leq 1)$

(ii) $P(Y = 0, X \leq 1)$

Solution:

(i) $$P(X = 0, Y \leq 1) = P_{XY}(0,0) + P_{XY}(0,1)$$
$$= \frac{1}{6} + \frac{1}{4} = \frac{5}{12}$$ Answer.

(ii) $$P(Y = 2, X \leq 1) = P_{XY}(0,2) + P_{XY}(1,2) = \frac{1}{8} + \frac{1}{6} = \frac{7}{24}$$

Example-2

A function f is given by $f(x,y) = cxy$ for $x = 1,2,3 ; y = 1,2,3$

Determine the value of c for which the above function $f(x,y)$ *validate as true* $p.m.f$.

Solution:

From the question given,

$$f(x,y) = cxy$$

Where, $x = 1,2,3$ *and* $y = 1,2,3$.

There are 9 possible pairs of X *and* Y, namely $(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2)$ *and* $(3,3)$. The probabilities associated with each of the pairs are:

$$f(1,1) = c(1)(1) = c$$
$$f(1,2) = 2c, \quad f(1,3) = 3c, \quad f(2,1) = 2c$$
$$f(2,2) = 4c, \quad f(2,3) = 6c, \quad f(3,1) = 3c$$
$$f(3,2) = 6c, \quad f(3,3) = 9c$$

For $f(x,y)$ to be a valid joint pmf,

$$\sum_x \sum_y f(x,y) = 1$$

Hence,

$$\sum_{x=1}^{3}\sum_{y=1}^{3} f(x,y) = c + 2c + 3c + 2c + 4c + 6c + 3c + 6c + 9c = 1$$

$$36c = 1$$

$$c = \frac{1}{36}$$

Thus, for, $c = \frac{1}{36}$ the given function is a valid probability mass function.

10.3.1 Conditional Probability

Conditional probability has already been discussed earlier. It is once again reiterated that conditional probability is a measure of the probability of an event occurring given that another event has already occurred.

Conditional probability is denoted by $P((A \mid B)$ where

$$P((A \mid B) = \frac{P(A \cap B)}{P(B)}; \qquad P(B) > 0$$

In this chapter, we deal in joint probability of two random variable *X* and *Y*. The conditional probability of which is given by

$$\boxed{P[X \in C \mid Y \in D] = \frac{P(X \in C, Y \in D)}{P(Y \in D)}}$$

where $C, D \subset R$.

For discrete random variables *X* and *Y*, the conditional PMFs of *X* given *Y* and *Y* given *by X* respectively are given by

$$P_{X|Y}(x_i / y_j) = \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)} \quad \{\text{for any } x_i \in R_X \text{ and}$$

$$P_{Y|X}(y_j / x_i) = \frac{P_{XY}(x_i, y_j)}{P_Y(x_i)} \quad y_j \in R_Y$$

Example-3

Consider two random variables *X* and *Y* with joint PMF as shown in the table below

	Y = 2	*Y* = 4	*Y* = 5
X = 1	1/12	1/24	1/24

$X = 2$	1/6	1/12	1/8
$X = 3$	1/4	1/8	1/12

Find the following –

(a) $P(X \le 2, Y \le 4)$

(b) $P(Y = 2 \mid X = 1)$

Solution:

(a) $$P(X \le 2, Y \le 4)$$
$$= P_{XY}(1,2) + P_{XY}(1,4) + P_{XY}(2,2) + P_{XY}(2,4)$$
$$= \frac{1}{12} + \frac{1}{24} + \frac{1}{6} + \frac{1}{8} = \frac{3}{8}$$

(b) $$P(Y = 2 \mid X = 1) = \frac{P(X = 1, Y = 2)}{P(X = 1)}$$
$$= \frac{P_{XY}(1,2)}{P_X(1)} = \frac{1}{12} \div \frac{1}{6} = \frac{1}{2}$$ Ans.

10.3.2 Independence of Random Variables

In the case of joint PMF, criteria for the independence of two discrete random variables X and Y are given by –

$$P_{XY}(x, y) = P_X(x) \cdot P_Y(y) \quad \forall\ x, y$$

The above condition must fulfil for two discrete random variables X and Y is independent.

Example-4

From the question in example 3, check if X and Y are independent.

Solution:

For X and Y to be independent.

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

for all $x_i \in R_X$ and for all $y_j \in R_Y$

$$P(X = 2, Y = 2) = \frac{1}{6}$$

$$P(X = x_i) \cdot P(Y = y_j) = \frac{3}{8} \times \frac{1}{2} = \frac{3}{16}$$

$$\frac{1}{6} \neq \frac{3}{16}$$

$\because$

$\therefore$ X and Y are not independent.

10.3.3 Marginal Probability Mass Functions

If (X, Y) are discrete variables, then marginal probability is the probability of a single event that occur independent of another event.

The marginal probability mass function of X_i is obtained from the joint PMF as shown below–

$$P_{X_i}(x) = \sum_{X_1 \ldots X_k} P_X(x_1, x_2, \ldots, x_k)$$

In words the marginal PMF of X_i at the point X is obtained by taking the sum of the joint PMF P_X out all the vectors that belong to R_X in such a way that is component is equal to X.

Example-5

Carrying forward from example 3, find the marginal PMFs of X and Y.

Solution

$$R_X = \{1, 2, 3\}, \quad R_Y = \{2, 4, 5\}$$

Marginal PMFs are given by

$$P_X(x) = \begin{cases} \frac{1}{6}, & \text{for } X = 1 \\ \frac{3}{8} & \text{for } X = 2 \\ \frac{11}{24} & \text{for } X = 3 \\ 0 & \text{Otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} \frac{1}{2}, & \text{for } Y = 2 \\ \frac{1}{4} & \text{for } Y = 4 \\ \frac{1}{4} & \text{for } Y = 5 \\ 0 & \text{Otherwise} \end{cases}$$

10.3.4 Expectation of a PMF

Let X and Y be a jointly distributed Random variable with probability mass function $P(x, y)$ with discrete variables. Then the expected value of function $g(x, y)$ is given by

$$E[g(X,Y)] = \sum_x \sum_y g(X,Y) \cdot P(x, y)$$

Example-6

Find E(XY) for data given in example 2

Solution:

$$E(XY) = \sum_x \sum_y xyP(x, y)$$

$$= \left(1 \times 2 \times \frac{1}{12}\right) + \left(1 \times 4 \times \frac{1}{24}\right) + \left(1 \times 5 \times \frac{1}{24}\right) + \left(2 \times 2 \times \frac{1}{6}\right) + \left(2 \times 4 \times \frac{1}{12}\right) + \left(2 \times 5 \times \frac{1}{8}\right) + \left(3 \times 2 \times \frac{1}{4}\right) + \left(3 \times 4 \times \frac{1}{8}\right) + \left(3 \times 5 \times \frac{1}{12}\right)$$

$$= \frac{177}{24} = 7.38.$$

IN–TEXT QUESTIONS

Answer the following MCQs

1. Let $U \in \{0,1\}$ and $V\{0, 1\}$ be two independent binary variables. If $P(U = 0) = P$ and $P(V = 0)$ = q, when $P(U + V) \geq 1$ is

   (a) $pq + (1 - p)(1 - q)$

   (b) pq

   (c) $p(1 - q)$

(d) $1 - pq$

2. If a variable can take certain integer values between two given points, then it is called–
   (a) Continuous random variable
   (b) Discrete random variable
   (c) Irregular random variable
   (d) Uncertain random variable

3. If $E(U) = 2$ and $E(V) = 4$ then $E(U - V) = ?$
   (a) 2
   (b) 6
   (c) 0
   (d) Insufficient data

4. Height is a discrete variable (T / F)

5. If X and Y are two events associated with the same sample space of a random experiment. then $P(X \mid Y)$ is given by
   (a) $P(X \cap Y) / P(Y)$ provided $P(Y) \neq 0$
   (b) $P(X \cap Y) / P(Y)$ provided $P(Y) = 0$
   (c) $P(X \cap Y) / P(Y)$
   (d) $P(X \cap Y) / P(X)$

6. Let X and Y be events of a sample space S of an experiment. If $P(S \mid Y) = P(Y \mid Y)$ then the value of $P(Y \mid Y)$ is
   (a) 0
   (b) –1
   (c) 1
   (d) 2

7. What are independent events? Choose the correct option:
   (a) If the outcome of one event does not affect the outcome of another.
   (b) If the outcome of one event affects the outcome of another.
   (c) Any one of the outcomes of one event does not affect the outcome of another.
   (d) Any one of the outcomes of one event does affect the outcome of the other.\

10.4 CONTINUOUS RANDOM VARIABLES

The probability that the observed value of a continuous random variable X lies in a one-dimensional set A, is obtained by integrating the probability density function (PDF) $f(x)$ over the set A.

Similarly, the probability that the pair (X, Y) of a continuous random variable fall in a two-dimensional set A is obtained by integrating the joint PDM.

Joint density function is a piecewise continuous function of two variables $f(x, y)$, such that for any "reasonable" two-dimensional set B

$$P(X, Y) \in A = \iint_A f(x, y)dydy \quad .$$

Definition: Let X and Y be continuous random variables. A joint density function $f(x)$ for these two variables is a function satisfying

(a) $f(x, y) \geq 0$
and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

then for two-dimensional set A

$$P[(X, Y) \in A] = \iint_A f(x, y)dydx$$

Example-7

The joint PDF of (X, Y) is given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

answer the following

(a) Verify that $f(x, y)$ is a legitimate PDF.

(b) Find $P\left(0 \leq x \leq \frac{1}{n},\ 0 \leq y \leq \frac{1}{n}\right)$

Solution: (a) Two conditions must be satisfied for $f(x, y)$ to be a legitimate PDF

(i) $f(x,y) \geq 0$ and

(ii) $\int_{-\infty}^{\infty o} \int_{-\infty}^{\infty} f(x,y)dxydy = 1$

the first condition is fulfilled as $f(x,y) \geq 0$ for the verification of the second condition –

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = \int_{0}^{1} \int_{0}^{1} \frac{6}{5}(x+y^2)dxdy$$

$$= \int_{0}^{1} \int_{0}^{1} \frac{6}{5}x\,dxdy + \int_{0}^{1} \int_{0}^{1} \frac{6}{5}y^2dxdy$$

$$= \int_{0}^{1} \frac{6}{5}x\,dxdy + \int_{0}^{1} \frac{6}{5}y^2dxdy$$

= 1. This second condition is also verified.

(b) $P\left(0 \leq X \leq \frac{1}{4}, 0 \leq Y \leq \frac{1}{4}\right)$

$$= \int_{0}^{1/4} \int_{0}^{1/4} \frac{6}{5}(x+y^2)dxdy$$

$$= \frac{6}{5} \int_{0}^{1/4} \int_{0}^{1/4} x\,dxdy + \frac{6}{5} \int_{0}^{1/4} y^2dxdy$$

$$= \frac{6}{20} \times \frac{x^2}{2}\Big|_{x=0}^{x=\frac{1}{4}} + \frac{6}{20}\frac{y^3}{3}\Big|_{y=0}^{y=\frac{1}{4}}$$

$$= \frac{7}{640}$$ Ans.

Example-8

Consider two continuous random variables X and Y with joint p.d.f.

$$f(x,y) = \begin{cases} \frac{2}{81}x^2y, & 0 < x < K, 0 < y < K \\ 0, & otherwise \end{cases}$$

a) find the value of K so that $f(x,y) is\ a\ valid\ \ p.d.f.$

b) find P(X>3Y)

Solution:

a) for $f(x,y)\ to\ be\ a\ valid\ \ p.d.f.$ conditions of continuous p.d.f. must satisfy

therefore, $\int_0^k \int_0^k \frac{2}{81}x^2ydxdy = 1 \quad = \frac{K^5}{243}$ => K = 3

b) P(X>3Y) = $\int_0^3 \left(\int_0^{\frac{x}{3}} \frac{2}{81}x^2ydy \right) dx$

$= \int_0^3 \frac{1}{729}x^4dx$

$= \frac{1}{15}$

10.4.1 Marginal Probability Density Function

Marginal PDF in the continuous distribution variable can be obtained in a similar manner as in the case of discrete variables.

MDF can be obtained by integrating the joint PDF of one variable keeping the other constant.

The marginal PDF of *X* and *Y* denoted by $f_X(x)$ and $F_Y(Y)$, respectively is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(x) = \int_{-\infty}^{\infty} f(x,y)dx \quad \text{for } -\infty < y < \infty$$

Example 9

Find the MDF $f_X(x)$ and $f_Y(y)$ in example 3.

Solution

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$= \int_0^1 \frac{6}{5}(x+y^2)dy = \frac{6x}{5} + \frac{2}{5}$$

$$f_X(x) = \begin{cases} \frac{6}{5}x + \frac{2}{5}, & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

$$= \int_0^1 \frac{6}{5}(x+y^2)dx$$

$$= \frac{6}{5}y^2 + \frac{3}{5}$$

$$\therefore \quad f_Y(y) = \begin{cases} \frac{6}{5}y^2 + \frac{3}{5} & for\ 0 \le y \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

10.4.2 Expected value of a PDF

Let *X* and *Y* be a continuous random variable with joint PDF $f(x,y)$. Let *g* be some function, then

$$E[g(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,ly),\ f(x,y)\,dxdy$$

Example-10

The length of a thread is 1 mm, and two points are chosen *U*niformly and independently along the thread. Find the expected distance between these two points.

Solution

Let *U* and *V* be the two points that are chosen. The joint PDF of *U* and *V* is

$$f(U,V)=\begin{cases}1 & 0<U,V<1\\ 0 & \text{otherwise}\end{cases}$$

$$E[U-V]=\int_{01}^{1}\int_{0}^{1}|U-V|\,dUdV$$

$$E[U-V]=\iint_{U\geq V}(U-V)\,dUdV+\int_{V>U}(U-V)dUDV$$

$$=\int_{0}^{1}\int_{0}^{1}(U-V)dUdV+\int_{0}^{1}\int_{0}^{1}(V-U)dUdV$$

$$E[U-V]=\frac{1}{3}$$

Example 11

The joint PDF of X and Y is given by

$$f(x,y)=\begin{cases}\frac{3}{7}(x+y) & 0<x<1\\ 0 & \text{otherwise}\end{cases}$$

find the expected value of X/Y^2.

Solution

$$E[X,Y^2]=\int_{1}^{2}\int_{0}^{1}\frac{3x(x+y)}{7y^2}dxdy$$

$$=\frac{3}{7}\int_{1}^{2}\left(\frac{1}{3y^2}+\frac{1}{y}\right)dy$$

$$E[X,Y^2]=\frac{3}{28}\ \text{Ans.}$$

10.4.3 Conditional Distributions

Conditional PDF of X, given that $Y=y$ is denoted by

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

and the conditional expected value of X given $Y = Y$ is given by

$$E[X \mid Y] = \int x f_{X|Y}(x \mid y) dx$$

Similarly, one can define the conditional PDF, expected value of Y given $X = X$ by interchanging the rate of X and Y.

Properties of Conditional PDFs

The conditional PDF for X, given $Y = Y$ is a valid PDF if two conditions are satisfied–

(1) (a) $0 \le f_{X|Y}(x,y)$

(b) $\int f_{X|Y}(x \mid y) dx = 1$

(2) The conditional distribution of X given Y does not equal the conditional distribution of Y given X.

i.e. $f_{X|Y} f(x \mid y) \neq f_{Y|X}(y \mid x)$

Example 12

If the joint PDF of U and V is given by

$$f(U,V) = \begin{cases} \frac{2}{3}(U+V) & 0 < U < 1,\ 0 < V < 1 \\ 0 & \text{Otherwise} \end{cases}$$

find the conditional mean of U given $V = 1/2$.

Solution:

Let U and V be the joint PDF

$$f(U \mid V) = \begin{cases} \frac{2U + 4V}{1 + 4V} & 0 < U < 1 \\ 0 & \text{elsewhere} \end{cases}$$

so that

$$f\left(U \middle| \frac{1}{2}\right) = \begin{cases} \frac{2}{3}(U+1) & 0 < U < 1 \\ 0 & \text{otherwise} \end{cases}$$

then

$$E\left[U\left|\frac{1}{2}\right.\right]=\int_0^1 \frac{2}{3}U(U+1)dV$$

$$E\left[U\left|\frac{1}{2}\right.\right]=\frac{5}{9}$$

IN TEXT QUESTIONS

8. A random variable that assume an infinite number of values is called
   (a) Continuous random variable
   (b) Discrete random variable
   (c) Irregular random variable
   (d) Uncertain random variable

9. For the function $f(x)=a+bx,\ 0\le x\le 1$ to be a valid PDF, which of the following statement is correct:
   (a) $a=0.5,\ b=1$
   (b) $a=1,\ b=4$
   (c) $a=1,\ b=-1$
   (d) $a=0,\ b=0$

10. Two random variables *X* and *Y* are distributed according to

$$f_{X,Y}(x,y)=\begin{cases}(x+y) & 0\le x\le 1,\ 0\le y\le 1\\ 0 & \text{otherwise}\end{cases}$$

The probability $P(X+Y\le 1)$ is
   (a) 0.66
   (b) 0.33
   (c) 0.5
   (d) 0.1

11. What are the two important conditions that must be satisfied for $f(x,y)$ to be a legitimate PDF.

12. When do the conditional density function get converted into the marginal density function?
   (a) Only if random variable exhibits statistical dependency.
   (b) Only if random variable exhibits statistical independency

(c) Only if random variable exhibit deviation from its mean value

(d) None of the above.

13. Let U and V be jointly distributed continuous variable and joint PDF is given as

$$f_{U,V}(U,V) = \begin{cases} 6e^{-(2U+3V)} \\ 0 \qquad \text{otherwise} \end{cases}$$

Answer the following

(a) Are U and V independent?

(b) Verify if $E[V|U > 2] = 1/3$

(c) Verify if $P(U > V) = 3/5$

10.5 SUMMARY

Joint probability distribution function refers to the combined probability distribution of more than one random variable. These variables may be discrete or continuous. Marginal probability distribution is obtained by adding probability distribution of one variable keeping the other variable as constant. $P(x, y)$ must satisfy the following conditions in the case of discrete variables to be a valid joint probability mass function-

(d) $0 \le P(x, y) \le 1$

(e) $\sum_x \sum_y P(x, y) = 1$

Respective counterpart is important in the case of continuous random variable. Conditional probability is the probability of happening one event when the other event has already occurred. X and Y are called independent if the joint p.d.f. is the product of the individual p.d.f.'s,

i.e., if $f(x, y) = f_X(x) \cdot f_Y(y)$ for all x, y.

10.6 GLOSSARY

Conditional Probability: a measure of the probability of an event occurring given that another event has already occurred

Independence of Random Variables: if $P_{XY}(x, y) = P_X(x) \cdot P_Y(y) \qquad \forall\ x, y$

Marginal probability Density Function: obtained by integrating the joint PDF of one variable keeping the other constant.

$$E[g(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,ly),\ f(x,y)\,dxdy$$

Expected Value of a PDF:

10.7 ANSWERS TO IN – TEXT QUESTIONS

1. d	8. a
2. b	9. a
3. a	10. b
4. False	11. Refer to example 5
5. a	12. b
6. c	13. a) Yes
7. a	b). Yes
	c) Yes

10.8 SELF – ASSESSMENT QUESTIONS

1. A fair coin is tossed 4 times. Let the random variable X denote the number of leads in the first 3 tosses and let the random variable Y denote the number of leads in the last 3 tosses. Answer the following.

   (a) What is the joint PMF of X and Y.

   (b) What is the probability of 2 or 3 leads appearing in the first three tosses and 1 or 2 leads appear in the last three tosses.

2. Let X and Y be random variables with joint PDF.

$$f_{XY}(x,y) = \begin{cases} \frac{1}{4} & -1 \le x,\ y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

   Find

   (a) $P(X^2 + Y^2 < 1)$

   (b) $P(2X - Y > 0)$

3. Let X and Y be two jointly distributed continuous random variable with joint PDF

$$f_{X,Y}(x,y)=\begin{cases}6xy & 0\le x\le 1,\ 0\le y\le\sqrt{x}\\ 0 & \text{otherwise}\end{cases}$$

(a) Find $f_X(x)$ and $f_Y(y)$

(b) Are X and Y independent?

(c) Find conditional PDF of X given Y

(d) Find $E[X\mid Y=y]$ for $0\le y\le 1$

4. The joint pdf of two random variables X and Y is given by:

$$f(u,v)=\begin{cases}24uv, & 0<u<1, 0<v<1, u+v<1\\ 0, & \text{otherwise}\end{cases}$$

Find $P(U+V)<\frac{1}{2}$.

10.9 REFERENCES

- Devore J. L. (2012). *Probability and statistics for engineering and the sciences* (8th ed.; First Indian reprint 2012). Brooks/Cole Cengage Learning.
- Rice J. A. (2007). *Mathematical statistics and data analysis* (3rd ed.). Thomson/Brooks/Cole.
- Johnson R. A. & Pearson Education. (2017). *Miller & Freund's probability and statistics for engineers* (Ninth edition Global). Pearson Education.
- Miller, I., Miller, M. (2017). J. Freund's Mathematical Statistics with Application, 8th ed., Pearson
- Hogg R. V. Tanis E. A. & Zimmerman D. L. (2021). *Probability and statistical inference* (10th Edition). Pearson.
- James McClave, P. George Benson, Terry Sincich (2017), Statistics for Business and Economcs, Pearson Publication

10.10 SUGGESTED READING

- Webster A. L. (1998). *Applied statistics for business and economics an essentials version* (Third). Irwin/McGraw-Hill.

LESSON 11

CORRELATION AND COVARIANCE

STRUCTURE

11.1 LEARNING OBJECTIVES

After reading this lesson, students will be able to :

1. Difference between covariance and correlation
2. Method of calculating covariance
3. Types of correlation and
4. Methods of calculating correlation

11.2 INTRODUCTION

In the previous units, you must have come across problems dealing with a single variable such as marks, weight, and height of students in a classroom in which you used statistical measures of central tendency such as mean, median, mode, standard deviation etc. All these measures focused on understanding the data set containing individual variables independently. However, in the real world, we have to analyse not only single variable but number of variables at the same time. In such a situation, the basic question that comes in our minds is whether there is any relationship between the two or more variables or not? And if there is a relationship, then what kind of relationship? How can we find out the presence of such relationship between the variables? What is the strength of such a relationship? The objective of this unit is to find the answer to such numerous questions that we come across while dealing with two or more variables simultaneously.

11.3. COVARIANCE

Covariance is one of the statistical measures of the relationship between two variables. In other words, it shows how two variables change simultaneously. Suppose in your classroom there are different students with different height and weight. So, if you want to know whether there is any relationship between the height and weight of students. In other words, whether weight of students varies simultaneously with the height or not. Then in such a case, we can use covariance to understand the relationship between the two variables such as height and weight in the present case.

The formula for covariance is given by:

$$\text{Cov (X, Y)} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N}$$

Where, N= sum of number of observations

$\overline{X}$ = Mean of X

X= value of observation X

$\overline{Y}$= Mean of Y

Y= value of observation Y

Example 1: Find the covariance between the height and weight of the following students.

Height (cm)	65	60	70	55	50
Weight (Kg)	73	82	50	40	55

Solution: Assuming height to be X and weight of students to be Y, we have

X	Y	$(X - \overline{X})$	$(Y - \overline{Y})$	$(X - \overline{X})(Y - \overline{Y})$
65	73	5	13	65
60	82	0	22	0
70	50	10	-10	-100
55	40	-5	-20	100
50	55	-10	-5	50
ΣX= 300	ΣY= 300			115

$\overline{X} = \frac{\Sigma X}{N} = \frac{300}{5} = 60$ and $\overline{Y} = \frac{\Sigma Y}{N} = \frac{300}{5} = 60$

Using Cov (X, Y) = $\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N}$

We have Cov(X, Y)= $\frac{115}{5}$= 23

Thus, the covariance between the height and weight is 23.

IN-TEXT QUESTIONS

1. _____ is the statistical measure of relationship between variables.
2. Find the covariance between the marks of students in English and mathematics in Grade 10.

English	60	40	41	55	67	23	19	70	75
Mathematics	50	60	75	55	87	65	70	45	33

3. Find the covariance between X and Y from the following table.

X	100	150	175	225	250
Y	700	600	500	400	800

11.4. CORRELATION

In earlier example, where we tried to find out the relationship between weight and height of students, and we used covariance to find that there is positive covariance between the two variables. Positive covariance suggested that the variable move in same direction i.e. when there is increase in height of a student, weight also rise simultaneously. However, in this example, we couldn't find through covariance, how much the weight increases with increase in the height.

Therefore, covariance tells us whether there is relationship between two variable or not. However, it fails to determine the strength of such a relationship. In other words, covariance doesn't inform about how closely two variables are related to each other. Thus, in order to determine the strength of relationship between two variables, we use correlation. In other words. If we need to determine how much one variable changes with respect to another variable, we use another measure known as correlation.

Correlation is defined as the degree of association between two variables. In simple words, it explains how far two variables are related to each other. In fact, coefficient of correlation is said to be a measure of covariance between two series of variables.

Correlation is an important statistical measure which helps in determining changes in one variable vis-à-vis another variable. For example, we know law of demand, according to which, quantity demanded is inversely related to the price of a commodity given all other things are constant. Similarly, Keynes physiological law of consumption, which says that if there is an increase in income, it will lead to increase in consumption but by less than the increase in the former. However, if we need to find out how much consumption changes with increase in income, we can again take the help of correlation coefficient to measure this relationship between income and consumption.

However, it is important to distinguish correlation from causation. Correlation simply informs us about the how much one variable varies with respect to changes in another variable. It doesn't necessarily mean causation. It means correlation doesn't not tell anything about the cause-and-effect relationship between two variables, it just only gives an understanding regarding the strength of relationship between the two variables. For example, in the following table, we have information regarding the demand and price of a commodity.

Demand	100	200	300	400	500
Price	60	50	40	30	20

In this case, there is a perfect negative correlation between the demand and price. However, it implies that decrease in price causes demand to rise. This is only explaining inverse relationship between the price and quantity demanded. In order to determine the cause-and-effect relationship, we need to use higher statistical measures such as regression analysis

11.5. TYPES OF CORRELATION

11.5.1. Positive and Negative Correlation

When the coefficient of correlation between the two variables is positive it means that both the variables move in the same direction. In other words, when one variable increases, then the other also increases, though the rate of increase could be different.

For example: The law of supply curve states that there is a one-to-one relationship between the price of a commodity and quantity supplied, given other things are constant. Such a relationship between the price and quantity supplied is positive indicating as there is rise in the price, the quantity supplied by the producer also rise.

Negative correlation between the two variables implies that both move in opposite direction i.e. when one variable increase, there is a decrease in other variable. Such an inverse relationship is found in the law of demand, which states that as the price of a commodity increase, there is fall in the quantity demanded of the commodity.

11.5.2. Linear and Non- linear correlation

When the relationship between the two variables is linear, then it is referred as linear correlation. In case of linear correlation, the amount of change in one variable tends to bear constant ratio to the amount of change in another variable as a result when two variables are plotted in a graph, we get a straight line.

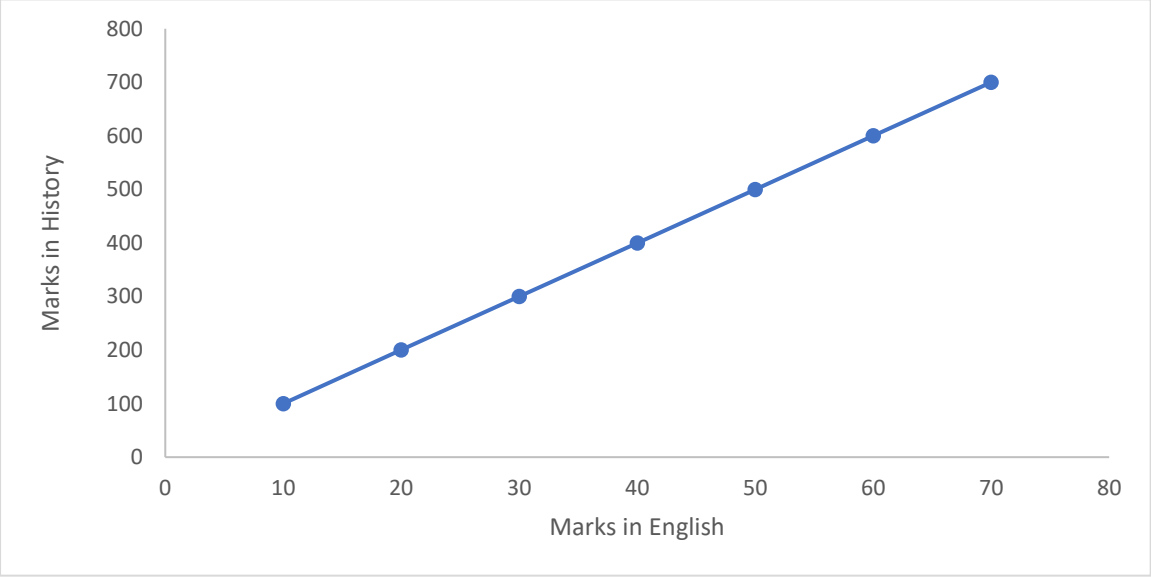


Fig.1: Linear Correlation

On the other hand, when the relationship between the two variables is non-linear, then it is referred as non- linear correlation. In this case, the amount of change in one variable does not bear a constant ratio to the amount of change in other variable. Thus, when we plot two variables in a graph, then we don't get a straight line but a curve.

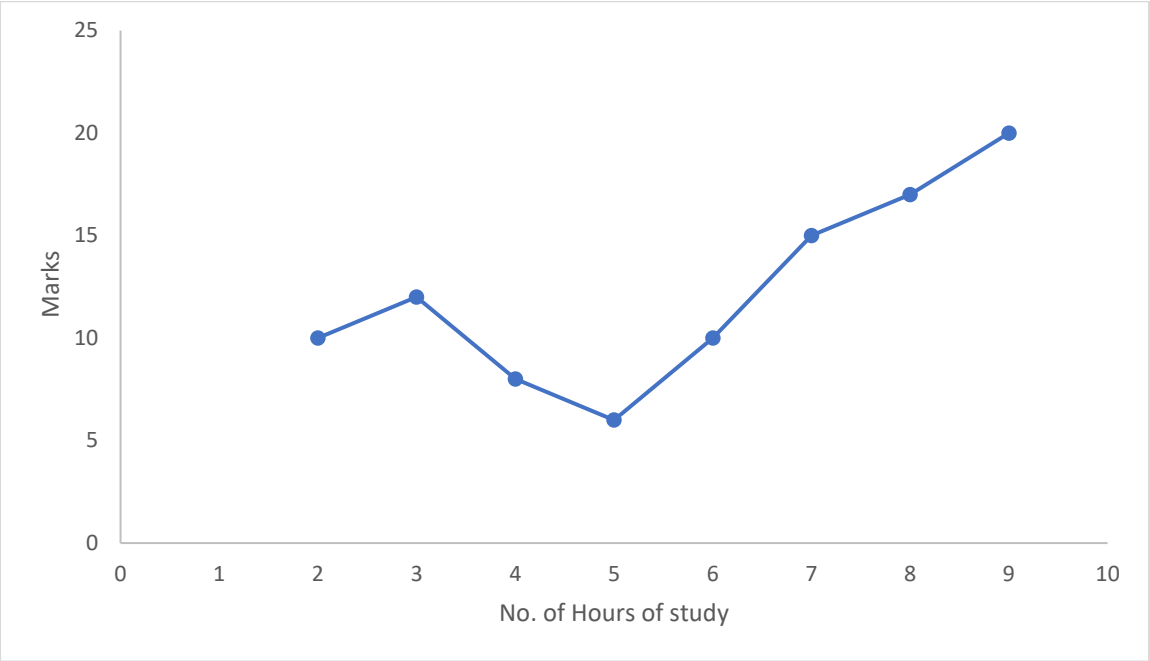


Figure 2: Non-Linear Correlation

11.5.3. Simple and Multiple correlations

When we try to find the relationship between only two variables, then it is a case of simple correlation. In the case of partial or multiple correlations, we are concerned with finding the correlation between two or more variables. For example, when we try to find out the relationship between the marks obtained by the students on a test and the number of hours of study done by the students and his/her IQ. Then such a case is an example of multiple correlation.

11.6. DIFFERENCE BETWEEN CORRELATION AND COVARIANCE

Since now you have studied correlation, you need to have a clear distinction between correlation and covariance. Table 1 elaborates on the same.

Table 1: Difference between correlation and covariance

Correlation	Covariance
It is the measure of strength of relationship between two variables.	It is a measure which shows how two random variables change with respect to each other.
The value of correlation lies between -1 and +1.	The value of covariance lies between -∞ and +∞.
It measures the direction as well as strength of the relationship between the given two variables.	It only indicates the direction of the relationship between the given two variables.
It is free from units of measurement.	It depends on units of measurement.

11.7. METHODS OF CALCULATING CORRELATION

There are three methods of calculating coefficient of correlation:

1) Scatter diagram
2) Karl Pearson's coefficient of correlation
3) Spearman's rank correlation

11.7.1. Scatter diagram

As the name suggests, in this method we will simply put the data into the graph in the form of scatter plot to find out the correlation between the two variables. If the scatter of the plotted points is dense, then the correlation between the two variables is higher. However, if the scatter of the plotted point is spread widely, then the correlations between the two variables is small.

This is one of the simplest methods of ascertaining relationship between two variables as it just requires plotting on graph and visualization.

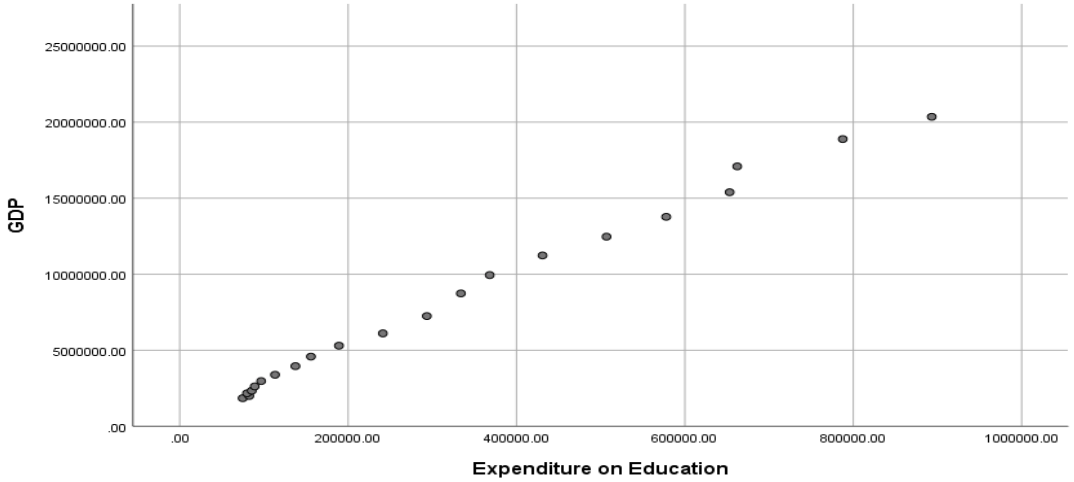


Figure 3: Scatter Plot Diagram showing positive correlation

The above graph plots the GDP at current price and public expenditure on education by the government of India since 1999-2000 to 2019-2020. In this case, the points lie closely on a upward sloping straight line from left to right, thus the correlation between the GDP and public expenditure on education is highly positive.

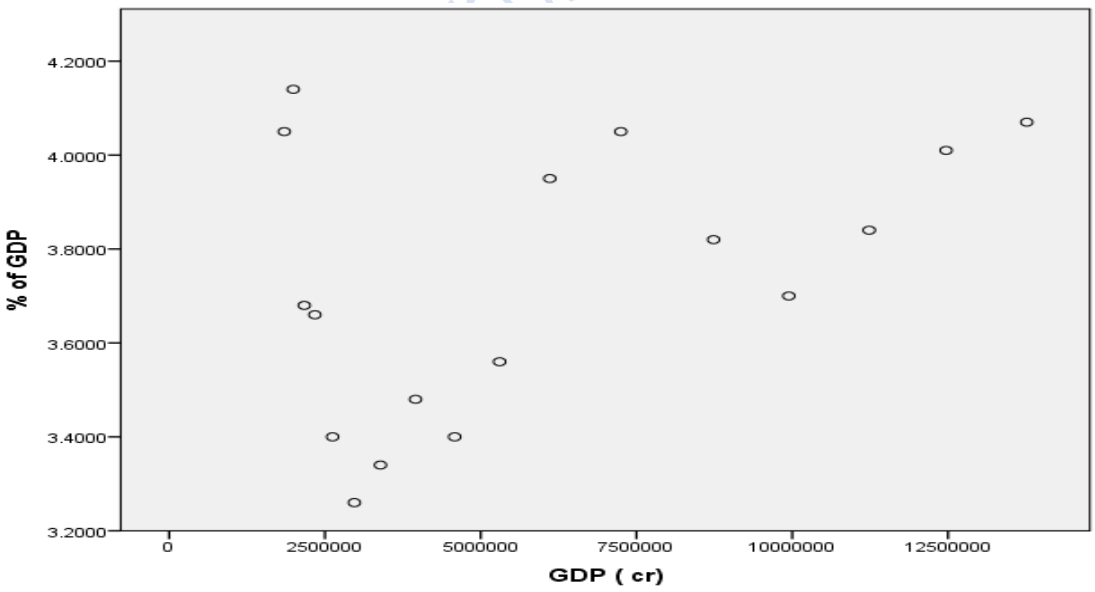


Figure 4: Scatter Plot Diagram showing weak correlation

The above graph plots the GDP and the expenditure on education as the percentage of GDP spent on education in India during 1999-2000 to 2019-20. In this case, the points are widely scattered in the graphs which indicates weak correlation between the GDP and expenditure on education as the percentage of GDP

11.7.2. Karl Pearson's Coefficient of Correlation

It is the mathematical method of calculating coefficient of correlation. The coefficient of correlation in case of Karl Pearson is represented by r. When both variables in a particular data set are normally distributed, it is the best method to use this method. However, extreme values can have an impact on this coefficient, which makes it undesirable when one or both of the variables are not normally distributed because they could exaggerate or weaken the strength of the association.

Assumptions of Karl Pearson's coefficient of correlation:

1) There exits linear relationship between two variables
2) The two variables are normally distributed
3) There is a presence of cause-and-effect relationship between the factors which affect the distribution of the two variables.

r = $\frac{\Sigma xy}{N\sigma x\sigma y}$

where x = (X – $\overline{X}$)

y = (Y – $\overline{Y}$)

N = No of items

σx= Standard deviation of X

σy= Standard deviation of Y

In simpler form, r = $\frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$

Where x = (X – $\overline{X}$)

y= (Y – $\overline{Y}$)

x^2= (X – $\overline{X}$)^2

y^2= (Y – $\overline{Y}$)^2

Example 1: Find the Karl Pearson's coefficient of correlation between the marks in economics and mathematics of the following students.

Economics	70	50	55	70	85	90
Mathematics	35	45	28	33	25	14

Solution: Let marks in economics by X and marks in mathematics be Y.

X	Y	$(X - \overline{X})$ = x	$(Y - \overline{Y})$ = y	$(X - \overline{X})^2$ $=x^2$	$(Y - \overline{Y})^2$= y^2	xy
70	35	0	5	0	25	0
50	45	-20	15	400	225	-300
55	28	-15	-2	225	2	30
70	33	0	3	0	9	0
85	25	15	-5	225	25	-75
90	14	20	-16	400	256	-320
ΣX = 420	ΣY = 180			Σx^2= 1250	Σy^2= 544	Σxy= -665

$$\overline{X} = \frac{\Sigma X}{N} = 420/6 = 70$$

$\overline{Y} = \frac{\Sigma Y}{N}$ = 180/6 = 30

Using, r = $\frac{\Sigma xy}{\sqrt{x^2 \times y^2}}$

We have

r = $\frac{-665}{\sqrt{1250 \times 544}}$

$= \frac{-665}{\sqrt{6,80,000}}$

$= \frac{-665}{824.62}$

= - 0.8064

= - 0.81(approx.)

Therefore, the marks of students in economics and mathematics are inversely correlated to each other as the Karl Pearson's coefficient of correlation is -0.81.

Direct method for finding coefficient correlation

In the previous method of finding correlation coefficient, we have taken deviations of items from mean. However, we can also find correlation without taking any deviation of items from mean by employing following formula:

$r = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$

Example 2: find out the coefficient of correlation from the following data set using direct method

A	1	6	9	3	4	5	8	2	1
B	12	9	5	6	3	7	15	11	9

Solution:

X	Y	X^2	Y^2	XY
1	12	1	144	12
6	9	36	81	54
9	5	81	25	45
3	6	9	36	18
4	3	16	9	12
5	7	25	49	35

8	15	64	225	120
2	11	4	121	22
1	9	1	81	9
$\Sigma X = 39$	$\Sigma Y = 77$	$\Sigma X^{2=} 237$	$\Sigma Y^2 = 771$	$\Sigma XY = 327$

$$r = \frac{327 - (39 \times 37)}{\sqrt{9 \times 237 - (39)^2}\sqrt{9 \times 771 - (77)^2}}$$

$$r = \frac{327 - 3003}{\sqrt{2133 - 1521}\sqrt{6939 - 5929}}$$

$$r = \frac{-2676}{\sqrt{612}\sqrt{1010}}$$

$$r = \frac{-2676}{24{,}74 \times 31.78}$$

$$r = \frac{-2676}{786.24}$$

r = - 3.40

11.7.3. Spearman's Coefficient of Rank Correlation

In the previous method of calculating correlation coefficient, important assumption was made that the variables under the study must be normally distributed so as to yield appropriate results. However, in actual circumstances, we often face a situation where the variables are not normally distributed but skewed. In such a situation, there is a need to use another method which doesn't make such unrealistic assumptions about the distribution of the variables in question. Such one method is Spearman's rank correlation, under which no assumption is to be followed for calculating coefficient of correlation between the two variables.

In Spearman's rank correlation, variables are ranked, and the calculations are made on the basis of ranks not the original observations in order to determine the coefficient of correlation.

The formula for spearman's rank correlation is

$$R = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

Where, R= rank corelation

D^2 = square of difference between the ranks

N = no of observation

Characteristics of spearman's rank correlation:

1) The difference in rank between any two variables will have sum equal to zero.
2) Spearman's rank correlation is non-parametric meaning it is not based on any assumption about the distribution of the variable.
3) The Karl Pearson coefficient of correlation between the rankings is the same as the spearman's rank correlation.

The above method of rank correlation is useful when there is no tie between the ranks of an observation. However, in many cases, we come across variables which are similar in size or other characteristics. In such situations, it becomes important to give equal ranks to such similar observation. Thus, in such cases, where observations in a variable set have equal ranks, the above method of rank correlation needs to be modified so as to be appropriate in cases of equal ranks.

Thus, when equal ranks are given, we will have modified version of spearman's rank correlation, which is =

$$R = 1 - \frac{6\Sigma D^2 - \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\ldots\ldots\ldots}{N(N^2-1)}$$

Where, R= rank corelation

D^2 = square of difference between the ranks

m = no. of items whose ranks are same.

N = no of observation

Example 3: Find out the spearman's rank correlation between the ranks of students in two section of class XI.

Student	1	2	3	4	5	6	7
Section A	4	5	7	1	3	2	6
Section B	1	3	2	7	5	6	4

Solution:

Students	Ranks in Section A (R_1)	Ranks in Section B(R_2)	(R_1-R_2)= D	(R_1-R_2) = D^2
1	4	1	3	9
2	5	3	2	4
3	7	2	5	25
4	1	7	-6	36
5	3	5	-2	4
6	2	6	-4	16
7	6	4	2	4
				ΣD^2= 98

Using,

$$R = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

Where, $\Sigma D^2 = 98$

$$R = 1 - \frac{6 \times 98}{7(7^2-1)}$$

$$R = 1 - \frac{588}{336}$$

R= 1- 1.75

R = -0.75

Example 4: Calculate coefficient of correlation between the rank of participants in a dance competition.

Participants	1	2	3	4	5
Score of Judge 1	55	70	80	70	75
Score of Judge 2	70	80	90	50	60

Solution: In the above case, we are given score of two judges. Since ranks are not given, we will give rank to the participants on the basis of scores of two judges.

Participants	Score of Judge 1	R_1	Score of Judge 1	R_2	R_1-R_2	$(R_1-R_2)^2 = D^2$
1	55	5	70	3	2	4
2	70	3.5	80	2	1.5	2.25
3	80	1	90	1	0	0
4	70	3.5	50	5	1.5	2.25
5	75	2	60	4	-2	4
						ΣD^2= 12.50

Using, $R = 1- \frac{6\Sigma D^2 - \frac{1}{12}(m_1^3 - m_1)}{N(N^2-1)}$

Where m = 2 as there are only 2 observation whose ranks are repeated.

$R = 1- \frac{6\times 12.5 - \frac{6}{12}}{5 \times 24}$

$R = 1- \frac{75-0.5}{120}$

$R = 1- \frac{74.5}{120}$

R = 1- 0.60

R= 0.40

IN-TEXT QUESTIONS

4. Scatter plot is simplest method of determining correlation. True/False
5. Karl Pearson's is a one of the methods of correlation. True/False
6. Scatter plot method involves plotting variables in graph. True/ False
7. Spearman's Rank correlation cannot be used in case of common ranks. True/False
8. Which of the following is the method of measuring correlation?

   A) Spearman's Rank method B) Standard Deviation

   C) Covariance D) Mode

9. Find out the correlation between X and Y using Karl Pearson's coefficient of correlation.

X	10	20	30	40	50	60	70	80
Y	100	150	100	250	300	200	200	300

10. In the following table ranking of 10 participants by two judges in a drawing competition is given.

Judge 1	2	5	10	1	9	8	4	3	7	6
Judge 2	2	10	7	4	9	5	9	1	8	3

11. Calculate the rank coefficient between the marginal utilities of two goods received by the 10 individuals.

Individuals	A	B	C	D	E	F	G	H	I	J
Marginal utility of Good X	70	50	60	60	77	80	90	15	25	45
Marginal Utility of Good Y	60	90	55	40	59	65	70	85	73	50

11.8 GLOSSARY

Covariance: it is the measure of relationship between two or more variables

Correlation: It refers to the degree of relationship between two or more variables.

11.9 SUMMARY

In this chapter, we dealt with related but different statistical measure of determining relationship between two or more variables. on the one hand, covariance informs us about how the variables move together or not. However, it doesn't inform anything about the amount of association between the variables. On the other hand, correlation shows the relationship as well as degree of relationship between two or more variables. Correlation can be further determined using following methods:

A) Scatter Plot method

B) Karl Pearson's Coefficient of Correlation

C) Spearman's Rank Correlation

11.10 ANSWER TO IN-TEXT QUESTIONS

1. Covariance
2. -135.67
3. – 600
4. True
5. True
6. True
7. False
8. Spearman's rank method
9. 0.73
10. 0.45
11. -0.22

11.11 SELF-ASSESSMENT QUESTIONS

Q.1. What is the difference between covariance and correlation?

Q.2. Discuss briefly the difference between different methods of calculating coefficient of correlation.

Q.3. Plot the following data on a scatter plot and comment about the correlation between the two variables.

X	10	5	1	3	2	7
Y	3	4	10	9	5	2

Q.4. Using scatter plot find out whether there is any correlation between the number of hours individuals exercise and their respective weight.

No of Exercise Hours	1	2	3	4	5	6
Weight	80	65	55	60	50	60

Q5. Find out the correlation between the following variables.

X	1	2	3	5	6
Z	2	5	6	1	2

11.12 REFERENCES

- Devore, J. (2012) Probability and Statistics for Engineers, 8th ed.. Cengage Learning
- John A. Rice (2007), Mathematical Statistics and Data Analysis, 3rd ed. Thomson Brooks/Cole
- Miller, I,Miller, M (2017, J. Freund's Mathematical Statistics with Applications. 8th ed Pearson
- Larsen, R, Marx (2011). An Introduction to Mathematical Statistics inference, 10th Edition, Pearson

11.13 SUGGESTED READINGS

- Godfrey, K. (1980). Correlation methods. *Automatica*, *16*(5), 527–534. https://doi.org/10.1016/0005-1098(80)90076-x
- Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001
- Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W., & van Diepen, M. (2021). Conducting correlation analysis: important limitations and pitfalls. *Clinical Kidney Journal*, *14*(11), 2332–2337. https://doi.org/10.1093/ckj/sfab085
- Correction to Lancet Respir Med 2021; published online April 9. https://doi. org/10.1016/S2213-2600(21)00160-0. (2021). *The Lancet Respiratory Medicine*, *9*(6), e55. https://doi.org/10.1016/s2213-2600(21)00181-8

978-93-95774-84-0

Department of Distance and Continuing Education
University of Delhi