IIT MADRAS **BS DEGREE**

STATISTICS

Descriptive Statistics

by
Prashant Sharma

About the author

Prashant Sharma

Prashant Sharma holds a Master's degree in Statistics with an outstanding academic record from the prestigious University of Delhi. He possesses a solid academic foundation in the subject that serves as the bedrock of his writing. He is constantly driven by a keen interest in expanding his knowledge and exploring the vast realm of statistics.

As an instructor, Prashant has conducted live sessions for students, delivering statistical concepts in a clear and engaging manner. He actively seeks opportunities to enhance his teaching skills and stay up-to-date with the latest statistical developments.

By leveraging his expertise and leveraging online teaching platforms, Prashant strives to empower students of IITM BS to grasp statistical concepts and apply them to their respective fields. Through his writing and sessions, Prashant continues to inspire students and making statistics an approachable and captivating subject for all.

Contents

1 Statistics **7**
- 1.1 Population and Sample . . . 7
- 1.2 Major branches of statistics . . . 7
- 1.3 Purpose of statistical analysis . . . 8

2 Data **9**
- 2.1 Unstructured and Structured Data . . . 9
  - 2.1.1 Variables and Cases . . . 11
- 2.2 Classification of Data . . . 11
  - 2.2.1 Categorical Data and Numerical Data . . . 11
    - 2.2.1.1 Categorical Data . . . 11
    - 2.2.1.2 Numerical Data . . . 12
  - 2.2.2 Time-series and cross-sectional Data . . . 12
  - 2.2.3 Scales of measurement . . . 12
    - 2.2.3.1 Nominal scale of measurement . . . 12
    - 2.2.3.2 Ordinal scale of measurement . . . 13
    - 2.2.3.3 Interval scale of measurement . . . 13
    - 2.2.3.4 Ratio scale of measurement . . . 13

3 Describing categorical data: Frequency distribution **17**
- 3.1 Frequency Distribution . . . 17
- 3.2 Relative frequency . . . 18
- 3.3 Charts of categorical data . . . 18
  - 3.3.1 Pie Chart . . . 18
  - 3.3.2 Bar Chart . . . 19
  - 3.3.3 Pareto Chart . . . 20
- 3.4 The Area Principle . . . 24
  - 3.4.1 Misleading graphs: violating area principle . . . 24
  - 3.4.2 Misleading graphs: truncated graphs . . . 26
  - 3.4.3 Manipulated y-axis . . . 27
  - 3.4.4 Indicating a y-axis break . . . 28
  - 3.4.5 Round-off errors . . . 28
- 3.5 Summarizing Categorical Data . . . 29
  - 3.5.1 Mode . . . 29
    - 3.5.1.1 Bimodal and Multimodal data . . . 31
  - 3.5.2 Median . . . 32

4 Describing Numerical data **35**
- 4.1 Types of variables . . . 35
  - 4.1.1 Discrete Variable . . . 35
  - 4.1.2 Continuous Variable . . . 35

4.2 Organizing Numerical Data . . . 35
4.2.1 Organizing Discrete Data (single value) . . . 36
4.2.2 Organizing Continuous Data . . . 37
4.2.2.1 Terminology . . . 37
4.3 Stem-and-leaf diagram . . . 38
4.3.1 Steps to construct a stemplot . . . 38
4.4 Descriptive Measures . . . 39
4.4.1 Measures of Central Tendency . . . 39
4.4.1.1 Mean . . . 39
4.4.1.2 Median . . . 43
4.4.1.3 Mode . . . 44
4.4.2 Measures of Dispersion . . . 45
4.4.2.1 Range . . . 46
4.4.2.2 Variance . . . 46
4.4.2.3 Standard Deviation . . . 49
4.5 Percentiles . . . 53
4.5.1 Computing Percentiles . . . 53
4.6 Quartiles . . . 54
4.7 Five Number Summary . . . 55
4.8 Interquartile Range (IQR) . . . 55

5 Association between two variables **56**
5.1 Association Between Two Categorical Variables . . . 56
5.1.1 Stacked Bar Chart . . . 61
5.1.2 100% Stacked Bar Chart . . . 61
5.2 Association Between Two Numerical Variables . . . 64
5.2.1 Scatter Plot . . . 64
5.2.1.1 Describing Association . . . 67
5.2.2 Measures of association between two numerical variables . . . 69
5.2.2.1 Covariance . . . 69
5.2.2.2 Correlation . . . 71
5.2.2.3 Fitting a line . . . 72
5.3 Association Between Categorical and Numerical Variables . . . 72
5.3.1 Point Bi-serial Correlation Coefficient . . . 72

6 Basic Principle of Counting **76**
6.1 Introduction . . . 76
6.1.1 Addition rule of counting . . . 76
6.1.2 Multiplication rule of counting . . . 76
6.1.2.1 Solved Examples: . . . 77
6.1.3 Unsolved Problems: . . . 79

7 Factorial **80**
7.1 Definition . . . 80
7.1.0.1 Simplifying expressions: . . . 80
7.1.0.2 Unsolved Problems: . . . 81

8 Permutation **82**
8.1 Definition . . . 82
8.1.0.1 Solved Examples: . . . 82
8.2 Permutation formula . . . 84
8.2.1 When repetition is not allowed. . . . 84
8.2.1.1 Solved examples by using permutation formula: . . . 84
8.2.1.2 Example: Application . . . 86
8.3 Permutation formula . . . 87
8.3.1 When repetition is allowed. . . . 87
8.3.1.1 Solved examples: . . . 88
8.4 Permutation formula . . . 89
8.4.1 Rearranging letters . . . 89
8.5 Circular Permutation . . . 90
8.5.1 Clockwise and anticlockwise are different . . . 90
8.5.2 Clockwise and anticlockwise are same . . . 91
8.5.2.1 Examples of calculating n and r. . . . 92
8.6 Unsolved Problems: . . . 94

9 Combination **95**
9.1 Definition . . . 95
9.1.0.1 Solved Examples: . . . 95
9.1.1 Drawing lines in a circle . . . 97
9.1.1.1 Some more solved examples on permutation and combination: 98
9.2 Unsolved Problems: . . . 100

Chapter 1

1 Statistics

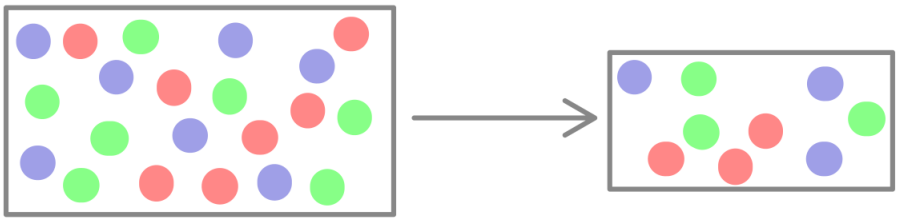
Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

1.1 Population and Sample

Population
The total collection of all the elements that we are interested in is called a population.

Sample
A subgroup of the population that will be studied in detail is called a sample.

We can understand about sample and population from the following picture:



Example:
Suppose a survey is conducted to know the prices of all houses in Tamil Nadu and 1000 houses were randomly selected from the urban areas of Tamil Nadu for this study. It is concluded that price of a house per square feet is roughly 5680 Rs. Then, the sample consists of the selected 1000 houses from the urban areas of Tamil Nadu and the population consists of all houses in Tamil Nadu.

1.2 Major branches of statistics

1. Descriptive Statistics Statistics
The part of statistics concerned with the description and summarization of data is called descriptive statistics.

- Summarization of data means numerical/graphical summary of data or to describe the main points of data.
- A descriptive study may be performed either on a sample or on a population data.

2. Inferential Statistics Statistics

The part of statistics concerned with drawing conclusions from the data is called inferential statistics.

1.3 Purpose of statistical analysis

- If the purpose of the analysis is to examine and explore information about the collected data only, then the study is descriptive.

  For Example: A class of 50 students gave an exam (of 100 marks) and the average marks of the class is calculated as 65. This type of study is called descriptive statistics because here we are just summarizing the data (calculating the average marks of whole class).

- If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions/inferences about the population, the study is inferential.

  For Example: A teacher wants to know the average marks of all students in the school. Since there is a large number of students in the school, the teacher collects a sample of students from the school and calculates the average marks of the selected students which is, say, 60 marks. Then, teacher made the conclusion (using statistical techniques) that average marks of all students in the school is 60. This type of study is called inferential statistics because here we are making conclusion about population based on the sample data.

Chapter 2

2 Data

Definition
Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

• **Purpose to collect the data :**
Generally, we collect the data when we are interested to understand the characteristics or attributes of some group or groups of people, places, things, or events.
For Example:

(1) To know about temperatures in a particular month in Chennai, India.

(2) To know about the marks obtained by students in their Class X.

2.1 Unstructured and Structured Data

Unstructured Data
Unstructured data is a dataset that is not organized in a predefined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. Also, unstructured data requires more work to process and understand.
For Example: You-tube comments, Image files, Social-media posts, lyrics of a song etc.

When data are scattered with no structure, i.e., not in any standard format, the information is of very little use.

Structured Data
Structured data is a standardized format for providing information about a dataset and it is clearly defined and searchable, as for the information in a dataset to be useful, we must know the context of the numbers and text it holds. Also, structured data is easy to analyze and understand. Hence, we need to organize the data.
Let's consider the following two examples:

(1) Dataset of students:

Name	Gender	Date of Birth	Marks in class 10^{th}	Board
Anjali	F	17 Feb, 2003	484	State Board
Pradeep	M	3 June, 2002	514	ICSE
Divya	F	22 Mar, 2003	397	State Board
Sarita	F	19 May, 2002	533	ICSE
Harsha	M	4 March, 2002	436	CBSE
Bhavana	F	7 Apr, 2003	526	State Board
Rohit	M	4 March, 2002	378	CBSE
Vikash	M	11 Oct, 2001	526	CBSE

Table 1: Student dataset

The student dataset shown in Table 1 can be considered as structured data because this data is in a tabular form and provides the information about Gender, Date of Birth, Marks in 10th class and Board of the students. Also, this data is easy to analyze and understand as we can easily get information about any student e.g. Anjali has scored 484 marks in class 10th of State board, Pradeep is Male and have date of birth as 3rd June, 2002 etc.

(2) Dataset of fertilizers:

Fertilizers	Types of Fertilizers	Area of fields (In acres)	Types of Crops	Amount of fertilizers (In Kg)
Nitrogen	Inorganic	1	Rice	200
Phosphorus	Inorganic	2	Wheat	400
Manure	Organic	1.5	Potato	300
Compost	Organic	1.3	Rice	260
Potassium	Inorganic	1.6	Pulse	320

Table 2 : Fertilizers dataset

Fertilizers dataset shown in table 2 can also be considered as structured data because this data is in a tabular form and provides the information about fertilizers. Also, this data is easy to analyze and understand as we can easily get information e.g. Potassium is an inorganic fertilizer and can be used for pulse in the amount of 320 Kg etc.

2.1.1 Variables and Cases

Case (observation) : A case/observation is a unit for which data is collected. Cases should uniquely identify each row in the dataset.

Variable : A variable is a characteristic or attribute that varies across all units. Intuitively, a variable is that "varies".

For Example:

In the table **1** of student dataset, each student, i.e., "Anjali, Pradeep, Divya etc." are cases as data is collected for every student and all the names uniquely identify each row in the dataset.

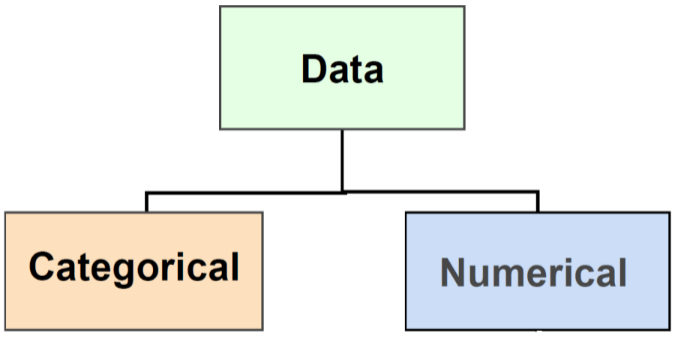
And, variables are "Name, Gender, Date of Birth, Board etc., as their values keeps on varying."

Note: The student dataset is in tabular form. If we want to organise a data in a tabular form, then following two points should take into consideration:

- Rows represent cases: For each case, same attribute is recorded.
- Columns represent variables: For each variable, same type of value for each case is recorded.

2.2 Classification of Data

Data is broadly classified into two categories; categorical data and numerical data.



2.2.1 Categorical Data and Numerical Data

2.2.1.1 Categorical Data

Categorical data are also called qualitative variables and it identifies the group membership. Also, we cannot perform any meaningful mathematical operations on it.

In the student dataset which is illustrated in Table **1**, Gender is a categorical variable because it has two categories as F and M. We can classify any observation into one of these two categories.

Also, Board is a categorical variable since it has three categories as State Board, ICSE and CBSE and any observation can be categorized into one of these three groups.

2.2.1.2 Numerical Data

Numerical data are also called quantitative variables. It describes the numerical properties of the data, i.e., we can perform mathematical operations on the data.

In the student dataset of table **1**, Marks is a numerical variable because we can describe the numerical properties of data as marks of Rohit is 378, marks of Pradeep is 514 or marks of Bhavana is more than marks of Harsha etc.

• **Measurement units**

Scale defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in centimeters, etc.
Also, the data that make up a numerical variable in a data table must share a common unit.

2.2.2 Time-series and cross-sectional Data

- If the data is recorded over a period of time, then it is called time-series data. Also, graph of a time series showing values in chronological order is known as Time-plot.
  Example:
  The data collected to observe the temperature in Delhi for seven different days is a time-series data. Because, data is recorded only for one place (i.e. Delhi) and it is recorded over a period of time (i.e. seven different days).

- If the data is observed at the same time, then it is called cross-sectional data.
  Example:
  The data collected to observe the temperature of Delhi, Chennai, Jaipur and Bhopal on a particular day is a cross-sectional data. Because, data is recorded at the same time and it is observed for several places.

2.2.3 Scales of measurement

We have four scales of measurement called nominal, ordinal, interval and ratio scale. Data collection requires any one of the scales of measurement.

2.2.3.1 Nominal scale of measurement

When the data for a variable consist of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a nominal scale.
Example: Name, Board, Gender, Blood group etc.

Note:

- Sometimes nominal variables might be numerically coded like we might code men as 1 and women as 2 or code men as 3 and women as 1.

- There is no ordering in the variable.

- In short " Nominal scale is just categories or labels which does not contain any order."

2.2.3.2 Ordinal scale of measurement

When data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered an ordinal scale.

Example:

Each customer who visits a restaurant provides a service rating of excellent, good, or poor. Here, the data obtained are the labels as excellent, good, or poor, i.e., the data have the properties of nominal data. Also, the data can be ranked/ordered, with respect to the service quality.

Note:

- We can code an ordinal scale of measurement, as bad can be coded as 1, good can be coded as 2 and excellent can be coded as 3. There is an order in $1, 2, 3$ but one thing need to understand is the distance between bad and good need not be same as the distance between good and excellent. It is just an order.
  As we know excellent is better than good, but we cannot say that the difference between good and excellent is the same as the difference between good and bad. Thus, we have just an order.

- In short " Ordinal scale is just categories or labels which contain an order."

2.2.3.3 Interval scale of measurement

If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is interval scale.

Note:

- Data with interval scale of measurement are always numeric and we can find out the difference between any two values.

- Ratios of values have no meaning here because the value of zero is arbitrary.

Example:

Consider an AC room where temperature is set at 20°C and the temperature outside the room is 40°C. It is correct to say that the difference in temperature is 20°C, but it is incorrect to say that the outdoor is twice as hot as indoor.

Also, temperature in degrees Fahrenheit or degrees centigrade has an interval scale of measurement, because it has no absolute zero. In the Celsius scale, 0 and 100 are set to be as the freezing point and the boiling point whereas, in Fahrenheit it is 32 and 212.

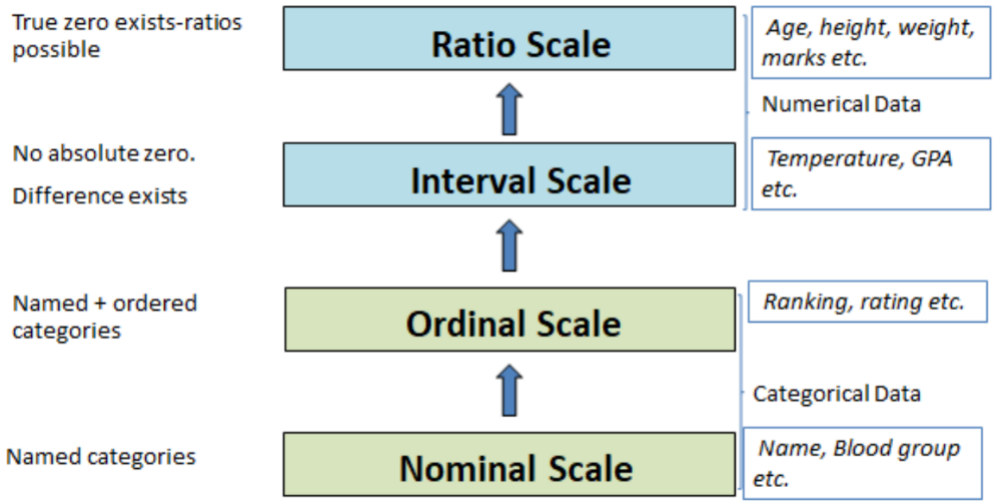
2.2.3.4 Ratio scale of measurement

If the data have all the properties of interval data and the ratio of two values is meaningful, then the scale of measurement is ratio scale.

Ratio scale of measurement has absolute zero property which is the key difference between

ratio and interval scale.
Example: Height (in cm), Weight (in kg) and Marks, etc. All such types of data like height, weight and marks can be added, subtracted and multiplied or divided as it all have absolute zero property.

A summary about all scales of measurement can be described as follows :



Unsolved Problems

(1) An analyst wants to conduct a survey for testing the maintenance of hospitals in a particular district in Bihar, for which he selects 25 hospitals randomly from that district. Identify the sample and population. [2 Marks]

(a) The population is all the hospitals in Bihar and the sample is all the hospitals in the district.

(b) The population is all the hospitals in Bihar and the sample is 25 selected hospitals in Bihar.

(c) The population is all hospitals in the district of Bihar and the sample is 25 selected hospitals in the district.

(d) None of the above

Answer: c

(2) In the 2011 Cricket ODI World Cup quarter-final match between India and Australia, a media organization estimated that Australia would beat India by 50 runs if Australia bats first, based on the information of matches played between the two teams previously. Which branch of statistics does the above analysis belong to?

Answer: Inferential Statistics

(3) Values of temperature and humidity of a room are measured for 24 hours at a regular time interval of 30 minutes. Based on this information, choose the correct option:

(a) It is a cross-sectional data.

(b) It is time-series data.

Answer: b

(4) What kind of data is "Social media posts"?

(a) Unstructured data

(b) Structured data

Answer: a

(5) What kind of variable is the qualification of a candidate sitting for a job interview?

(a) Numerical/ Quantitative

(b) Categorical/ Qualitative

(c) Numerical and discrete

(d) Numerical and continuous

Answer : b

(6) If addition, subtraction can be performed on a variable, then the scale(s) of measurement of the variable could be:

(a) Ordinal

(b) Ratio

(c) Interval

(d) Nominal

Answer : b, c

(7) Which of the following variable(s) have nominal scale of measurement?

(a) Education qualification of a person.

(b) Hair color

(c) Brand name of mobile phone

(d) Number plate of cars

Answer: b, c, d

Chapter 3

3 Describing categorical data: Frequency distribution

3.1 Frequency Distribution

A frequency distribution of qualitative data is a listing of the distinct values and their frequencies.
Each row of a frequency table lists a category along with the number of cases in this category.

Example: Let's construct a frequency table for the following data.

(1) A, A, B, C, A, D, A, B, D, C

Category	Tally mark	Frequency
A	𝍸	4
B	𝍸	2
C	𝍸	2
D	𝍸	2
Total		10

(2) A, A, B, C, A, D, A, B, D, C, A, B, C, D, A

Category	Tally mark	Frequency
A	𝍸	6
B	𝍸	3
C	𝍸	3
D	𝍸	3
Total		15

(3) A, B, B, C, A, D, B, B, D, C, A, B, C, D, B

Category	Tally mark	Frequency
A	𝍸	3
B	𝍸	6
C	𝍸	3
D	𝍸	3
Total		15

(4) A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D

Category	Tally mark	Frequency
A	𝍸	6
B	𝍷	3
C	𝍸	4
D	𝍸	5
Total		18

3.2 Relative frequency

The ratio of the frequency to the total number of observations is called relative frequency.

Note: Relative frequency plays an important role for comparing two data sets because relative frequencies always fall between 0 and 1, they provide a standard for comparison.

Examples: Let us find the relative frequencies for the following data.

(1) A, A, B, C, A, D, A, B, D, C

Category	Frequency	Relative Frequency
A	4	0.4
B	2	0.2
C	2	0.2
D	2	0.2
Total	10	1

(2) A, A, B, C, A, D, A, B, D, C, A, B, C, D, A

Category	Frequency	Relative Frequency
A	6	0.4
B	3	0.2
C	3	0.2
D	3	0.2
Total	15	1

3.3 Charts of categorical data

The two most common displays of a categorical variable are a bar chart and a pie chart.

3.3.1 Pie Chart

A pie chart is a circle divided into pieces proportional to the relative frequencies of the qualitative data and it is used to show the proportions of a categorical variable. And, a pie chart is a good way to show that one category makes up more than half of the total.

Example: Consider the frequency table of the dataset A, A, B, C, A, D, A, B, D, C.

Category	Frequency	Relative Frequency
A	4	0.4
B	2	0.2
C	2	0.2
D	2	0.2
Total	10	1

Table 3.1

Figure 3.1 is the pie chart representation of the dataset in Table 3.1:

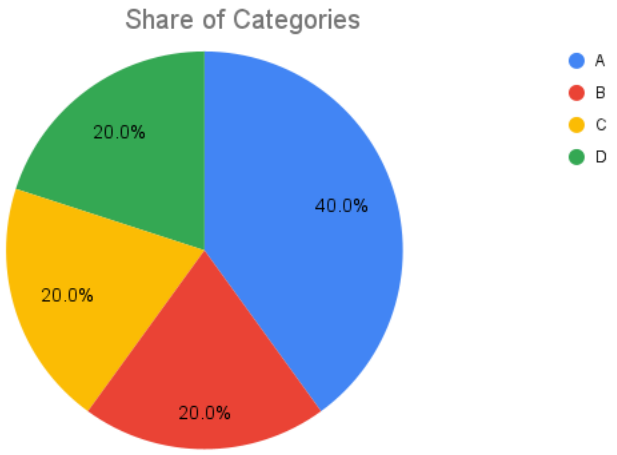


Figure 3.1: Pie chart representation

As pie chart gives us the share of a pie, share of category A is 40%, category B is 20%, category C is 20% and category D is 20%.

3.3.2 Bar Chart

A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (or frequencies or percents) of those values on a vertical axis. The frequency/relative frequency of each distinct value is represented by a vertical bar whose height is equal to the frequency/relative frequency of that value. The bars should be positioned so that they do not touch each other.
Bar chart is most appropriate to represent the count of a particular category and it can be oriented either horizontally or vertically.

Example: A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D

Category	Frequency	Relative frequency
A	6	0.33
B	3	0.17
C	4	0.22
D	5	0.28
Total	18	1

Table 3.2

Figure 3.2 represents the bar chart of the dataset in Table 3.2 as follows:

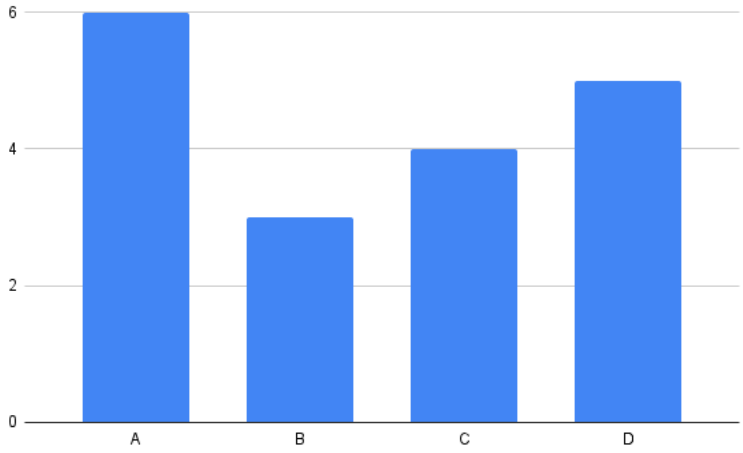


Figure 3.2: Bar chart representation

3.3.3 Pareto Chart

When the categories in a bar chart are sorted by frequency, the bar chart is sometimes called a Pareto chart. Pareto charts are popular in quality control to identify problems in a business process.
Example: A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D

Category	Frequency	Relative frequency
A	6	0.33
B	3	0.17
C	4	0.22
D	5	0.28
Total	18	1

Table 3.3

Figure 3.3 is the pareto chart representation of the dataset in Table 3.3 as follows:

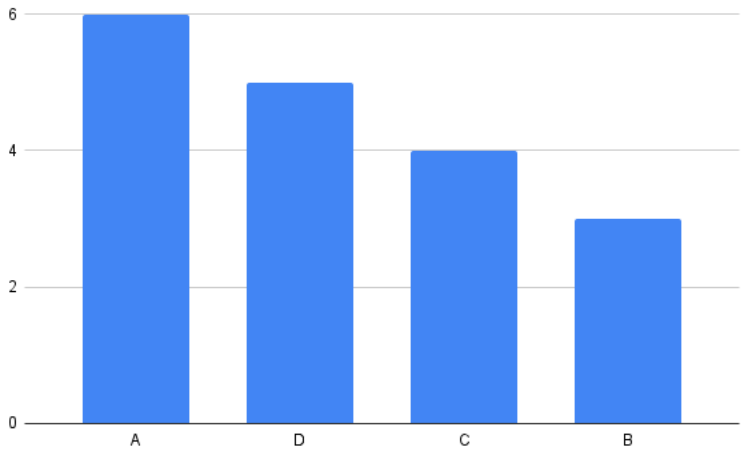


Figure 3.3: Pareto chart representation

Note: If the categorical variable is ordinal, then the bar chart must preserve the ordering.
For example:
The T-shirt sizes L, M, M, S, L, S, S, M, L, M, M, S, S, L, M, S, M, S, L, M of twenty students is listed in Table 3.4:

Size	Frequency	Relative frequency
Small	7	0.35
Medium	8	0.40
Large	5	0.25
Total	20	1

Table 3.4

Dataset of Table 3.4 is ordinal. So, we have preserved the order of the data.
And, bar chart representation for the dataset of Table 3.4 is given as follows:

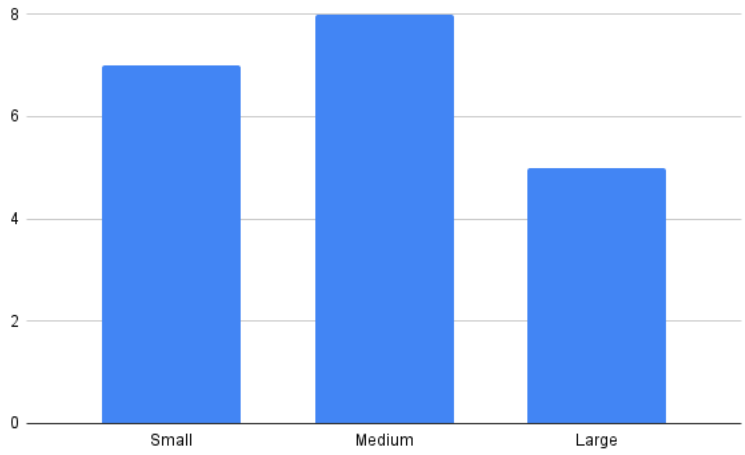


Figure 3.4: Bar chart of Ordinal data

Purpose of using charts

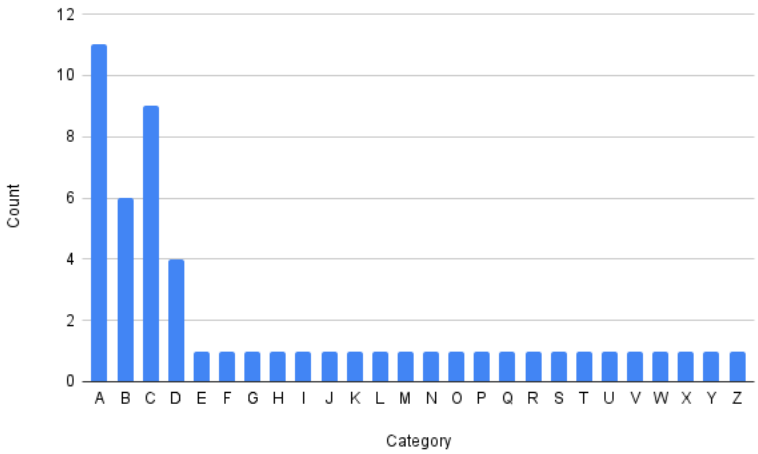
(1) Pie charts are best to use when we are trying to compare parts of a whole.

(2) Bar graphs are used to compare things between different groups.

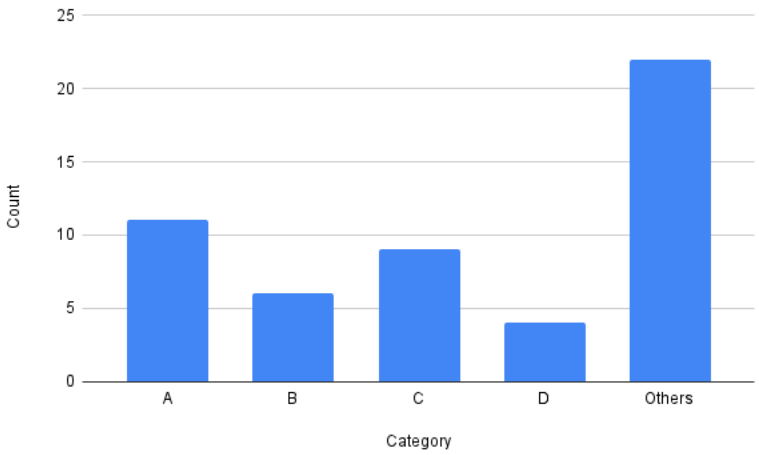
Many Categories:

A bar chart or pie chart with too many categories might conceal the more important categories. In some case, grouping other categories together might be done.

Now, let's consider the following bar chart with too many categories:



Now, we can do grouping of other categories together as follows:



Grouping other categories together in a major category conveys two important things.

(1) We are not excluding any data.

(2) We have a significant number that comes from smaller categories.

3.4 The Area Principle

The area principle says that the area occupied by a part of the graph should correspond to the amount of data it represents.
Display of data must obey the rule of area principle and violations of the area principle are a common way to mislead with statistics.

3.4.1 Misleading graphs: violating area principle

(1) Decorated graphs: Sometimes charts are decorated to attract attention which often violate the area principle.
For Example: Figure 3.5 is an example of decorated graph:

Figure 3.5: Decorated graph

Figure 3.5 gives us the total wine exports in UK, Canada, Japan and Italy. But, there is no baseline and the chart shows bottles on top of labeled boxes of various sizes and shapes.

Now, Figure 3.6 represents the chart which is not decorated:

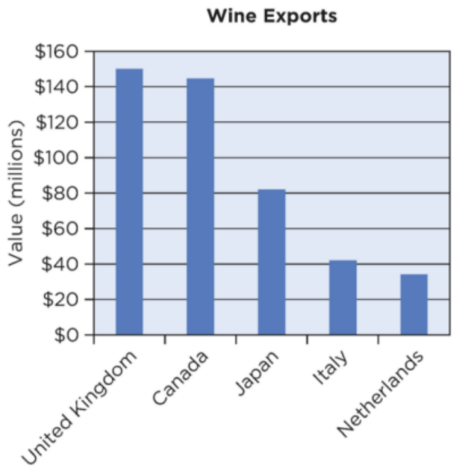


Figure 3.6

We have labeled each one of the categories. It is accurate and it has a baseline. This chart is actually consistent and the width of the bars for each countries are equal. Also, the area occupied by the graph is proportional to the data that is being presented.

(2) Violation of area principle in a pie chart

Figure 3.7 represents the pie chart of the sales distribution of mobile phones of different company.

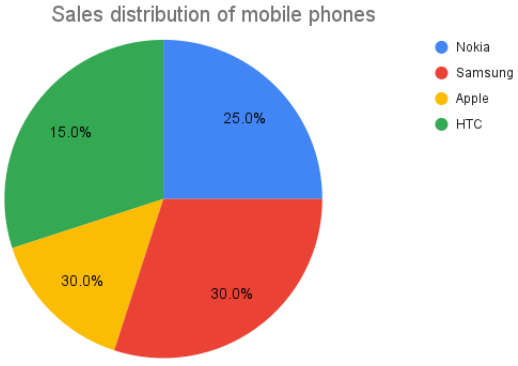


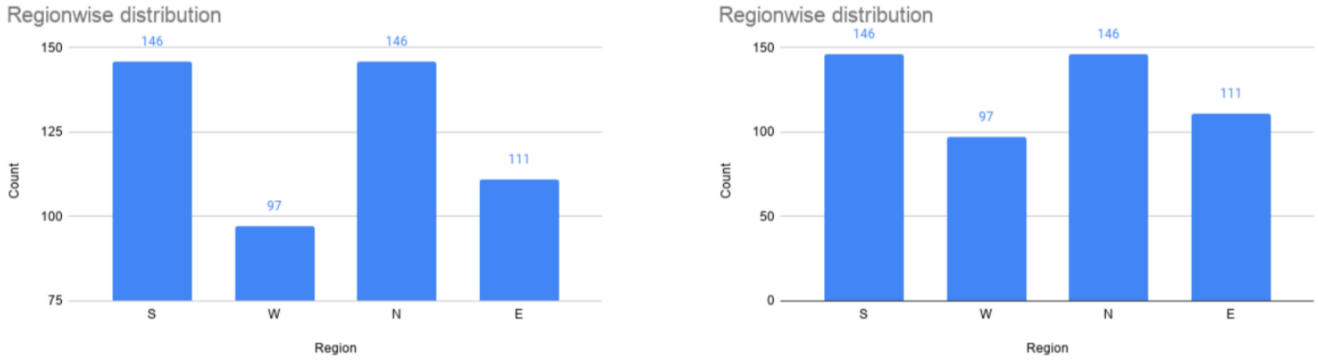
Figure 3.7

The pie chart of the Figure 3.7 is violating the area principle as areas occupied by sales distribution of HTC and Apple do not correspond to the amount of data it represent.

3.4.2 Misleading graphs: truncated graphs

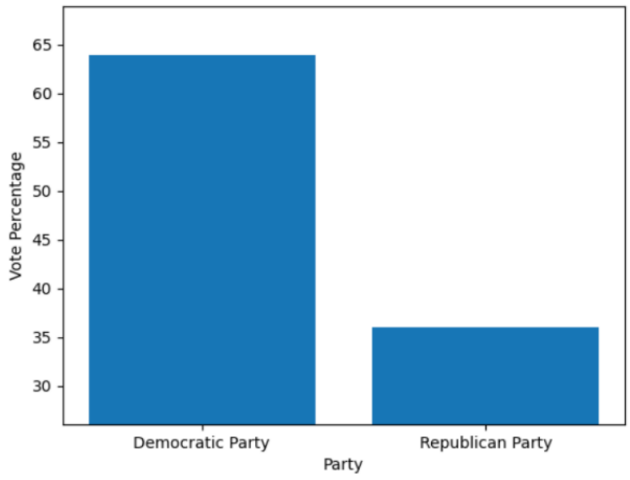
Another common violation is when the baseline of a bar chart is not at zero.

(1) Consider the following two bar chart:



Left graph exaggerates the number as it is not at zero. But, the graph on right side shows same data with the baseline at zero.

(2) The following figure represents the share of votes in an election in USA.



From the length of the bar we observe that Republic party voting percentage is less than half of the Democratic party but if we consider the actual number this is not the case.

3.4.3 Manipulated y-axis

Expanding or compressing the scale on a graph that can make changes in the data seem less significant than they actually are, is known as the manipulation of y-axis.
For example: Following bar charts represent the number of sales of smart phone A and B of a local shop.

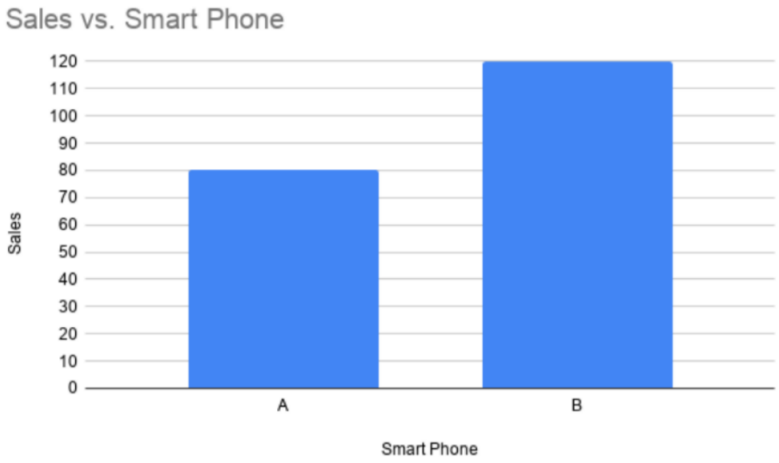


Figure : 3.8

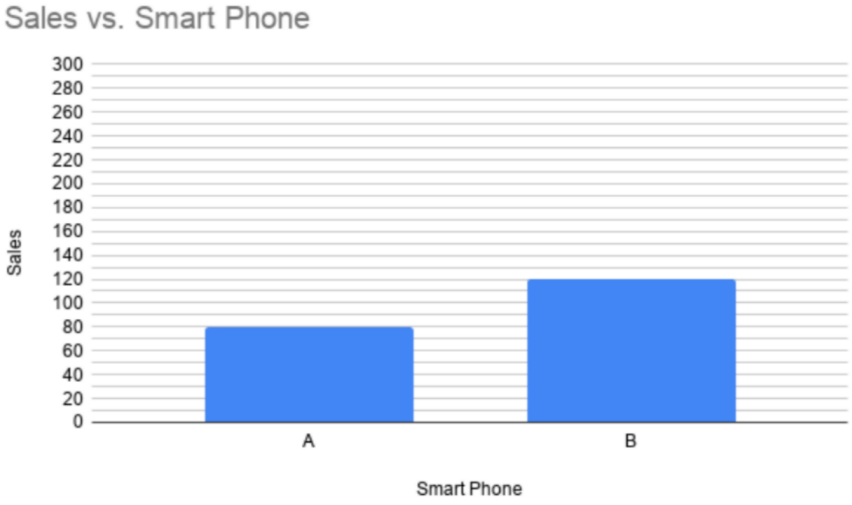


Figure : 3.9

From the figure 3.8 we are getting the information that a significant amount of sales is being done of both the smart phones but from the figure 3.9 it seems that the sales is very low of the smart phone A and B. So, the graph in figure 3.9 is misleading because it has manipulated y-axis.

3.4.4 Indicating a y-axis break

We can indicate a y-axis break in a bar chart in the following way:

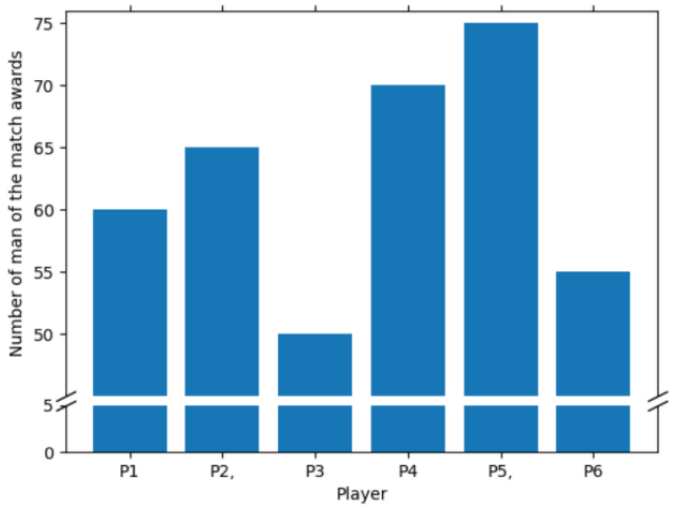


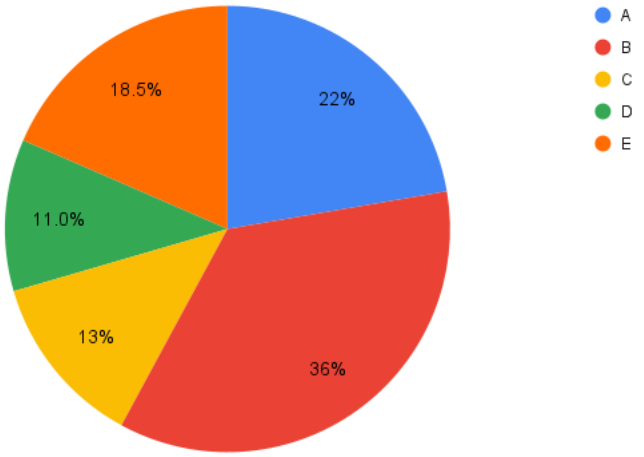
Figure : 3.10

3.4.5 Round-off errors

It is important to check for round-off errors. Round-off errors occur when table entries are percentages or proportions, the value of total sum may slightly differ from 100% or 1. This might result in a pie chart.
For Example: Consider the following table:

Category	Percentage
A	22.3
B	35.6
C	12.6
D	11
E	18.5
Total	100

In the table, the value of total sum is 100%.
Suppose, we round off the values and draw a pie chart as follows:



In this pie chart has round-off errors because total sum of all entries is 100.5% which is different from 100%.

3.5 Summarizing Categorical Data

- Bar chart and Pie chart are graphical summaries of categorical data.
- Numbers that are used to describe data sets are called descriptive measures.
- Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of central tendency.

3.5.1 Mode

The mode of a categorical variable is the most common category, the category with the highest frequency.
Mode labels the longest bar in a bar chart, the widest slice in a pie chart and the first category shown in a Pareto chart.
Example: Let's consider the dataset A, A, B, C, A, D, A, B, C, C, A, B, C, D, A.
Here, category A is the mode of the data as it occurs with the highest frequency.
Now, figure 3.11, 3.12 and 3.13 represent the bar chart, pie chart and pareto chart for the dataset as follows:

(1) Bar chart representation for the above dataset is:

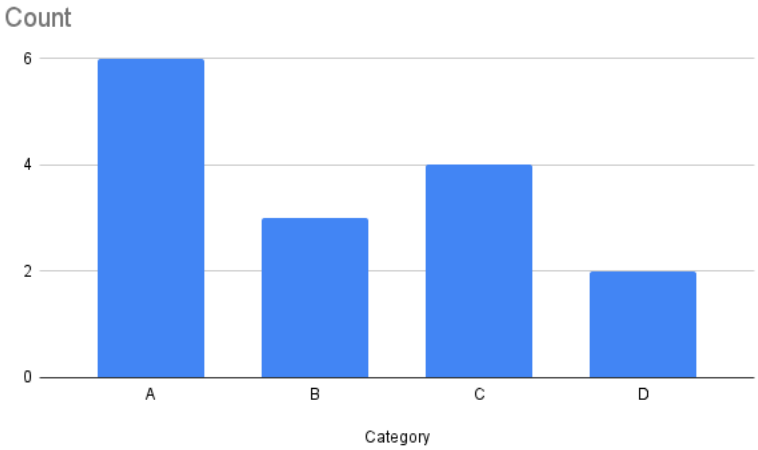


Figure : 3.11

In the figure 3.11, category A has the longest bar. Thus, mode of the dataset is category "A".

(2) Pie chart representation of the above dataset is:

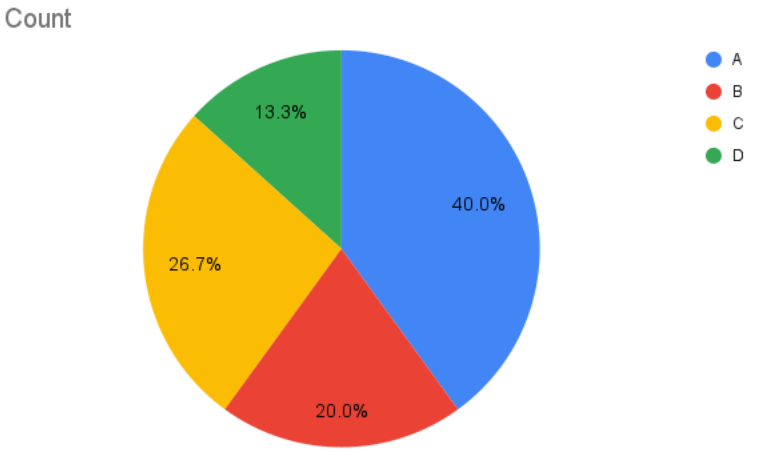


Figure : 3.12

In the above pie chart, category A has the widest slice. Thus, mode of the dataset is category "A".

(3) Pareto chart for the above dataset is:

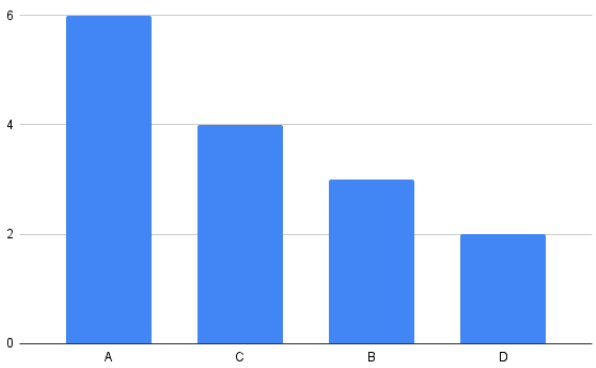


Figure : 3.13

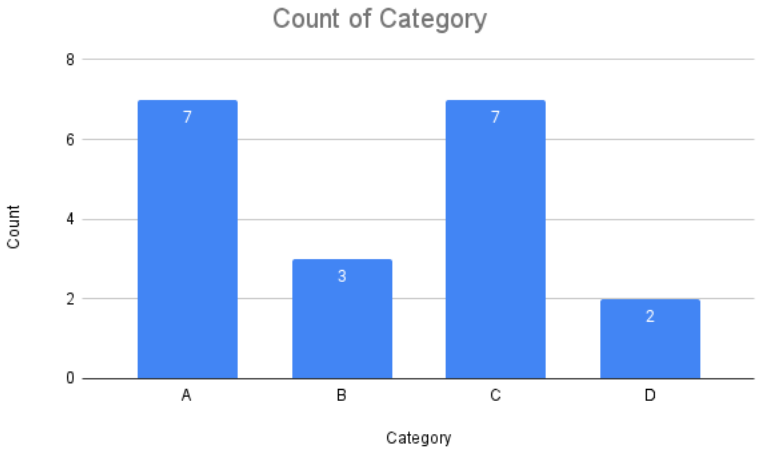
In the above pareto chart, first bar is for category A. Thus, mode of the dataset is category "A".

3.5.1.1 Bimodal and Multimodal data

If two or more categories tie for the highest frequency, the data is called bimodal (in the case of two) or multimodal (more than two).

Example:

Let's consider the dataset A, A, B, C, A, C, A, B, C, C, A, C, C, D, A, A, C, D, B.
Here both categories "A" and "C" have highest frequency. Thus, this data is bimodal.
Now, we can consider the following bar chart also.



In the above bar chart, both categories "A" and "C" have highest frequency.

3.5.2 Median

The median of an ordinal variable is the category of the middle observation of the sorted values.
If there are an even number of observations, then we can choose the category on either side of the middle of the sorted list as the median.

Examples:

(1) When number of observations is odd:
Let's consider the grades of 15 students as A, B, B, C, A, D, B, B, A, C, B, B, C, D, A.
Now to find the median of the categorical data, we need to order the data. So, the ordered data is A, A, A, A, B, B, B, B, B, B, C, C, C, D, D.
Hence, the median grade is the category associated with the 8^{th} observation which is "B".

(2) When number of observations is even:
Let's consider the grades of 14 students which is listed as A, B, B, C, A, D, B, B, A, C, B, B, C, D.
Now, the ordered data is A, A, A, B, B, B, B, B, B, C, C, C, D, D.
The median grade is the category associated with the 7^{th} or 8^{th} observation which is "B".
In the example (1), mode of the dataset is also category "B". Here, mode and median both are same.

(3) Consider the grades of 15 students which is listed as A, B, B, C, A, D, A, B, A, C, B, A, C, D, A.
The ordered data is A, A, A, A, A, A, B, B, B, B, C, C, C, D, D.
The median grade is the category associated with the 8^{th} observation which is "B".
The most common grade is "A", hence mode is "A". In this example both mode and median are the different.

Note: Median can be defined only for ordinal data whereas mode can be defined for both nominal as well as ordinal data.

Unsolved Problems

(1) If an analyst wants to represent the revenues of various companies using graphs, then which of the following graphical representation/s is/are most appropriate for the purpose?(More than one option can be correct)

(a) A pie chart with a pie/slice for each company and the width corresponding to its revenue in crore rupees.

(b) A bar chart with a bar for each company on the x-axis and the length corresponding to its revenue in crore rupees on the y-axis.

(c) A bar chart with a bar for each company on the y-axis and the length corresponding to its revenue in crore rupees on the x-axis.

(d) A bar chart with the minimum revenue as a baseline.

Answer: b, c

(2) Mode of a categorical variable is:(More than one option can be correct)

(a) The last bar in ascending order of a Pareto chart.

(b) The middle-most bar in a Pareto chart.

(c) The longest bar in a bar chart.

(d) The widest slice in a pie chart.

Answer: a, c, d

(3) Which of the following can be defined for both nominal and ordinal data?

(a) Mean

(b) Median

(c) Mode

(d) All of the above

Answer: c

A total of 2000 cases of Covid-19 have been registered on 5^{th} May 2020 in 5 key districts of Maharashtra. The proportion (out of 5 districts) of cases in each district has been listed in Table 2.1.A. Based on the information given, answer questions (4) and (5).

District	Relative Frequency
Mumbai	0.35
Pune	0.20
Nagpur	x
Thane	0.25
Nashik	0.08

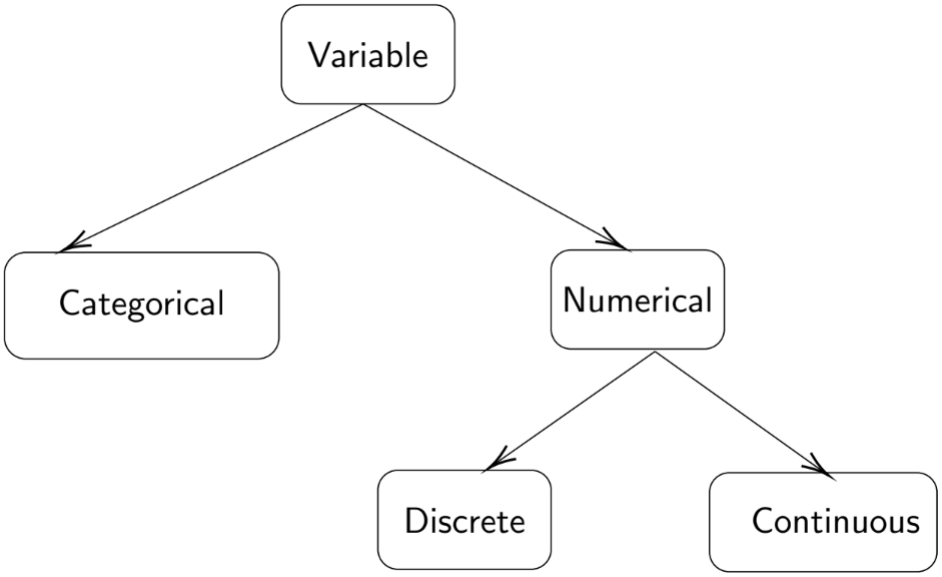
(4) Find the relative frequency of district Nagpur.
Answer: 0.12

(5) How many cases were registered in Pune on 5^{th} May?
Answer: 400

Chapter 4

4 Describing Numerical data

4.1 Types of variables



4.1.1 Discrete Variable

A discrete variable usually involves a count of something.

For example: Number of people in a household, Number of spelling mistakes in a report, Number of accidents in a month in a particular city etc.

4.1.2 Continuous Variable

A continuous variable usually involves a measurement of something.

For example: Weight of person, Height of a person, Speed of a vehicle etc.

4.2 Organizing Numerical Data

We can do the following procedures for organizing the numerical data.

(1) Group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of quantitative data.

(2) Once we group the quantitative data into classes, we can construct frequency and relative-frequency distributions of the data.

4.2.1 Organizing Discrete Data (single value)

We can proceed in the following ways for organizing the discrete data.

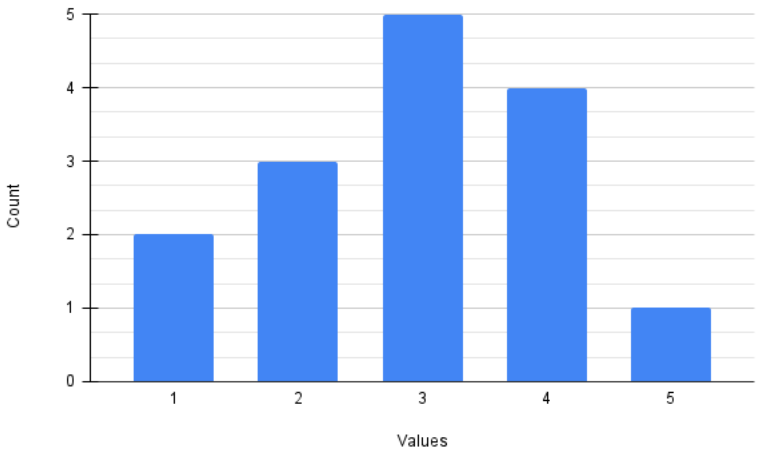
(1) If the data set contains only a relatively small number of distinct, or different, values, it is convenient to represent it in a frequency table.

(2) Each class represents a distinct value (single value) along with its frequency of occurrence.

For Example:
Suppose the dataset reports the number of people in a household and data of the response from 15 individuals is $2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4$.
The distinct values the variable, number of people in each household, takes is $1, 2, 3, 4, 5$.
The frequency distribution table is:

Value	Frequency	Relative frequency
1	2	0.13
2	3	0.2
3	5	0.33
4	4	0.27
5	1	0.07
Total	15	1

Here each value is considered as a category.
Now, let's consider the graph of the above data:



Since values are distinct, therefore we can't connect the bars. And, in the graph, we have just listed out about height of each bar.

4.2.2 Organizing Continuous Data

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed.

(1) Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.

(2) Each observation should belong to some class and no observation should belong to more than one class.

(3) It is common, although not essential, to choose class intervals of equal length.

4.2.2.1 Terminology

(1) Lower class limit: The smallest value that could go in a class.

(2) Upper class limit: The largest value that could go in a class.

(3) Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.

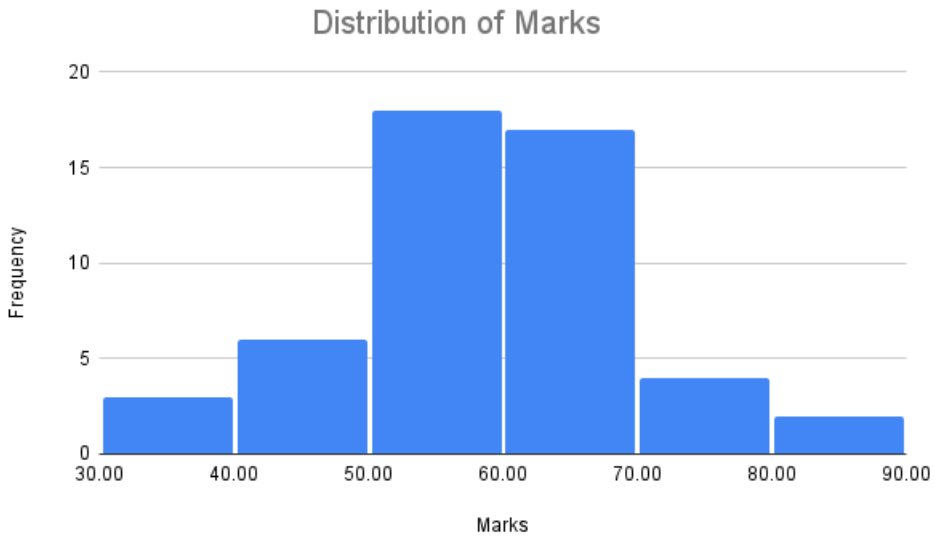
(4) Class mark: The average of the two class limits of a class.

(5) A class interval contains its left-end but not its right-end boundary point.

Example:
Consider the marks obtained by 50 students in a particular course which are as follows:
$68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53,$
$55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 54, 69, 58, 52, 42, 66, 42, 56, 58.$
Frequency table for the above dataset is:

Class Interval	Frequency	Relative frequency
$30-40$	3	0.06
$40-50$	6	0.12
$50-60$	18	0.36
$60-70$	17	0.34
$70-80$	4	0.08
$80-90$	2	0.04
Total	50	1

Graph for the above dataset is:



4.3 Stem-and-leaf diagram

In a stem-and-leaf diagram (or stemplot), each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.

For Example:

If the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.

The value 75 is expressed as:

Stem	Leaf
7	5

Here, 7 | 5 represents 75.

The two values 75, 78 is expressed as:

Stem	Leaf
7	5 8

Here, 7 | 5 represents 75.

4.3.1 Steps to construct a stemplot

(1) Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

(2) Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

(3) Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

(4) Arrange the leaves in each row in ascending order.

Example: Draw a stem-and-leaf plot for the dataset $15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48$ which are the ages of 11 patients admitted in a certain hospital.
Stem-and-leaf plot for the above dataset is :

Stem	Leaf				
1	0	5			
2	2	3	5	8	9
3	1	6			
4	5	8			

Here, 1 | 0 represents 10 years.

4.4 Descriptive Measures

Descriptive measures are quantities whose values are determined by the data and can be used to summarize a data set.

Types of Descriptive Measures
Most commonly used descriptive measures can be categorized as:
• **Measures of central tendency**: These are measures that indicate the most typical value or center of a data set.
• **Measures of dispersion**: These measures indicate the variability or spread of a dataset.

4.4.1 Measures of Central Tendency

4.4.1.1 Mean
The mean of a data set is the sum of the observations divided by the number of observations. And, mean is the most commonly used measure of central tendency.

- The mean is usually referred to as average.
- In arithmetic average, we have to divide the sum of the values by the number of values which is another typical value.
- Mean formula for discrete observations:

  (1) Sample mean $(\bar{x}) = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$

  (2) Population mean $(\mu) = \dfrac{x_1 + x_2 + \ldots + x_N}{N}$

Example(1):

(a) Mean of the observations $2, 12, 5, 7, 6, 7, 3$ can be computed as

$$\bar{x} = \frac{2 + 12 + 5 + 7 + 6 + 7 + 3}{7} = 6$$

(b) Mean of the observations $2, 105, 5, 7, 6, 7, 3$ can be computed as

$$\bar{x} = \frac{2 + 105 + 5 + 7 + 6 + 7 + 3}{7} = 19.29$$

(c) Mean of the observations $2, 105, 5, 7, 6, 3$ can be computed as

$$\bar{x} = \frac{2 + 105 + 5 + 7 + 6 + 3}{6} = 21.33.$$

Example(2):
Suppose the marks obtained by ten students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$. The sample mean is:

$$\frac{68 + 79 + 38 + 68 + 35 + 70 + 61 + 47 + 58 + 66}{10} = \frac{590}{10} = 59$$

- **Mean for grouped data: discrete single value data**

Mean formula for grouped data in case of discrete single value data is:

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{n}$$

Example(3):
Let's consider the dataset $2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4$ which are responses from 15 individuals.
Here, we can make the frequency table for the above dataset as follows:

Value(x_i)	Tally Mark	Frequency(f_i)	$f_i x_i$
1	| |	2	2
2	| | |	3	6
3	$\cancel{\text{\| \| \| \| \|}}$	5	15
4	| | | |	4	16
5	|	1	5
Total		15	44

Mean $= \frac{44}{15} = 2.93$

• **Mean for grouped data: continuous data**

Mean formula for grouped data in case of continuous data is:

$$\bar{x} = \frac{f_1 m_1 + f_2 m_2 + \ldots + f_n m_n}{n}$$

where m_i, $i = 1, 2, \ldots, n$, is the mid-point of i^{th} class-interval.

Example:

Class interval	Tally Mark	Frequency(f_i)	Mid point (m_i)	$f_i m_i$
$30 - 40$	𝍸	3	35	105
$40 - 50$	𝍸|	6	45	270
$50 - 60$	𝍸𝍸𝍸	18	55	990
$60 - 70$	𝍸𝍸𝍸||	17	65	1105
$70 - 80$	||||	4	75	300
$80 - 90$	||	2	85	170
Total		50		2940

By applying the formula, average $= \frac{2940}{50} = 58.8$ which is an approximate not exact value of the mean.

• **Adding a constant**

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and mean of the dataset is $\bar{x}$.
Let $y_i = x_i + c$, where c is a constant, then $\bar{y} = \bar{x} + c$.
Now,

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\implies \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

$$\implies \bar{y} = \frac{(x_1 + c) + (x_2 + c) + \ldots + (x_n + c)}{n}$$

$$\implies \bar{y} = \frac{(x_1 + x_2 + \ldots + x_n) + nc}{n}$$

$$\implies \bar{y} = \frac{x_1 + x_2 + \ldots + x_n}{n} + \frac{nc}{n}$$

$$\implies \bar{y} = \bar{x} + c$$

Example:

Suppose the marks obtained by 10 students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$ and average marks is 59. If teacher decided to add 5 marks to each students, then find the mean of new dataset.

Solution:
By the property of adding a constant, mean of the new dataset is

$$\bar{y} = \bar{x} + c = 59 + 5 = 64.$$

Also, we can verify as follows:
Since, marks of 10 students are $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$ and after adding 5 to each observations, new dataset will be $73, 84, 43, 73, 40, 75, 66, 52, 63, 71$, therefore mean of the new dataset is:

$$\bar{y} = \frac{73 + 84 + 43 + 73 + 40 + 75 + 66 + 52 + 63 + 71}{10} = \frac{640}{10} = 64 = 59 + 5.$$

• Multiplying a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and mean of the dataset is $\bar{x}$.
Let $y_i = x_i c$, where c is a constant, then $\bar{y} = \bar{x}c$.
Proof:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\implies \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

$$\implies \bar{y} = \frac{(x_1 c) + (x_2 c) + \ldots + (x_n c)}{n}$$

$$\implies \bar{y} = \frac{(x_1 + x_2 + \ldots + x_n)c}{n}$$

$$\implies \bar{y} = \bar{x}c$$

Example:

Suppose the marks obtained by 10 students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$ and average marks is 59.If teacher has decided to scale down each mark by 40%, i.e., each mark is multiplied by 0.4, then find the mean of new dataset.

Solution:
By the property of adding a constant, mean of the new dataset is

$$\bar{y} = \bar{x}c = 59 \times 0.4 = 23.6.$$

Also, we can verify as follows:
Since, marks of 10 students are $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$ and after multiplying 0.4 to each observations, new dataset will be $27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4$, therefore mean of the new dataset is:

$$\bar{y} = \frac{27.2 + 31.6 + 15.2 + 27.2 + 14 + 28 + 24.4 + 18.8 + 23.2 + 26.4}{10} = \frac{236}{10} = 23.6 = 59 \times 0.4.$$

4.4.1.2 Median

The median of a data set is the middle value in its ordered list. In other words, median of a data set is the number that divides the bottom 50% of the data from the top 50%.

• Steps to obtain median

Arrange the data in increasing order. Let n be the total number of observations in the dataset.

(1) If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e., $\left(\frac{n+1}{2}\right)^{th}$ observation.

(2) If the number of observations is even, then the median is the mean of the two middle observations in the ordered list, i.e., mean of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observation.

Examples:

(1) Compute the median of the dataset $2, 12, 5, 7, 6, 7, 3$.
Step(1): Arrange the data in increasing order: $2, 3, 5, 6, 7, 7, 12$
Step(2): Here, $n = 7$ which is odd.
So, median of the data will be $\left(\frac{n+1}{2}\right)^{th} = \left(\frac{8}{2}\right)^{th} = 4^{th}$ observation.
Thus, median is 6.

(2) Compute the median of the dataset $2, 105, 5, 7, 6, 7, 3$.
Step(1): Data in increasing order is: $2, 3, 5, 6, 7, 7, 105$
Step(2): Here, $n = 7$ which is odd.
So, median of the data is $\left(\frac{7+1}{2}\right)^{th} = 4^{th}$ observation which is 6.

(3) Compute the median of the dataset $2, 105, 5, 7, 6, 3$.
Step(1): Data in increasing order is: $2, 3, 5, 6, 7, 105$
Step(2): Here, $n = 6$ which is even.
So, median of the data will be average of $\left(\frac{6}{2}\right)^{th} = 3^{rd}$ and $\left(\frac{6}{2}+1\right)^{th} = 4^{th}$ observation.
Thus, median is $\frac{(5+6)}{2} = 5.5$.

• Adding a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i + c$, where c is a constant then, new median = old median+c .

Example:

Let's again consider the marks obtained by 10 students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$.

If teacher has decided to add 5 marks to each student, then find the median of new dataset.
Solution:
First, arrange the data in ascending order as $35, 38, 47, 58, 61, 66, 68, 68, 70, 79$.
Here, we have $n = 10$. So, Median of the dataset is $\frac{61+66}{2} = 63.5$.
Now, after adding 5 to each observations, new dataset in ascending order will be
$40, 43, 52, 63, 66, 71, 73, 73, 75, 84$.
Median of the new dataset is $\frac{66+71}{2} = 68.5 = 63.5 + 5 =$ old median + 5.

• **Multiplying a constant**

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i c$, where c is a constant then, new median = old median $\times$ c .

Example:

Let's again consider the marks obtained by 10 students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$.
If teacher has decided scale down each mark by 40%, i.e., each mark is multiplied by 0.4., then find the median of new dataset.
Solution:
As we know that median of this dataset is 63.5 from the previous example. After multiplying by 0.4 to each observation, new dataset will be $27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4$.
Ascending order of dataset is $14, 15.2, 18.8, 23.2, 24.4, 26.4, 27.2, 28, 31.6$.
Median of the new dataset is $\frac{24.4+26.4}{2} = 25.4 = 0.4 \times 63.5 =$ old median $\times$ 0.4.

• **Note:** "Mean is sensitive to outliers, whereas the median is not sensitive to outliers."

Example:

(1) For the dataset $2, 12, 5, 7, 6, 7, 3$
Mean $= \frac{2+3+5+6+7+7+12}{7} = 6$.
Now, arrange the data in ascending order: $2, 3, 5, 6, 7, 7, 12$
Median = 6.

(2) For the dataset $2, 117, 5, 7, 6, 7, 3$
Mean $= \frac{2+3+5+6+7+7+117}{7} = 21$.
Now, arrange the data in ascending order: $2, 3, 5, 6, 7, 7, 117$
Median = 6.

4.4.1.3 Mode

The mode of a dataset is its most frequently occurring value.

Examples:

(1) Find the mode of dataset $2, 12, 5, 7, 6, 7, 3$.
Mode is 7 as it occurs twice.

(2) Find the mode of dataset $2, 105, 5, 7, 6, 7, 3$.
Mode is 7.

(3) Find the mode of dataset $2, 105, 5, 7, 6, 3$.
There is no mode for the above dataset as no value occurs more than once.

• Adding a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i + c$, where c is a constant then, new mode = old mode + c .

Example:

Consider the marks obtained by 10 students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$. If teacher has decided to add 5 marks to each student, then find the mode of new dataset.

Solution:

Mode of the old dataset is 68. After adding 5 to each observation, new dataset will be $73, 84, 43, 73, 40, 75, 66, 52, 63, 71$.
Mode of new dataset is $73 = 68 + 5$=old mode + 5.

• Multiplying a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i c$, where c is a constant then, new mode = old mode $\times$ c .

Example:

Consider the marks obtained by 10 students in an exam is $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$. If teacher has decided to decided to scale down each mark by 40%, i.e., each mark is multiplied by 0.4, then find the mode of new dataset.

Solution:

Mode of the old dataset is 68. After multiplying 0.4 to each observation, new dataset will be $27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4$.
Mode of new dataset is $27.2 = 0.4 \times 68$ = old mode $\times$ 0.4.

4.4.2 Measures of Dispersion

Measure of dispersion indicates the amount of variation, or spread, in a dataset. These measures also known as measures of variation, or measures of spread.

Some of measures of dispersion are:

(1) Range

(2) Variance

(3) Standard Deviation

(4) Interquartile range

4.4.2.1 Range

The range of a dataset is the difference between its largest and smallest values.
The range of a dataset is given by the formula:

$$Range = Max - Min$$

Where, Max and Min represent the maximum and minimum values of dataset respectively.
Examples:

(1) Find the range of the dataset $3, 3, 3, 3, 3$.
Solution:
Here, maximum value of the dataset is 3 and minimum value is also 3.
Therefore,

$$Range = Max - Min$$
$$Range = 3 - 3 = 0.$$

(2) Find the range of the dataset $1, 2, 3, 4, 5$.
Solution:
Here, maximum value of the dataset is 5 and minimum value is 1.
Therefore,

$$Range = Max - Min$$
$$Range = 5 - 1 = 4.$$

• Effect of outliers on Range

Range is sensitive to outliers as it takes into consideration only the minimum and maximum value of the dataset.
For example:

(1) Dataset 1 : $1, 2, 3, 4, 5$.
Range of the dataset $1 = 5 - 1 = 4$.

(2) Dataset 2: $1, 2, 3, 4, 15$.
Range of the dataset $2 = 15 - 1 = 14$.

The above two datasets differ only in one point and this point changes the value of Range significantly. And, this significant change happens because range depends only on the maximum and minimum value of the dataset.

4.4.2.2 Variance

Variance measures the variability of a data set and considers the deviations of the data values from the central value.
Since, Range is also a measure of dispersion and it takes into account only minimum and maximum value of the dataset whereas variance takes into account all the observations.

Population variance and sample variance

The two variances can be computed using the following formulae:

- Population variance $(\sigma^2) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$
- Sample variance $(s^2) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Example: Consider the dataset $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$.

(1) Compute population variance of the dataset.
Solution:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
68	9	81
79	20	400
38	-21	441
68	9	81
35	-24	576
70	11	121
61	2	4
47	-12	144
58	-1	1
66	-7	49
$\sum x_i = 590$		$\sum(x_i - \bar{x})^2 = 1898$

Mean $(\bar{x}) = \dfrac{590}{10} = 59$

Population variance $= \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \dfrac{1898}{10} = 189.8$

(2) Compute the sample variance of the dataset.

sample variance $= \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \dfrac{1898}{9} = 210.89$

• Adding a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i + c$, where c is a constant then, new variance = old variance.

Proof:
Let population variance of old dataset $x_1, x_2, \ldots, x_n$ is σ^2_{old} and for new dataset is σ^2_{new}.
Now,

$$\sigma^2_{new} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

On substituting the values of $y_i = x_i + c$ and $\bar{y} = \bar{x} + c$ in the above equation, we get

$$\sigma^2_{new} = \frac{\sum_{i=1}^{n}(x_i + c - (\bar{x} + c))^2}{n}$$

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

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

$$= \sigma^2_{old}$$

Hence, there is no change in the variance of new dataset on adding a constant to each observations of old dataset, i.e., new variance = old variance.

For example:

Consider the dataset in the example **1** of population variance, if we add 4 to each observations then population variance of new dataset will be:

x_i	$y_i = x_i + 4$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
68	72	9	81
79	83	20	400
38	42	-21	441
68	72	9	81
35	39	-24	576
70	74	11	121
61	65	2	4
47	51	-12	144
58	62	-1	1
66	70	-7	49
$\sum x_i = 590$	$\sum y_i = 630$		$\sum(y_i - \bar{y})^2 = 1898$

Mean $(\bar{y}) = \dfrac{630}{10} = 63$

Population variance $= \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \dfrac{1898}{10} = 189.8$, which is same as variance of old dataset.

• Multiplying a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i \times c$, where c is a constant then, new variance $= c^2 \times$ old variance.

Proof:

Let population variance of old dataset $x_1, x_2, \ldots, x_n$ is σ^2_{old} and for new dataset is σ^2_{new}.
Now,

$$\sigma^2_{new} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

On substituting the values of $y_i = x_i \times c$ and $\bar{y} = \bar{x} \times c$ in the above equation, we get

$$\sigma^2_{new} = \frac{\sum_{i=1}^{n}(cx_i - c\bar{x})^2}{n}$$
$$= \frac{\sum_{i=1}^{n}(c(x_i - \bar{x}))^2}{n}$$
$$= \frac{c^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$
$$= c^2 \times \sigma^2_{old}$$

For example:

Consider the dataset in the example **1** of population variance, if we multiplied by 0.5 to each observations then population variance of new dataset will be:

x_i	$y_i = 0.5 \times x_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
68	34	4.5	20.25
79	39.5	10	100
38	19	−10.5	110.25
68	34	4.5	20.25
35	17.5	−12	144
70	35	5.5	30.25
61	30.5	1	1
47	23.5	−6	36
58	29	−0.5	0.25
66	33	3.5	12.25
$\sum x_i = 590$	$\sum y_i = 295$		$\sum(y_i - \bar{y})^2 = 474.5$

Mean $(\bar{y}) = \frac{295}{10} = 29.5$

Population variance $= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{474.5}{10} = 47.45 = 0.5^2 \times 189.8$

4.4.2.3 Standard Deviation

Standard deviation is also the measure of dispersion and it is square root of the variance.

Formulas of standard deviation

The population standard deviation and sample standard deviation can be computed by using the following formulae:

- Population standard deviation $(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$
- Sample standard deviation $(s) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Examples:

(1) Consider the dataset in the example **1** of variance and value of population standard deviation can be computed as follows:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
68	9	81
79	20	400
38	-21	441
68	9	81
35	-24	576
70	11	121
61	2	4
47	-12	144
58	-1	1
66	-7	49
$\sum x_i = 590$		$\sum(x_i - \bar{x})^2 = 1898$

Mean $(\bar{x}) = \dfrac{590}{10} = 59$

Population variance $= \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \dfrac{1898}{10} = 189.8$

So, population standard deviation $= \sqrt{189.8} = 13.78$.

(2) In the example **2** of variance, value of sample variance is computed as 210.89.
So the sample standard deviation will be $\sqrt{210.89} = 14.52$.

Units of standard deviation

Variance is expressed in units of square units as units of original variable while standard deviation is expressed in the same units as original data.

For Examples:

(1) If we have a dataset of weights of 10 students which is measured in kg, then the unit of variance will be $(kg)^2$ and units of standard deviation will be kg.

(2) If we have a dataset of age of 10 students which is measured in $year$, then the unit of variance will be $(year)^2$ and units of standard deviation will be $year$.

• Adding a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i + c$, where c is a constant then, new standard deviation = old standard deviation.

Proof:

Let population standard deviation of old dataset $x_1, x_2, \ldots, x_n$ is σ_{old} and for new dataset is σ_{new}.

Now,

$$\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

On substituting the values of $y_i = x_i + c$ and $\bar{y} = \bar{x} + c$ in the above equation, we get

$$\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n}(x_i + c - (\bar{x} + c))^2}{n}}$$

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

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

$$= \sigma_{old}$$

Similarly, we can prove for sample standard deviation.
Hence, there is no change in the standard deviation of new dataset on adding a constant to each observations of old dataset, i.e., new standard deviation = old standard deviation.

For example:
Consider the dataset in the example **1** of population standard deviation, if we add 4 to each observations then population standard deviation of new dataset will be:

x_i	$y_i = x_i + 4$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
68	72	9	81
79	83	20	400
38	42	-21	441
68	72	9	81
35	39	-24	576
70	74	11	121
61	65	2	4
47	51	-12	144
58	62	-1	1
66	70	-7	49
$\sum x_i = 590$	$\sum y_i = 630$		$\sum(y_i - \bar{y})^2 = 1898$

Mean $(\bar{y}) = \dfrac{630}{10} = 63$

Population standard deviation $= \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\dfrac{1898}{10}} = \sqrt{189.8} = 13.78$, which is same as the population standard deviation of old dataset.

• Multiplying a constant

Suppose $x_1, x_2, \ldots, x_n$ are observations of a dataset and let $y_i = x_i \times c$, where c is a constant

then, new standard deviation = $c \times$ old standard deviation.

Proof:
Let population standard deviation of old dataset $x_1, x_2, \ldots, x_n$ is σ_{old} and for new dataset is σ_{new}.
Now,

$$\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

On substituting the values of $y_i = x_i \times c$ and $\bar{y} = \bar{x} \times c$ in the above equation, we get

$$\begin{aligned}
\sigma_{new} &= \sqrt{\frac{\sum_{i=1}^{n}(cx_i - c\bar{x})^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}(c(x_i - \bar{x}))^2}{n}} \\
&= \sqrt{\frac{c^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \\
&= c \times \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \\
&= c \times \sigma_{old}
\end{aligned}$$

For example:
Consider the dataset in the example **1** of population standard deviation, if we multiplied by 0.5 to each observations then population standard deviation of new dataset will be:

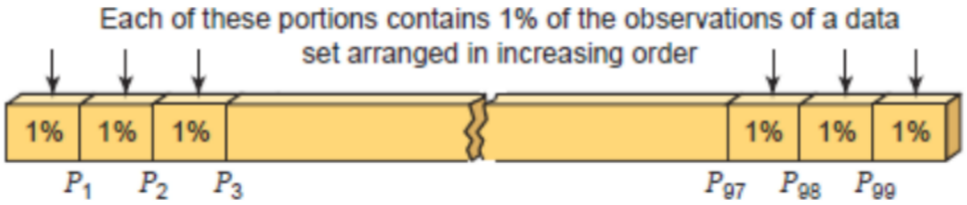
x_i	$y_i = 0.5 \times x_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
68	34	4.5	20.25
79	39.5	10	100
38	19	-10.5	110.25
68	34	4.5	20.25
35	17.5	-12	144
70	35	5.5	30.25
61	30.5	1	1
47	23.5	-6	36
58	29	-0.5	0.25
66	33	3.5	12.25
$\sum x_i = 590$	$\sum y_i = 295$		$\sum(y_i - \bar{y})^2 = 474.5$

Mean $(\bar{y}) = \dfrac{295}{10} = 29.5$

Population standard deviation $= \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{47.45} = 6.89 = 0.5 \times 13.78.$

4.5 Percentiles

The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.
We can understand the percentiles from the following figure:



In the above figure, we can interpret that P_9 which is 99^{th} percentiles would have 100×0.99, i.e., 99% of the data is less than it, but 1% is greater than it. Similarly, P_1 says 1% is less than it whereas, 99% is greater than or equal to it.
Thus, the percentiles tells us that value in the dataset below which we have $100 \times p$ which are less than or equal to it and $100 \times (1–p)$ which are greater than or equal to it. And, If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these values.

4.5.1 Computing Percentiles

To find the sample $100p$ percentile of a data set of size n, we need to follow the following steps:

(1) Arrange the data in increasing order.

(2) If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.

(3) If np is an integer, then the average of the values in positions np and $np+1$ is the sample $100p$ percentile.

Examples:

Consider the dataset $68, 38, 66, 79, 61, 47, 68, 35, 70, 58$.

(1) Compute the 25^{th} percentiles of the dataset.

Solution:

First, arrange data in ascending order $35, 38, 47, 58, 61, 66, 68, 68, 70, 79$.
Here, $n = 10$ and $p = 0.25$
$np = 10 \times 0.25 = 2.5$
Since np is in decimal and the smallest integer greater than 2.5 is 3. So, 3^{rd} observation of the dataset will be 25^{th} percentile which is 47.

(2) Compute the 75^{th} percentiles of the dataset.

Solution:

First, arrange data in ascending order $35, 38, 47, 58, 61, 66, 68, 68, 70, 79$.
Here, $n = 10$ and $p = 0.75$
$np = 10 \times 0.75 = 7.5$
Since np is in decimal and the smallest integer greater than 7.5 is 8. So, 8^{th} observation of the dataset will be 75^{th} percentile which is 68.

(3) Compute the 10^{th} percentiles of the dataset.

Solution:

First, arrange data in ascending order $35, 38, 47, 58, 61, 66, 68, 68, 70, 79$.
Here, $n = 10$ and $p = 0.10$
$np = 10 \times 0.10 = 1$
Since np is in integer, so we need to take the average of 1^{st} observation and 2^{nd} observation. So, 10^{th} percentile of the dataset is $\frac{35 + 38}{2} = 36.5$.

(4) Compute the 50^{th} percentiles of the dataset.

Solution:

First, arrange data in ascending order $35, 38, 47, 58, 61, 66, 68, 68, 70, 79$.
Here, $n = 10$ and $p = 0.50$
$np = 10 \times 0.50 = 5$
Since np is in integer, so we need to take the average of 5^{th} observation and 6^{th} observation. So, 50^{th} percentile of the dataset is $\frac{61 + 66}{2} = 63.5$.

4.6 Quartiles

Quartiles are the values that divide a given dataset into four parts by three quarters.
The sample 25^{th} percentile is called the first quartile, the sample 50^{th} percentile is called the median or the second quartile and the sample 75^{th} percentile is called the third quartile.
In other words, the quartiles break up a data set into four parts with about 25 percent of the data values being less than the first(lower) quartile, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third(upper) quartiles, and about 25 percent being larger than the third quartile.
Also, Q_1 represents the first quartile, Q_2 represents the second quartile and Q_3 represents the third quartile of the dataset.

For examples:

(1) In the example **1** of percentile, value of 25^{th} percentile is 47 which is Q_1(first quartile).

(2) In the example **4** of percentile, value of 50^{th} percentile is 63.5 which is Q_2(second quartile).

(3) In the example **2** of percentile, value of 75^{th} percentile is 68 which is Q_3(first quartile).

4.7 Five Number Summary

Five number summary is a very good way of summarizing a dataset and it is a set of descriptive statistics that provides information about the dataset.
Five number summary are as follows:
Minimum
Q_1 : First Quartile
Q_2 : Second Quartile or Median
Q_3 : Third Quartile
Maximum

For example:

Find the five number summary of the dataset $18, 28, 16, 29, 11, 27, 26, 35, 37, 28$.
Solution:
First, arrange data in ascending order $11, 16, 18, 26, 27, 28, 28, 29, 35, 37$.
Minimum value of the dataset is 11.
$n = 10$ and $p = 0.25$, $np = 10 \times 0.25 = 2.5$. Thus, 3^{rd} observation of the dataset is the value of Q_1 which is 18.
Now, $np = 10 \times 0.50 = 5$.
Thus, average of 5^{th} observation and 6^{th} observation of the dataset is the value of Q_2 which is $\dfrac{27 + 28}{2} = 27.5$.
Now, $np = 10 \times 0.75 = 7.5$. Thus, 8^{th} observation of the dataset is the value of Q_3 which is 29.
Maximum value of the dataset is 37.
Hence, five number summary of the dataset are $11, 18, 27.5, 29, 37$.

4.8 Interquartile Range (IQR)

The interquartile range, IQR, is the difference between the first and third quartiles.

$$IQR = Q_3 - Q_1$$

For example:

In the example **1** of percentile, value of 25^{th} percentile is 47 which is Q_1(first quartile) and in the example **2** of percentile, value of 75^{th} percentile is 68 which is Q_3(first quartile).
Thus, IQR of the dataset will be $Q_3 - Q_1 = 68 - 47 = 21$.

Chapter 5

5 Association between two variables

Association between two variables means knowing information about one variable provides information about the other variable.

5.1 Association Between Two Categorical Variables

To find the association between two categorical variables, first we have to make a contingency table and need to consider the following criteria.

- If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.
- If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.

Note: To know the association between two categorical variables from the contingency table, need to calculate either row relative frequencies or column relative frequencies.

Examples:

(1) A market research firm is interested in finding out whether ownership of a smartphone is associated with gender of a student. For this, a group of 100 college going children were surveyed about whether they owned a smart phone or not and following information is received.

(i) There are 44 female and 56 male students.

(ii) 76 students owned a smartphone and 24 did not own.

(iii) 34 female students owned a smartphone and 42 male students owned a smartphone.

Now, the given data can be organized in a contingency table as follows:

Gender	Own a smartphone		
	No	Yes	Row Total
Female	10	34	44
Male	14	42	56
Column Total	24	76	100

Table 5.1

Now, we can find the table for row relative frequency by dividing each cell frequency in a row by its row total:

Gender	Own a smartphone		
	No	Yes	Row Total
Female	$\frac{10}{44}$	$\frac{34}{44}$	44
Male	$\frac{14}{56}$	$\frac{42}{56}$	56
Column Total	$\frac{24}{100}$	$\frac{76}{100}$	100

Table 5.2

Gender	Own a smartphone		
	No	Yes	Row Total
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column Total	24.00%	76.00%	100

Table 5.3

In the above table, we can easily observe that row relative frequencies are the same for all the rows. So, we can say that two categorical variables, i.e., Gender and Smartphone ownership are not associated with each other.

We can also find the association between two categorical variables by column relative frequency which can be computed by by dividing each cell frequency in a column by its coulmn total.
And, column relative frequency for the dataset in Table 5.1 is as follows:

Gender	Own a smartphone		
	No	Yes	Row Total
Female	$\frac{10}{24}$	$\frac{34}{76}$	$\frac{44}{100}$
Male	$\frac{14}{24}$	$\frac{42}{76}$	$\frac{56}{100}$
Column Total	24	76	100

Table 5.4

Gender	Own a smartphone		
	No	Yes	Row Total
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100

Table 5.5

In the above table, we can easily observe that column relative frequencies are the same for all the columns. So, we can say that two categorical variables, i.e., Gender and Smartphone ownership are not associated with each other.

Hence, we can observe from Table 5.3 and Table 5.5, if the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.

(2) An analyst is interested in finding out whether ownership of a smartphone is associated with the income of an individual. For this, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not and following information is received.

(i) There are 20 high income, 66 medium income and 14 low income individuals.

(ii) 62 individuals owned a smartphone and 38 did not own.

(iii) 18 High income individuals owned a smartphone, 39 Medium income individuals owned a smartphone, and 5 Low income individuals owned a smartphone.

Now, the given data can be organized in a contingency table as follows:

Income Level	Own a smartphone		
	No	Yes	Row Total
High	2	18	20
Medium	27	39	66
Low	9	5	14
Column Total	38	62	100

Table 5.6

Now, row relative frequency table for the dataset in Table 5.6 is as follows:

Income Level	Own a smartphone		
	No	Yes	Row Total
High	$\frac{2}{20}$	$\frac{18}{20}$	20
Medium	$\frac{27}{66}$	$\frac{39}{66}$	66
Low	$\frac{9}{14}$	$\frac{5}{14}$	14
Column Total	$\frac{38}{100}$	$\frac{62}{100}$	100

Table 5.7

Income Level	Own a smartphone		
	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Table 5.8

Here, the row relative frequencies are different for some rows. Hence, we can say that the two categorical variables, i.e, income level and smartphone ownership are associated with each other.

Similarly, we can find the association between "Income level" and "smartphone ownership" by computing the column relative frequency as follows:

Income Level	Own a smartphone		
	No	Yes	Row Total
High	$\frac{2}{38}$	$\frac{18}{62}$	$\frac{20}{100}$
Medium	$\frac{27}{38}$	$\frac{39}{62}$	$\frac{66}{100}$
Low	$\frac{9}{38}$	$\frac{5}{62}$	$\frac{14}{100}$
Column Total	38	62	100

Table 5.9

Income Level	Own a smartphone		
	No	Yes	Row Total
High	5.26%	29.03%	20.00%
Medium	71.05%	62.90%	66.00%
Low	23.68%	8.06%	14.00%
Column Total	38	62	100

Table 5.10

Here, the column relative frequencies are different for some columns. Hence, we can say that the two categorical variables, i.e, income level and smartphone ownership are associated with each other.

(3) $10,000$ students have given IITM Online degree qualifier paper. Table 5.11 shows the distribution of students who passed the qualifier exam.

Gender	Qualified	
	Yes	No
Male	4800	1200
Female	3196	804

Table 5.11.

Is there any association between "Qualification" and "Gender"?

Solution:
To know the association between two categorical variables from the contingency table, we need to find out row relative frequencies or column relative frequencies.
Now, row relative frequencies for the dataset shown in Table 5.11 is as follows:

Gender	Qualified		
	Yes	No	Row total
Male	$\frac{4800}{6000}$	$\frac{1200}{6000}$	6000
Female	$\frac{3196}{4000}$	$\frac{804}{4000}$	4000
Column total	$\frac{7996}{10000}$	$\frac{2004}{10000}$	10000

Table 5.12.

Gender	Qualified		
	Yes	No	Row total
Male	80%	20%	6000
Female	79.9%	20.1%	4000
Column total	79.96%	20.04%	10000

Table 5.13.

From above table, it is clear that row relative frequencies are same for all the rows. Hence, we can conclude that qualification in the exam is not associated with gender of the person.

5.1.1 Stacked Bar Chart

A stacked bar chart represents the counts for a particular category. In addition, each bar is further broken down into smaller segments, with each segment representing the frequency of that particular category within the segment. A stacked bar chart is also referred to as a segmented bar chart.

5.1.2 100% Stacked Bar Chart

A 100% stacked bar chart is useful to part-to-whole relationships.

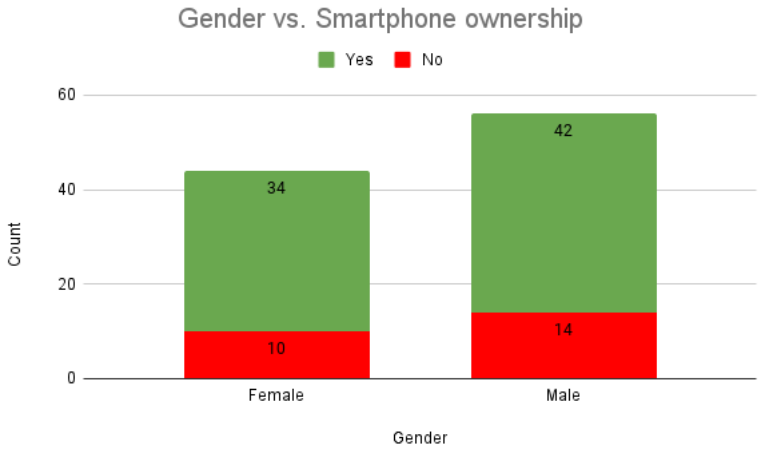
Examples:

(1) Consider the dataset in the Table 5.1 of example **1** which is as follows:

Gender	Own a smartphone		
	No	Yes	Row Total
Female	10	34	44
Male	14	42	56
Column Total	24	76	100

Table 5.1

Stacked bar chart for the dataset in Table 5.1. is:



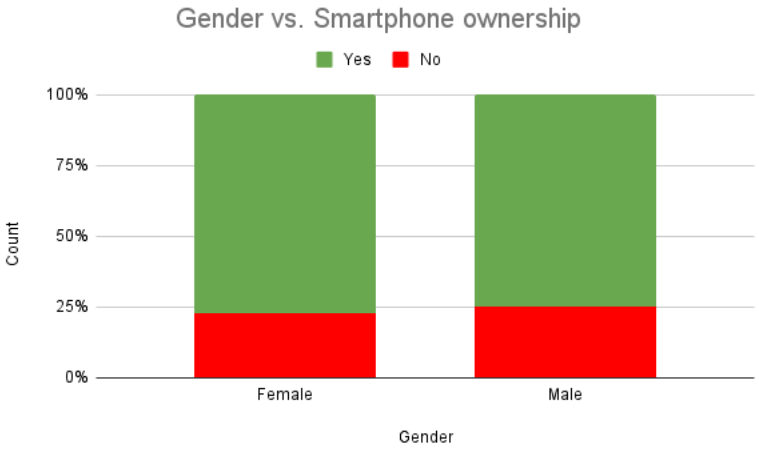
Here, first and second bar represent the count of female (44) and count of male (56) respectively. And, further each bars are broken down into 2 segments, one segment for count of smartphone owners and another segment for smartphone non-owners in each category.

As 100% stacked bar chart is useful to part-to-whole relationship. So, consider the row relative frequency Table 5.3 of example **1** as follows:

Gender	Own a smartphone		
	No	Yes	Row Total
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column Total	24.00%	76.00%	100

Table 5.3

A 100% stacked bar chart for the dataset in Table 5.3 is as follows:

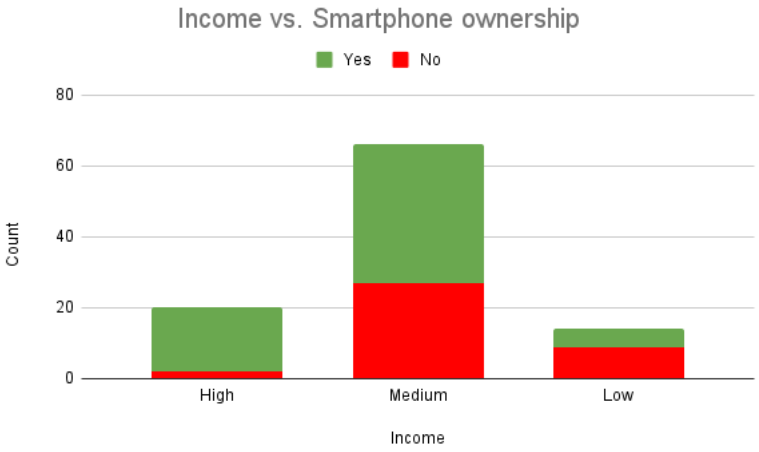


(2) Consider the dataset in the Table 5.6 of example **2** which is as follows:

Income Level	Own a smartphone		
	No	Yes	Row Total
High	2	18	20
Medium	27	39	66
Low	9	5	14
Column Total	38	62	100

Table 5.6

Stacked bar chart representation for the dataset in Table 5.6. is:

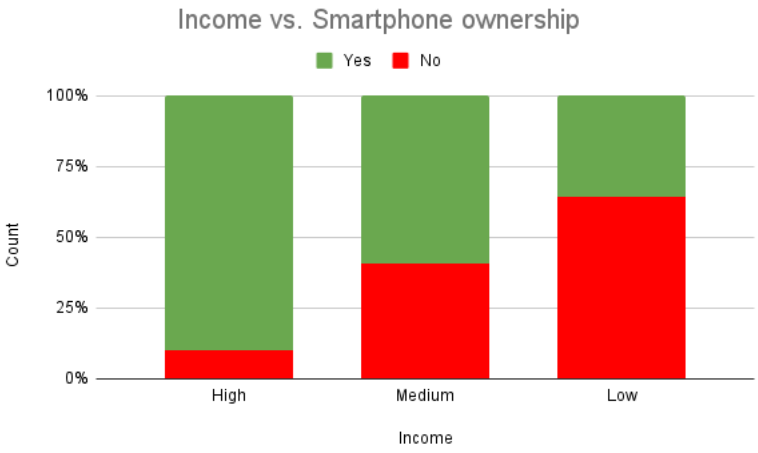


As 100% stacked bar chart is useful to part-to-whole relationship. So, consider the row relative frequency Table 5.8 of example **2** as follows:

Income Level	Own a smartphone		
	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Table 5.8

100% Stacked bar chart representation for the dataset in Table 5.8. is:



5.2 Association Between Two Numerical Variables

Generally, we use scatter plot to get an idea about association between two numerical variables and it is visual test for association.

5.2.1 Scatter Plot

A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane. It is a two-dimensional plot, i.e., we need a variable on x-axis (called independent variable) and a variable on y-axis (called dependent or response variable).

Examples:

(1) The following table represents the dataset of Age and Height of 5 childs.

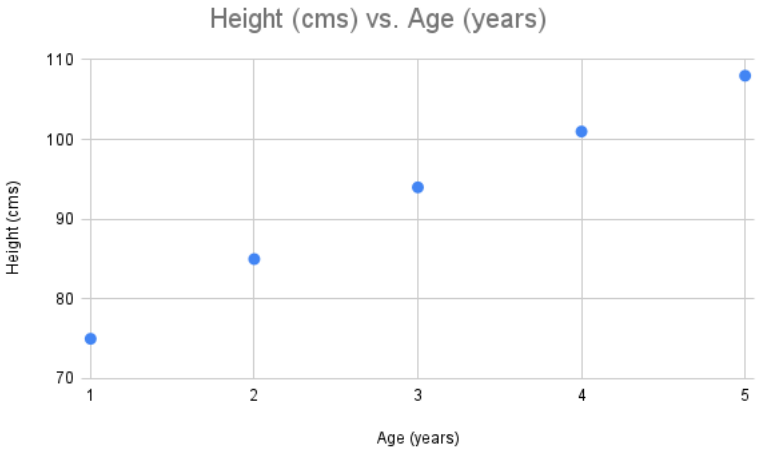
Age (years)	Height (cms)
1	75
2	85
3	94
4	101
5	108

Table 5.2.1

Is there any association between age and height?

Solution:

As we can use scatter plot to get an idea about association between two numerical variables.
Scatter plot representation for the dataset in Table 5.2.1 is:



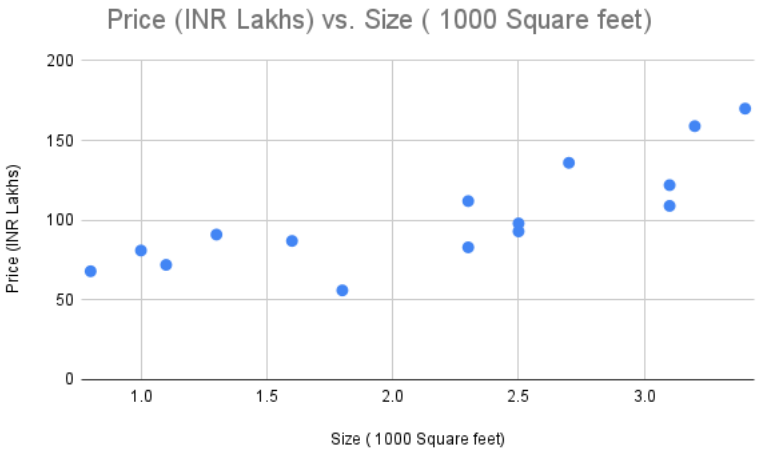
From the above scatter plot, it is clear that the height (cms) increase linearly with age (years).

(2) A real estate agent collected the prices of different sizes of homes. He wanted to see what was the relationship between the price of a home and size of a home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way? To answer the question, he collected data on 15 homes and tabulated the data in the following manner:

Size (1000 Square feet)	Price (INR Lakhs)
0.8	68
1	81
1.1	72
1.3	91
1.6	87
1.8	56
2.3	83
2.3	112
2.5	93
2.5	98
2.7	136
3.1	109
3.1	122
3.2	159
3.4	170

Table 5.2.2

Scatter plot representation for the dataset in Table 5.2.2 is:



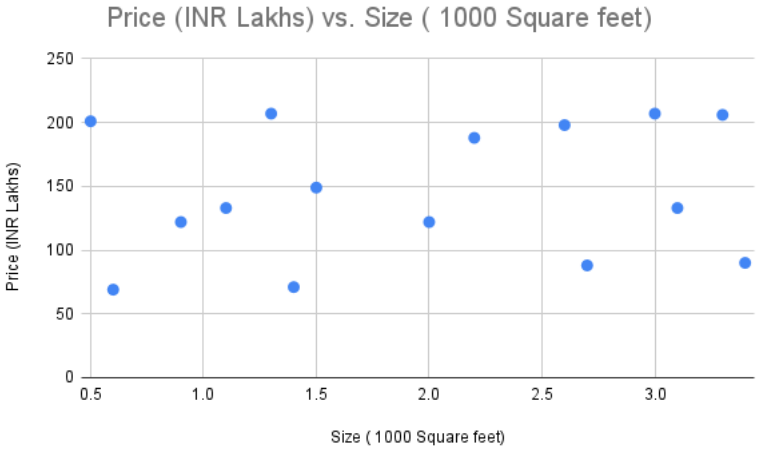
From the above scatter plot representation, we can observe that as sizes of the homes are increasing, the prices are also exhibiting some sort of a trend or it increases.

(3) Consider the collection of different dataset of 15 houses for the problem (2) and data is as follows:

Size (1000 Square feet)	Price (INR Lakhs)
0.5	201
0.6	69
0.9	122
1.1	133
1.3	207
1.4	71
1.5	149
2	122
2.2	188
2.6	198
2.7	88
3	207
3.1	133
3.3	206
3.4	90

Table 5.2.3

Scatter plot representation for the dataset in Table 5.2.3 is:



We cannot see any clear pattern in the above scatter-plot.

Thus, we can get an idea about association between two numerical variables by using scatter plot as elaborated in the above three examples.

5.2.1.1 Describing Association

To describe association between two numerical variables in a scatter plot, the following points need to be considered.

(1) Direction: Needs to check whether the pattern trend up, down or exhibits some sort of a trend.

(2) Curvature: Needs to check whether the pattern appear to be linear or any curve.

(3) Variation: Needs to check whether the points are tightly clustered along the pattern.

(4) Outliers: Needs to check whether there is any point which seems to be unexpected.

Describing Association: Direction

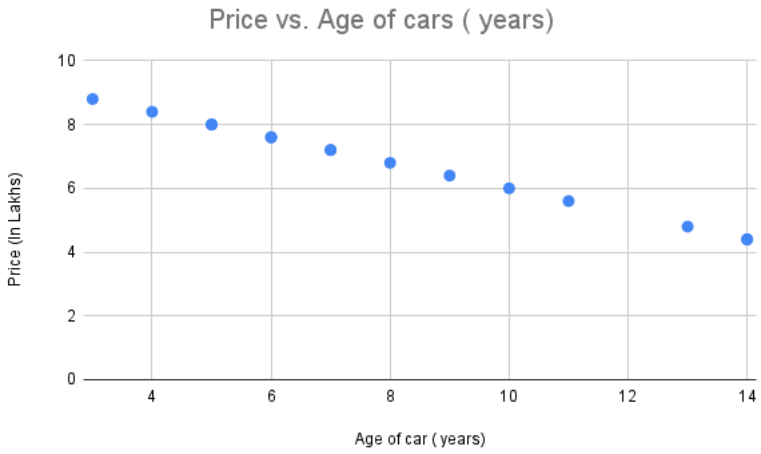
(1) By observing the scatter plot of dataset of example **2**, we can see an upward trend as sizes of homes are increasing the prices are also increasing.

(2) Suppose an analyst is interested to know the association between age of cars (years) and price of cars (in lakhs). For this, he collected data of 15 cars and tabulated as follows:

Age of cars(years)	Price (In Lakhs)
3	8.8
4	8.4
5	8
5	8
6	7.6
6	7.6
7	7.2
7	7.2
8	6.8
9	6.4
10	6
11	5.6
13	4.8
14	4.4
14	4.4

Table 5.2.4

Now, the scatter plot representation for the dataset of the above table is :



In this scatter plot, direction of the pattern is downward, i.e., as age of cars is increasing, price is decreasing.

5.2.2 Measures of association between two numerical variables

Measures by which we can measure the strength of association between two numerical variables are:

- Covariance
- Correlation

5.2.2.1 Covariance

Covariance quantifies the strength of the linear association between two numerical variables. Suppose x and y are two numerical variables and let x_i denote the i^{th} observation of variable x, and y_i denote the i^{th} observation of variable y, $i = 1, 2, \ldots, n$. Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset. The Covariance between the variables x and y is given by:

- Population covariance ; $Cov(x, y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$
- Sample covariance ; $Cov(x, y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$

Units of Covariance

The size of the covariance is difficult to interpret because the covariance has units and the units of the covariance are those of the $x-$variable times those of the $y-$variable.
For example: If one of the variable is measured in kilogram and other is measured in metre, then unit of covariance of two variables will be $kilogram \times meter$.

Important Points:

- If large (small) values of x tend to be associated with large (small) values of y, then the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be same.
- If large (small) values of x tend to be associated with small (large) values of y, then the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

Example: Consider the dataset of 15 cars of table **2** for computing the covariance.

Age (years) x_i	Price (In Lakhs) y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
3	8.8	-5.13	2.05	-10.5165
4	8.4	-4.13	1.65	-6.8145
5	8	-3.13	1.25	-3.9125
5	8	-3.13	1.25	-3.9125
6	7.6	-2.13	0.85	-1.8105
6	7.6	-2.13	0.85	-1.8105
7	7.2	-1.13	0.45	-0.5085
7	7.2	-1.13	0.45	-0.5085
8	6.8	-0.13	0.05	-0.0065
9	6.4	0.87	-0.35	-0.3045
10	6	1.87	-0.75	-1.4025
11	5.6	2.87	-1.15	-3.3005
13	4.8	4.87	-1.95	-9.4965
14	4.4	5.87	-2.35	-13.7945
14	4.4	5.87	-2.35	-13.7945
Total = 122	101.2			-71.8935

Table 5.2.5

$\bar{x} = \frac{122}{15} = 8.13$ and $\bar{y} = \frac{101.2}{15} = 6.75$

Population covariance $= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{-71.8935}{15} = -4.7929$ and,

Sample covariance $= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-71.8935}{14} = -5.13525$

5.2.2.2 Correlation

Correlation is a measure of linear association between two numerical variables and it is derived from covariance.

The pearson correlation coefficient, r, between two numerical variables x and y is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x, y)}{s_x s_y}$$

Note:

- Correlation coefficient, r, is an unit less measure because the units of the standard deviations cancel out the units of covariance.
- Correlation coefficient, r, always lies between -1 and $+1$.

Example: Consider the same dataset from the table 5.2.5 of above example of covariance topic **5.2.2.1**.

Age (years) x_i	Price (In Lakhs) y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
3	8.8	26.3169	4.2025	-10.5165
4	8.4	17.0569	2.7225	-6.8145
5	8	9.7969	1.5625	-3.9125
5	8	9.7969	1.5625	-3.9125
6	7.6	4.5369	0.7225	-1.8105
6	7.6	4.5369	0.7225	-1.8105
7	7.2	1.2769	0.2025	-0.5085
7	7.2	1.2769	0.2025	-0.5085
8	6.8	0.0169	0.0025	-0.0065
9	6.4	0.7569	0.1225	-0.3045
10	6	3.4969	0.5625	-1.4025
11	5.6	8.2369	1.3225	-3.3005
13	4.8	23.7169	3.8025	-9.4965
14	4.4	34.4569	5.5225	-13.7945
14	4.4	34.4569	5.5225	-13.7945
Total = 122	101.2	179.7335	28.7575	-71.8935

Table 5.2.6

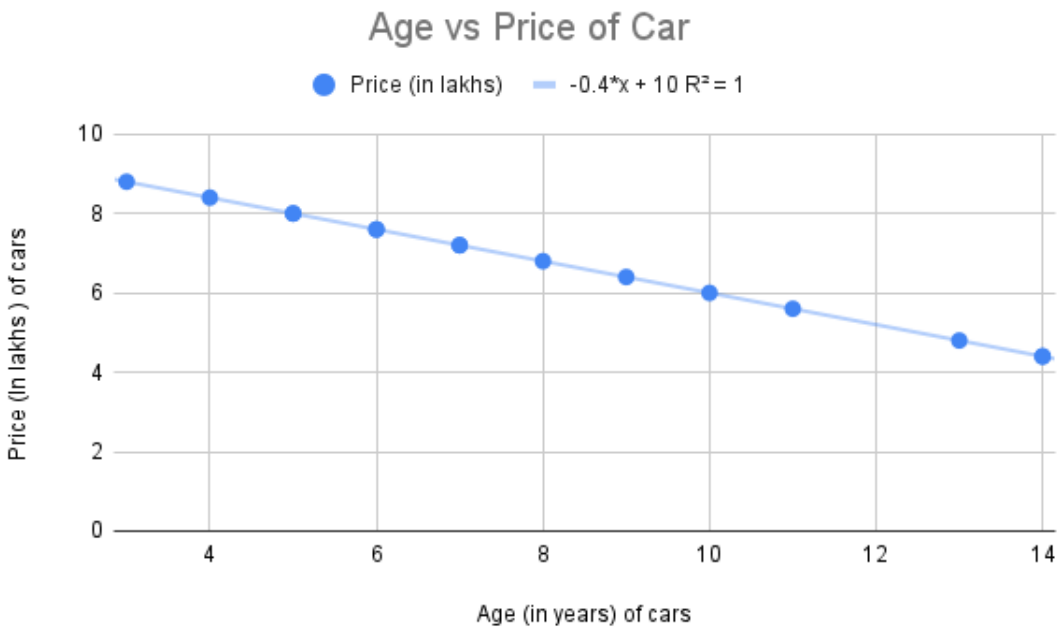
From the above table, we have

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -71.8935, \sum_{i=1}^{n}(x_i - \bar{x})^2 = 179.7335 \text{ and } \sum_{i=1}^{n}(y_i - \bar{y})^2 = 28.7575$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-71.8935}{\sqrt{179.7335}\sqrt{28.7575}} = -1.$$

5.2.2.3 Fitting a line

The linear association between two numerical variables can be described using the equation of a line which can be done by using google sheet.
Equation of the line for the data of car is given as below:



Equation of the line is : $Price = -4 \times Age + 10$ and $R^2 = 1 = -1^2 = r^2$.

where, R^2 captures the proportion of variance in the data set which is captured by the line. It is also referred as a goodness of fit measure and it takes the values between 0 and 1. If the value of R^2 is close to 1, then fit is good fit for our data. And, if the value is close to 0, then the fit is not a good fit for our data.

5.3 Association Between Categorical and Numerical Variables

Point Bi-serial correlation coefficient is a measure of association between categorical and numerical variable.

5.3.1 Point Bi-serial Correlation Coefficient

• Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).

• The following steps are used for calculating the Point Bi-serial correlation between these two variables.

- Step-1: Group the data into two sets based on the value of the dichotomous variable Y . That is, assume that the value of Y is either 0 or 1.

- Step-2 : Calculate the mean values of two groups. That is, let $\bar{Y}_0$ and $\bar{Y}_1$ be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.
- Step-3: Let p_0 and p_1 be the proportion of observations in a group with $Y = 0$ and $Y = 1$, respectively, and σ_x be the population standard deviation of the random variable X.

The correlation coefficient is :

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{\sigma_x} \right) \sqrt{p_0 p_1}$$

where, $p_0 = \dfrac{n_0}{n} = \dfrac{\text{observations coded with 0}}{\text{total observations}}$ and $p_1 = \dfrac{n_1}{n} = \dfrac{\text{observations coded with 1}}{\text{total observations}}$.

Example: A teacher was interested in knowing if female students performed better than male students in her class. She collected data from twenty students and the marks they obtained on 100 in the subject. Tabulated data is given as follows:

	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90

Table 5.2.7

Code 'F' as 1 and 'M' as 0 and vice-versa as gender is a dichotomous variable.

	Gender	Marks	Gender-coded
1	F	71	1
2	F	67	1
3	F	65	1
4	M	69	0
5	M	75	0
6	M	83	0
7	F	91	1
8	F	85	1
9	F	69	1
10	F	75	1
11	M	92	0
12	F	79	1
13	M	71	0
14	M	94	0
15	F	86	1
16	F	75	1
17	F	90	1
18	M	84	0
19	F	91	1
20	M	90	0

Table 5.2.8

From the above table, we have

$$p_0 = \frac{8}{20} = 0.4, p_1 = \frac{12}{20} = 0.6 \text{ and } \sigma_x = 9.33$$

$$\bar{Y}_0 = \frac{69+75+83+92+71+94+84+90}{8} = 82.25 \text{ and}$$

$$\bar{Y}_1 = \frac{71+67+65+91+85+69+75+79+86+75+90+91}{12} = 78.67$$

Now,

$$r_{pb} = \left(\frac{82.25-78.67}{9.33}\right) \times \sqrt{0.4 \times 0.6} = 0.188$$

Remarks:

- Absolute point bi-serial correlation coefficient can be computed as

$$r_{pb} = \left(\frac{|\bar{Y}_0 - \bar{Y}_1|}{\sigma_x} \right) \sqrt{p_0 p_1}$$

- Another formula of point bi-serial correlation coefficient is in terms of sample standard deviation is as follows:

$$r_{pb} = \left(\frac{|\bar{Y}_0 - \bar{Y}_1|}{\sigma_x} \right) \sqrt{p_0 p_1} \quad \ldots \ *$$

Since, relationship between population variance σ_x^2 and sample variance s_x^2 is

$$n\sigma_x^2 = (n-1)s_x^2$$

$$\implies \sigma_x = \sqrt{\frac{(n-1)s_x^2}{n}} = \sqrt{\frac{n-1}{n}} s_x$$

Now, on substituting the value of σ_x in equation (*), we get

$$r_{pb} = \left(\frac{|\bar{Y}_0 - \bar{Y}_1|}{\sqrt{\frac{n-1}{n}} s_x} \right) \sqrt{p_0 p_1}$$

More simplest form can be obtained by substituting the values of p_0 and p_1:

$$r_{pb} = \left(\frac{|\bar{Y}_0 - \bar{Y}_1|}{\sqrt{\frac{n-1}{n}} s_x} \right) \sqrt{\left(\frac{n_0}{n}\right)\left(\frac{n_1}{n}\right)}$$

$$r_{pb} = \left(\frac{|\bar{Y}_0 - \bar{Y}_1|}{s_x} \right) \sqrt{\left(\frac{n_0}{n-1}\right)\left(\frac{n_1}{n}\right)}$$

Chapter 6

6 Basic Principle of Counting

6.1 Introduction

6.1.1 Addition rule of counting

If an action A can occur in n_1 different ways, another action B can occur in n_2 different ways, then the total number of ways of occurrence of either actions A or B is $n_1 + n_2$.

Example: You have a gift card from a major retailer which allows you to buy "one" item, either a shirt or a pant. If the choices at the retailer are 1 yellow shirt, 1 blue shirt, 1 green shirt, 1 red shirt, 1 black pant, 1 blue pant and 1 brown pant, then in how many ways can you use your card?

Solution:
There are four choices to buy a shirt. You can buy either a yellow shirt or a blue shirt or a green shirt or a red shirt.
There are three choices to buy a pant. You can buy either a black pant or a blue pant or a brown pant.
If you choose to buy a shirt (pant), then you cannot buy a pant (shirt) because gift card allows you to buy "one" item, either a shirt or a pant.
Hence, the total number of choices available is $4 + 3 = 7$.

6.1.2 Multiplication rule of counting

If an action A can occur in n_1 different ways, another action B can occur in n_2 different ways, then the total number of ways of occurrence of actions A and B together is $n_1 \times n_2$.

Example: You have a gift card from a major retailer which allows you to one shirt and one pant. If the choices at the retailer are 1 yellow shirt, 1 blue shirt, 1 green shirt, 1 red shirt, 1 black pant, 1 blue pant and 1 brown pant, then in how many ways can you use your card?

Solution:
Here, card allows to buy one shirt and one pant, i.e, if you can buy a shirt (pant), then you can also buy a pant (shirt).
Thus, total choices will be as follows:

1. yellow shirt-black pant
2. yellow shirt-blue pant
3. yellow shirt-brown pant
4. blue shirt-black pant

5. blue shirt-blue pant
6. blue shirt-brown pant
7. green shirt-black pant
8. green shirt-blue pant
9. green shirt-brown pant
10. red shirt-black pant
11. red shirt-blue pant
12. red shirt-brown pant

Hence, you have 12 options in total to buy a shirt and pant together which is nothing but the application of multiplication rule as $4 \times 3 = 12$.

Generalization of Multiplication Rule:
Suppose that r actions are to be performed in a definite order. Further, suppose that there are n_1 possibilities for the first action and that corresponding to each of these possibilities are n_2 possibilities for the second action, and so on. Then there are $n_1 \times n_2 \times \ldots \times n_r$ possibilities altogether for the r actions.

Example: You have a gift card from a major retailer which allows you to buy one shirt, one pant and one pair of shoes. And, the choices at the retailer are 1 yellow shirt, 1 blue shirt, 1 green shirt, 1 red shirt, 1 black pant, 1 blue pant, 1 brown pant and 2 pairs of shoes (1 black pair of shoes and 1 brown pair of shoes), then in how many ways can you use your card?

Solution:
Here, three actions are to be performed, where the first action is buying a shirt, second action is buying a pant and third action is buying a pair of shoes.
Since there are 4 possibilities for buying a shirt, 3 possibilities for buying a pant and 2 possibilities for buying a pair of shoes, there are $4 \times 3 \times 2 = 24$ possibilities altogether for the three actions.
Hence, there are 24 ways in total in which card can be used.

6.1.2.1 Solved Examples:

Q1. Suppose you are asked to create a six digit alphanumeric password with the requirement that the password should have first two digits as alphabets (upper case) followed by four numbers, then in how many ways password can be created if the repetition is allowed?

Solution:
We are given that password should have first two letters followed by four numbers and

there are a total of 26 upper case alphabets and 10 digits $(0, 1, \ldots, 9)$ which can be used for the password.

Since the repetition is allowed, we can use any of the 26 alphabets for the first place. Similarly, for the second place we can use any of the 26 alphabets. Now, for the third place we can use any of the 10 digits, for the fourth place, again we can use any of the 10 digits, and so on.

So, here 6 actions are to be performed: first action is to choose an alphabet for the first place, second action is to choose an alphabet for the second place, third action is to choose a number for the third place, and so forth.

Thus, by the multiplication rule of counting, total number of ways of occurrence of all the 6 actions together is

$$26 \times 26 \times 10 \times 10 \times 10 \times 10 = 6,760,000$$

Hence, there are a total of $6,760,000$ ways in which password can be created.

Q2. Suppose you are asked to create a six digit alphanumeric password with the requirement that the password should have first two digits as alphabets (upper case) followed by four numbers, then in how many ways password can be created if the repetition is not allowed?

Solution:
We are given that password should have first two letters followed by four numbers and there are a total of 26 upper case alphabets and 10 digits $(0, 1, 2, \ldots, 9)$ which can be used for the password.
Since the repetition is not allowed, we can use any of the 26 alphabets for the first place. For the second place we can use any of the 25 alphabets (as one alphabet is already used for the first place). Now, for the third place we can use any of the 10 digits and for the fourth place we can use any of the 9 digits (as one digit is already used for the third place) and so on.
Thus, by the multiplication rule of counting, total number of ways $= 26 \times 25 \times 10 \times 9 \times 8 \times 7 = 3,276,000$.
Hence, there are total $3,276,000$ ways in which password can be created.

Q3. There are eight athletes who take part in a 100 m race. What are the possible ways the athletes can finish the race (assuming no ties)?

Solution:
There are a total of 8 people who are participating in the race, so we will have eight different positions of finishing the race.

For the first place, we can have any of the 8 athletes. For the second place, we can have anyone from the remaining 7 athletes, and so on.

Thus, by multiplication rule of counting, total number of ways is

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40,320$$

Hence, there are a total of $40,320$ ways in which athletes can finish the race.

6.1.3 Unsolved Problems:

Q1. Narendra is going to purchase some sports items at a sport shop. He has the choice of cricket bats from MRF, Spartan, SS or Mongoose, choice of cricket balls from SG, Acorn, Jaspo or Kookaburra, choice of cricket stumps from 8 brands, choice of cricket jerseys from 5 brands and choice of a sport shoes from 10 brands. How many ways can Narendra purchase the sports items? (Assume that he purchases only one item from either of all categories)

Q2. Narendra is going to purchase some sports items at a sport shop. He has the choice of cricket bats from MRF, Spartan, SS or Mongoose, choice of cricket balls from SG, Acorn, Jaspo or Kookaburra, choice of cricket stumps from 8 brands, choice of cricket jerseys from 5 brands and choice of a sport shoes from 10 brands. How many ways can Narendra purchase the sports items? (Assume that he purchases only one item from each category)

Q3. In a class, there are 60 students out of which the cricket captain and class-representative needs to be elected. A student can take only one position at a time. What are the total number of possible ways in which students can be elected for these positions?

Q4. In a class, there are 60 students out of which the cricket captain and class-representative needs to be elected. A student can take both the positions of cricket captain and position of class-representative at the same time. What are the total number of possible ways in which students can be elected for these positions?

Q5. How many words of five letters word can be formed using lower case alphabets such that the words start with vowels and end with alphabet A? (Assume that words can be meaningless and repetition is not allowed)

Chapter : 7

7 Factorial

7.1 Definition

The product of the first n positive integers (counting numbers) is called n factorial and is denoted $n!$. In symbols,

$$n! = n \times (n-1) \times \ldots \times 1$$

Remark:
By convention $0! = 1$

Example: There are eight athletes who take part in a 100 m race. What are the possible ways the athletes can finish the race (assuming no ties)?

Solution:
There are a total of 8 people who are participating in the race, so we will have eight different positions of finishing the race.
For the first place, we can have any of the 8 athletes. For the second place, we can have anyone from the remaining 7 athletes, and so on.
Thus, total number of ways $= 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 8!$ (by definition of factorial).

Note:

$$n! = n \times (n-1)!$$

For example:
If $n = 5$, then $5! = 5 \times 4!$

In general:
For $i \leq n$ we have,

$$n! = n \times (n-1) \times \ldots \times (n-i+1) \times (n-i)!$$

For example:
$5! = 5 \times 4! = 5 \times 4 \times 3!$

7.1.0.1 Simplifying expressions:

1. $\dfrac{6!}{3!} = \dfrac{6 \times 5 \times 4 \times 3!}{3!} = 6 \times 5 \times 4 = 120.$
2. $\dfrac{6! \times 5!}{3! \times 4!} = \left(\dfrac{6 \times 5 \times 4 \times 3!}{3!}\right)\left(\dfrac{5 \times 4!}{4!}\right) = 6 \times 5 \times 4 \times 5 = 600.$
3. Express $25 \times 24 \times 23$ in terms of factorials:

$$\frac{25 \times 24 \times 23 \times 22 \times \ldots \times 1}{22 \times 21 \times \ldots \times 1} = \frac{25!}{22!}$$

7.1.0.2 Unsolved Problems:

Q1. If $\frac{(n-1)!}{(n-3)!} = 6$, then find the value of n?

Q2. Express $\frac{(7 \times 6)}{(4 \times 3)}$ in terms of Factorials.

Q3. If $\frac{1}{(n-4)!} = \frac{20}{(n-2)!}$, then calculate the value of n?

Q4. Find the value of the expression $\frac{6 \times 5 \times 4!}{4}$.

Chapter : 8

8 Permutation

8.1 Definition

A permutation is an ordered arrangement of all or some of n objects.

8.1.0.1 Solved Examples:

1. What are all the possible arrangements of A, B and C when we take all of them at a time?

   Solution:

First place	Second place	Third place
A	B	C
A	C	B
B	A	C
B	C	A
C	A	B
C	B	A

2. What are all the possible arrangements of A, B and C when we take two of them at a time?

   Solution:

First place	Second place
A	B
A	C
B	A
B	C
C	A
C	B

3. What are all the possible arrangements of A, B, C and D when we take all of them at a time?

   Solution:

   If we fix A in the first place, then we have to arrange B, C and D at the remaining three places i.e. second, third and the fourth place. And, the number of ways in which we can fill these three places with B, C and D is 6. So, if we fix A in the first place, we get 6 ways of getting hold of different possible arrangements.

   Now, if we fix B in the first place, then we have to arrange A, C and D at the remaining

three places. And, number of ways to fill these three places is 6.
Similarly, if we fix C in the first place, then we have to arrange A, B and D at the remaining three places. And again, there are 6 ways for the same.
If we fix D in the first place, we have to arrange A, B and C at the remaining three places. And there are again 6 ways for the same. Finally, we have total of $6+6+6+6 = 24$ arrangements.

Also, we can list all the arrangements as follows:

First place	Second place	Third place	Fourth place
A	B	C	D
A	B	D	C
A	C	B	D
A	C	D	B
A	D	B	C
A	D	C	B
B	A	C	D
B	A	D	C
B	C	A	D
B	C	D	A
B	D	A	C
B	D	C	A
C	A	B	D
C	A	D	B
C	B	A	D
C	B	D	A
C	D	A	B
C	D	B	A
D	A	B	C
D	A	C	B
D	B	A	C
D	B	C	A
D	C	A	B
D	C	B	A

4. What are all the possible arrangements of A, B, C and D when we take two of them at a time?

   Solution:

First place	Second place
A	B
A	C
A	D
B	A
B	C
B	D
C	A
C	B
C	D
D	A
D	B
D	C

8.2 Permutation formula

8.2.1 When repetition is not allowed.

The number of possible permutations of r objects from a collection of n distinct objects is given by the formula

$$n \times (n-1) \times \ldots \times (n-r+1)$$

and is denoted by ${}^nP_r$.

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

Special cases

1. ${}^nP_0 = \dfrac{n!}{(n-0)!} = \dfrac{n!}{n!} = 1$. There is only one ordered arrangement of 0 objects.

2. ${}^nP_1 = \dfrac{n!}{(n-1)!} = n$. There are n ways of choosing one object from n objects.

3. ${}^nP_n = \dfrac{n!}{(n-n)!} = \dfrac{n!}{0!} = n!$. We can arrange n distinct objects in $n!$ ways - multiplication principle of counting.

8.2.1.1 Solved examples by using permutation formula:

5. What are all the possible arrangements of A, B and C when we take all of them at a time?

   Solution:

First place	Second place	Third place
A	B	C
A	C	B
B	A	C
B	C	A
C	A	B
C	B	A

Here, $n = 3$, $r = 3$. So, ${}^nP_r = \dfrac{n!}{(n-r)!} = \dfrac{3!}{(3-3)!} = \dfrac{3!}{0!} = 3! = 6$. (As $0! = 1$)

6. What are all the possible arrangements of A, B and C when we take two of them at a time?

   Solution:

First place	Second place
A	B
A	C
B	A
B	C
C	A
C	B

   Here, $n = 3$, $r = 2$. So, ${}^nP_r = \dfrac{n!}{(n-r)!} = \dfrac{3!}{(3-2)!} = \dfrac{3!}{1!} = 6$.

7. What are all the possible arrangements of A, B, C and D when we take all of them at a time?

   Solution:
   Here, $n = 4$, $r = 4$.
   So, ${}^nP_r = \dfrac{n!}{(n-r)!} = \dfrac{4!}{(4-4)!} = \dfrac{4!}{0!} = 4! = 24$.

   Thus, there are a total 24 possible arrangements and all the arrangements are listed in example **3**.

8. What are all the possible arrangements of A, B, C and D when we take two of them at a time?

   Solution:

First place	Second place
A	B
A	C
A	D
B	A
B	C
B	D
C	A
C	B
C	D
D	A
D	B
D	C

Here, $n = 4$, $r = 2$. So, ${}^nP_r = \dfrac{n!}{(n-r)!} = \dfrac{4!}{(4-2)!} = \dfrac{4!}{2!} = \dfrac{4 \times 3 \times 2!}{2!} = 12$.

8.2.1.2 Example: Application

1. From a committee of 8 persons, in how many ways can we choose a chairman and a vice chairman assuming one person can not hold more than one position?

   Solution:
   Suppose, A, B, C, D, E, F, G, H, are the 8 persons. A could be a chairman, B could be a vice chairman or B could be a chairman, A could be a vice chairman. So, these two arrangements are different (i.e. order matters).
   Here, we have 8 persons ($n = 8$) in total and we want to know number of ways in which we can choose a chairman and vice chairman. So, $r = 2$.
   Thus, total number of ways $= {}^8P_2 = \dfrac{8!}{(8-2)!} = \dfrac{8!}{6!} = \dfrac{8 \times 7 \times 6!}{6!} = 8 \times 7 = 56$.

2. Find the number of 4-digit numbers that can be formed using the digits 1, 2, 3, 4, 5 if no digit is repeated.

   Solution:
   Since no digit is repeated. Therefore, we have $n = 5$ distinct digits in total. Also, we have to form number of 4-digit numbers (i.e. we have 4 blanks). So, $r = 4$.
   Total number of ways $= {}^5P_4 = \dfrac{5!}{(5-4)!} = \dfrac{5!}{1!} = 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$.

   • **How many of these will be even?**
   If we want to count the number of even numbers, then last digit can either be 2 or 4. We will fix the last digit here. If we fix the last digit, we have 3 blank spaces(i.e. we have to choose 3 digits).
   If we fix the last digit as 2, then we have digits 1, 3, 4 and 5 in total, i.e., $n = 4$ and we have to choose 3 digits from these 4 digits.

And, that can be done in ${}^4P_3 = \dfrac{4!}{(4-3)!} = \dfrac{4!}{1!} = 4! = 4 \times 3 \times 2 \times 1 = 24$ ways.

Similarly, if we fix the last digit as 4, then we have digits 1, 2, 3 and 5 in total, i.e., $n = 4$ and we have to choose 3 digits from these 4 digits.

And, that can be done in ${}^4P_3 = \dfrac{4!}{(4-3)!} = \dfrac{4!}{1!} = 4! = 4 \times 3 \times 2 \times 1 = 24$ ways.

Therefore, total number of ways $= 24 + 24 = 48$.

3(i). Six people go to the cinema. They sit in a row with ten seats. Find how many ways can this be done if they can sit anywhere.

Solution:

We have 10 available seats and 6 people can sit anywhere in the row of ten seats. Thus, $n = 10$ and $r = 6$.

Now, total number of arrangements $= {}^{10}P_6 = \dfrac{10!}{(10-6)!} = \dfrac{10!}{4!} = 151200$.

Hence, six people can sit in a row with ten seats in 151200 ways when they can sit anywhere.

3(ii). Six people go to the cinema. They sit in a row with ten seats. Find how many ways can this be done if all the empty seats are next to each other.

Solution:

We have 10 seats in total. Since, six people have to sit in a row of 10 seats. Therefore, we will have 4 seats as empty.

Now, the condition is that all four empty seats are next to each other. So, we can consider these four empty seats as a single block. If these 4 empty seats are considered as 1 distinct object, then the total number of distinct objects are 1, 2, 3, 4, 5, 6 and 7. So, we have 7 places available and we have to set 6 people in these 7 places.

Thus, total number of ways$= {}^7P_6 = \dfrac{7!}{(7-6)!} = \dfrac{7!}{1!} = 7! = 5040$ ways.

Hence, six people can sit in a row with ten seats in 5040 ways when all the empty seats are next to each other.

8.3 Permutation formula

8.3.1 When repetition is allowed.

The number of possible permutations of r objects from a collection of n distinct objects when repetition is allowed is given by the formula

$$n \times n \times \ldots \times n$$

and is denoted by n^r.

8.3.1.1 Solved examples:

1. What are all the possible arrangements of A, B and C when we take all of them at a time, if repetition is allowed?

   Solution:
   Since repetition is allowed. Therefore, first place can be filled with A or B or C, second place can be filled with A or B or C, third place can be filled with A or B or C. So, there are 3 choices available for the first place, 3 choices available for second place and 3 choices available for the third place, because all the choices are available for all the places. So, the total number of arrangements $= 3 \times 3 \times 3 = 27$.

   Also, we can list all the arrangements as follows;

First place	Second place	Third place
A	A	A
A	A	B
A	A	C
A	B	A
A	B	B
A	B	C
A	C	A
A	C	B
A	C	C
B	A	A
B	A	B
B	A	C
B	B	A
B	B	B
B	B	C
B	C	A
B	C	B
B	C	C
C	A	A
C	A	B
C	A	C
C	B	A
C	B	B
C	B	C
C	C	A
C	C	B
C	C	C

2. What are all the possible arrangements of A, B and C when we take two of them at a time, if repetition is allowed?

   Solution:
   Since, we have A, B and C in total and we have to find all the possible arrangements by taking two of them with repetition. Therefore, $n = 3$ and $r = 2$.
   Thus, all the possible arrangements $= n^r = 3^2 = 9$.

   Also, we can list all the arrangements as follows;

First place	Second place
A	A
A	B
A	C
B	A
B	B
B	C
C	A
C	B
C	C

8.4 Permutation formula

8.4.1 Rearranging letters

(1). The number of permutations of n objects when p of them are of one kind and rest distinct is equal to $\frac{n!}{p!}$.

For example:
Suppose we want to rearrange the letters in the word "DATA". How many ways can it be done?
Solution:
In "DATA", we have 4 objects and 2 of them, i.e., A which are of one kind. Thus, we have $n = 4$ and $p = 2$.
Hence, the total number of ways the letters in "DATA" can be arranged is $\frac{4!}{2!} = \frac{4 \times 3 \times 2!}{2!} = 4 \times 3 = 12$.

Hence, all the possible arrangements can be listed as follows:

First place	Second place	Third place	Fourth place
A	D	T	A
A	D	A	T
A	T	D	A
A	T	A	D
A	A	D	T
A	A	T	D
D	A	T	A
D	A	A	T
D	T	A	A
T	A	D	A
T	A	A	D
T	D	A	A

(2). The number of permutations of n objects where p_1 is of one kind, p_2 is of second kind, and so on p_k of k^{th} kind is given by

$$\frac{n!}{p_1!p_2!\dots p_k!}$$

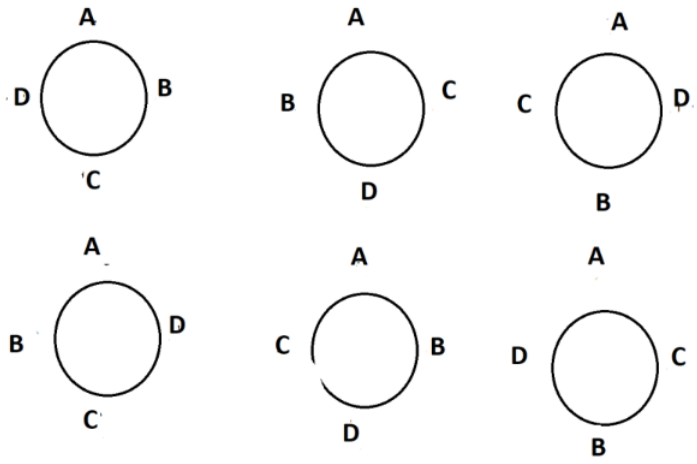
For example:
Suppose we want to rearrange the letters in the word "STATISTICS". How many ways can it be done?
Solution:
Total of ten letters of which there are five distinct letters: S, T, A, I, C. Also, "S" appears 3 times, "T" appears 3 times, "A" once, "I" twice, and "C" once.
Now, applying the formula to the word "STATISTICS", we get
$n = 10, p_1 = 3, p_2 = 3, p_3 = 1, p_4 = 2, p_5 = 1.$
Hence, total number of ways $= \dfrac{10!}{3!3!1!2!1!} = 50,400.$

8.5 Circular Permutation

8.5.1 Clockwise and anticlockwise are different

The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are different) is equal to $(n-1)!$.

For Example:
How many ways can four people sit in a round table?
Solution:
Consider four people are A, B, C and D and they are distinct. Now, arrangement of these 4 people in a circle (when clockwise and anticlockwise are different) can be shown as:

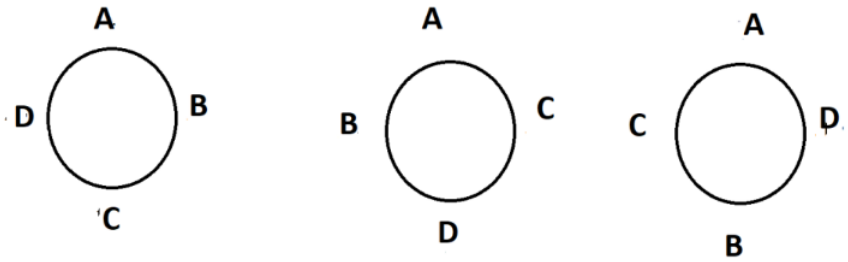


Thus, we have total 6 ways for circular arrangement (when clockwise and anticlockwise are different) of four distinct people.
Also, we have $n = 4$. So, by definition, number of ways 4 distinct people can be arranged in a circle (clockwise and anticlockwise are different) is equal to $(n-1)! = (4-1)! = 3! = 6$.

8.5.2 Clockwise and anticlockwise are same

The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are same) is equal to $\frac{(n-1)!}{2}$.

For Example:
How many ways can four people sit in a round table when clockwise and anticlockwise are same?
Solution:
Consider four people are A, B, C and D and they are distinct. Now, arrangement of these 4 people in a circle (when clockwise and anticlockwise are same) can be shown as:



Thus, we have total 3 ways for circular arrangement (when clockwise and anticlockwise are same) of four distinct people.
Also, we have $n = 4$. So, by definition, number of ways 4 distinct people can be arranged in a circle (clockwise and anticlockwise are same) is equal to $\frac{(n-1)!}{2} = \frac{(4-1)!}{2} = \frac{3!}{2} = \frac{6}{2} = 3$.

8.5.2.1 Examples of calculating n and r.

1. Find value of n if ${}^nP_4 = 20\,{}^nP_2$.

   Solution:
   We are given that, ${}^nP_4 = 20\,{}^nP_2$
   Now, apply the formula of permutation, we get

$$\frac{n!}{(n-4)!} = 20 \times \frac{n!}{(n-2)!}$$
$$\implies \frac{1}{(n-4)!} = 20 \times \frac{1}{(n-2)!}$$
$$\implies \frac{(n-2)!}{(n-4)!} = 20$$
$$\implies \frac{(n-2)(n-3)(n-4)!}{(n-4)!} = 20$$
$$\implies (n-2)(n-3) = 20$$
$$\implies n^2 - 5n + 6 = 20$$
$$\implies n^2 - 5n - 14 = 0$$
$$\implies n^2 - 7n + 2n - 14 = 0$$
$$\implies n(n-7) + 2(n-7) = 0$$
$$\implies (n-7)(n+2) = 0$$
$$\implies n = 7 \quad or \quad n = -2$$

   As n cannot take negative values. Therefore, value of $n = 7$.

2. If $\frac{{}^nP_4}{{}^{n-1}P_4} = \frac{5}{3}$, then find the value of n.

   Solution:
   We are given that,

$$\frac{{}^nP_4}{{}^{n-1}P_4} = \frac{5}{3}$$
$$\implies {}^nP_4 = {}^{n-1}P_4 \times \left(\frac{5}{3}\right)$$
$$\implies \frac{n!}{(n-4)!} = \frac{(n-1)!}{(n-5)!} \times \left(\frac{5}{3}\right)$$
$$\implies \frac{n \times (n-1)!}{(n-4) \times (n-5)!} = \frac{(n-1)!}{(n-5)!} \times \left(\frac{5}{3}\right)$$
$$\implies \frac{n}{(n-4)} = \frac{5}{3}$$

$$\implies 3n = 5n - 20$$
$$\implies 20 = 2n$$
$$\implies 2n = 20$$
$$\implies n = 10.$$

Hence, value of n is 10.

3. If ${}^5P_r = 2 \times {}^6P_{r-1}$, then find the value of r.

Solution:

We are given that,

$${}^5P_r = 2 \times {}^6P_{r-1}$$
$$\implies \frac{5!}{(5-r)!} = 2 \times \frac{6!}{(6-(r-1))!}$$
$$\implies \frac{5!}{(5-r)!} = 2 \times \frac{6!}{(6-r+1)!}$$
$$\implies \frac{5!}{(5-r)!} = 2 \times \frac{6!}{(7-r)!}$$
$$\implies \frac{5!}{(5-r)!} = 2 \times \frac{6 \times 5!}{(7-r) \times (6-r) \times (5-r)!}$$
$$\implies 1 = 2 \times \frac{6}{(7-r) \times (6-r)}$$
$$\implies (7-r) \times (6-r) = 2 \times 6$$
$$\implies r^2 - 13r + 42 = 12$$
$$\implies r^2 - 13r + 30 = 0$$
$$\implies r^2 - 10r - 3r + 30 = 0$$
$$\implies r(r-10) - 3(r-10) = 0$$
$$\implies (r-10)(r-3) = 0$$
$$\implies r = 10 \quad or \quad r = 3$$

Since $r \leq n$, therefore $r = 10$ is eliminated and we get $r = 3$.

8.6 Unsolved Problems:

Q1. A teacher is creating a quiz paper of 11 questions from a test bank of 20 questions. In how many ways can he select and arrange the questions?

Q2. A box contains 3 distinct white balls, 4 distinct black balls, and 3 distinct red balls. Find the number of ways in which three balls can be drawn from the box so that all the three drawn balls will have different colours.

Q3. In how many ways can we order (in line) the 26 letters of English alphabet so that no two vowels (a, e, i, o, and u) occur consecutively ?

Q4. Find the number of rearrangements of the letters in the word "CRICKET".

Q5. Seven people are going to sit at a round table. How many different ways can this be done?

Q6. If ${}^{n+2}P_3 = 6 \times {}^{n}P_1$, then find the value of n?

Chapter : 9

9 Combination

9.1 Definition

The number of possible combinations of r objects from a collection of n distinct objects is denoted by ${}^nC_r$ and is given by

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

Another common notation is $\binom{n}{r}$ which is also referred to as the binomial coefficient.

Note: Number of possible combinations of r objects from a number collection of n distinct object is denoted by ${}^nC_r$. Since, each combination of r objects from n objects give rise to $r!$ arrangements. So, ${}^nC_r \times r! = {}^nP_r \implies {}^nC_r = \frac{{}^nP_r}{r!} \implies {}^nC_r = \frac{n!}{r!(n-r)!}$ which is the formula of ${}^nC_r$.

• **Some useful results**

1. ${}^nC_r = \frac{n!}{r!(n-r)!} = \frac{n!}{(n-r)!r!} = {}^nC_{(n-r)}$

In other words, selecting r objects from n objects is the same as rejecting $n-r$ objects from n objects.

2. ${}^nC_n = 1$ and ${}^nC_0 = 1$ for all values of n.
3. ${}^nC_r = {}^{n-1}C_{r-1} + {}^{n-1}C_r \ ; 1 \leq r \leq n$

9.1.0.1 Solved Examples:

1. How many ways can we select two students from a group of three students?
   Solution:
   Let A, B, and C be the three students. We can select AB, AC, BC or BA, CA, CB. Both of the selections are same because in this case, the concern is only which of the 2 objects are chosen and not in the order in which they are chosen.
   Thus, we have total 3 ways in which we can select two students from a group of three students.
   Also, we can find the same by using combination formula as ${}^3C_2 = \frac{3!}{2! \times (3-2)!} = 3.$

2. In an examination, a question paper consists of 12 questions divided into two parts i.e., Part I and Part II, containing 7 and 5 questions, respectively. A student is required to attempt 8 questions in all, selecting at least 3 from each part. In how many ways can a student select the questions ?
   Solution:
   We are given that a student has to attempt 8 questions in all and there are two parts, part I consists of 7 and part II consists of 5 questions.
   Since, the condition is he has to choose at least 3 from each part. Therefore, he can

choose the questions in the following manner:
(i) 3 questions from part I and 5 questions from part II.
(ii) 4 questions from part I and 4 questions from part II.
(iii) 5 questions from part I and 3 questions from part II.
He cannot choose 6 questions from part I and 2 questions from part II because condition will be violated.
Now,
(i) As 3 questions can be chosen out of 7 questions in ${}^7C_3$ ways and 5 questions can be chosen out of 5 questions in ${}^5C_5$ ways. So, ${}^7C_3 \times {}^5C_5$ is the total number of ways in which student can choose 3 questions from part I and 5 questions from part II.
(ii) As 4 questions can be chosen out of 7 questions in ${}^7C_4$ ways and 4 questions can be chosen out of 5 questions in ${}^5C_4$ ways. So, ${}^7C_4 \times {}^5C_4$ is the total number of ways in which student can choose 4 questions from part I and 4 questions from part II.
(iii) As 5 questions can be chosen out of 7 questions in ${}^7C_5$ ways and 3 questions can be chosen out of 5 questions in ${}^5C_3$ ways. So, ${}^7C_5 \times {}^5C_3$ is the total number of ways in which student can choose 5 questions from part I and 3 questions from part II.
Total number of ways student can select questions is

$$= \left({}^7C_3 \times {}^5C_5\right) + \left({}^7C_4 \times {}^5C_4\right) + \left({}^7C_5 \times {}^5C_3\right)$$

$$= \left(\frac{7!}{3! \times 4!}\right) \times \left(\frac{5!}{5! \times 0!}\right) + \left(\frac{7!}{4! \times 3!}\right) \times \left(\frac{5!}{4! \times 1!}\right) + \left(\frac{7!}{5! \times 2!}\right) \times \left(\frac{5!}{3! \times 2!}\right)$$

$$= 35 + 175 + 210 = 420.$$

3(a). In how many ways you can choose 4 cards from a pack of 52 playing cards ?
Solution:
A "standard" deck of playing cards consists of 52 Cards in each of the 4 suits of Spades, Hearts, Diamonds, and Clubs. Each suit contains 13 cards: Ace, $2, 3, 4, 5, 6, 7, 8, 9, 10$, Jack, Queen, King. And, Hearts and Diamonds are red faced cards whereas Spades and clubs are black faced card. Thus, there are 26 red cards and 26 black cards in a pack of 52 cards. Also, Jack, Queen and King are known as face cards.
Now, we have to choose just 4 cards from a pack from 52 cards, without placing any other condition, which can be done in ${}^{52}C_4 = \dfrac{52!}{4! \times 48!} = 270725$ ways.

3(b). In how many ways you can choose 4 cards from a pack of 52 playing cards if all four cards are of same suit ?
Solution:
Since, selection of all four cards should be from same suit. Therefore, first we have to choose one suit out of four suits and then, have to choose the 4 cards from that suit.
Now, we can choose one suit out of four suits in ${}^4C_1$ ways and within each suits we have 13 of each kind and we need to choose 4 cards from 13 cards and that can be done in ${}^{13}C_4$ ways. So, the total number of ways we can choose all the 4 cards from the same suits is ${}^4C_1 \times {}^{13}C_4 = 2860$ ways.

3(c). In how many ways you can choose 4 cards from a pack of 52 playing cards if all four cards are of same colour?

Solution:

As we know that there are cards of two colour (red and black) in a pack of 52 cards. First, we have to choose 1 colour from 2 colours which can be done in ${}^2C_1$ ways. Now, within each colour we have 26 cards and we need to choose 4 from each colour which can be done in ${}^{26}C_4$ ways.

Hence, total number of ways in which we can choose 4 cards of the same colour are ${}^2C_1 \times {}^{26}C_4 = 29900$.

4. Select a cricket team of eleven from 17 players in which only 5 players can bowl. The requirement is the cricket team of 11 must include exactly 4 bowlers. How many ways can the selection be done?

Solution:

Total number of players available for selection is 17 in which 5 are bowlers. The requirement is the there should be exactly 4 bowlers, so we need to select 4 bowlers out of 5 which can be done in ${}^5C_4$ ways. Now, remaining 7 players can be selected from remaining 12 players in ${}^{12}C_7$ ways.

Thus, total number of ways the selection can be done is ${}^5C_4 \times {}^{12}C_7 = 5 \times 792 = 3960$.

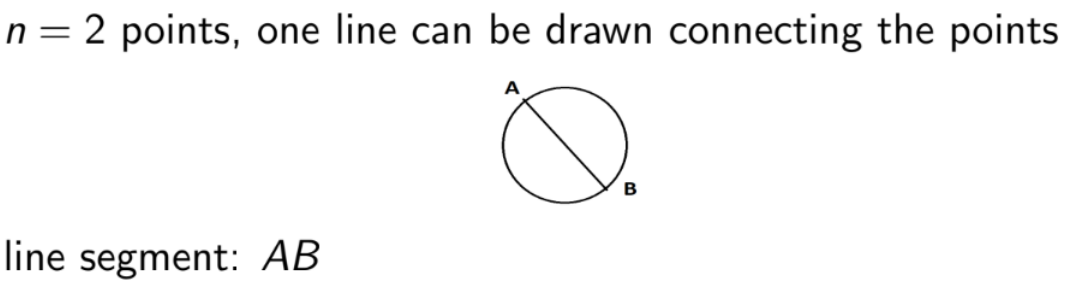
9.1.1 Drawing lines in a circle

Given n points on a circle, number of line segments that can be drawn connecting the points is ${}^nC_2$.

1. If 2 points are given on a circle, then how many lines can be drawn connecting these points ?

   Solution:

   We are given that 2 points on a circle, i.e, $n = 2$. So, number of line segments that can be drawn connecting the points is ${}^2C_2 = 1$, which can be illustrated as follows:

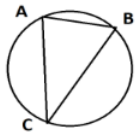


2. If 3 points are given on a circle, then how many lines can be drawn connecting these points ?

   Solution:

We are given that 3 points on a circle, i.e, $n = 3$. So, number of line segments that can be drawn connecting the points is ${}^3C_2 = 3$, which can be illustrated as follows:

$n = 3$ points, three line can be drawn connecting the points



line segments: AB, AC, and BC

Note: The point to distinguish between situations involving combinations and situations involving permutations is, we have to use Permutation when "order matters" and we have to use Combination when "order does not matter".

9.1.1.1 Some more solved examples on permutation and combination:

5(a). Consider the situation of eight athletes participating in a 100m race in a competition with several rounds. How many different ways can you award the Gold, Silver, and Bronze medals?

Solution:

Suppose 8 athletes are A, B, C, D, E, F, G and H. We have to award the Gold, Silver and Bronze medals among these 8 athletes. Here order matters in awarding the medals. Because, order of awarding Gold medal to A, order of awarding Silver medal to B and order of awarding Bronze medal to C is different from order of awarding Gold medal to C, order of awarding Silver medal to A and order of awarding Bronze medal to B. So, we need permutation. Thus, number of ways in which we can award the Gold, Silver, and Bronze medals is ${}^8P_3 = 336$.

5(b). Consider the situation of eight athletes participating in a 100m race in a competition with several rounds. How many different ways can you choose the top three athletes to proceed to the next round in the competition?

Solution:

We have to choose the top three athletes to proceed to the next round in the competition. Now, suppose E, F and G are top three athletes then it does not matter who came at first, second or third position. All three athletes will proceed to the next round in the competition.

Similarly, top three athletes can be D, G and H or A, E and F and so on. Hence, order does not matter. So, we need to use combination.

Thus, number of ways in which we can choose the top three athletes out of eight athletes is ${}^8C_3 = 56$.

6(a). Consider the situation of a class with forty students. In how many different ways can we choose two leaders?

Solution:

We have to just choose two students out of 40 students for being leader. Here, order is not important. So, we need to use combination.
Therefore, the number of ways in which we can choose two leaders is ${}^{40}C_2 = 780$.

6(b). Consider the situation of a class with forty students. In how many different ways can we choose a captain and vice captain?

Solution:

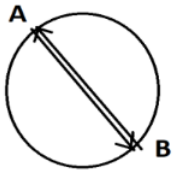
We have to choose a captain and a vice captain. Here, order is important. So, we need to use permutation.
Therefore, the number of ways in which we can choose a captain and vice captain is ${}^{40}P_2 = 1560$.

• Important Note

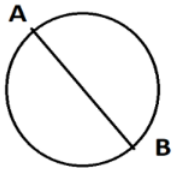
Given n points on a circle, how many lines can be drawn connecting these points?

Solution:

1. If the segment has a direction line segment AB is different from BA. Order is important. Hence, total number of ways is ${}^{n}P_2$.
Also, we can illustrate the same in figure as follows:



2. If segment has no direction. Line segment AB. Order is not important. Hence, total number of ways is ${}^{n}C_2$.
Also, we can illustrate the same in figure as follows:



9.2 Unsolved Problems:

Q1. If ${}^nC_2 = {}^nC_3$, then calculate the value of n?

Q2. Out of a group of six men and eight women, you need to form a committee of three men and five women. In how many ways can the committee be formed?

Q3. Out of six consonants and four vowels, how many words of three consonants and two vowels can be formed?

Q4. 20 points are chosen in the plane so that no three of them are collinear. How many triangles do they determine?

Answers of Unsolved Problems

Chapter : 6

1. 31
2. 6400
3. 3540
4. 3600
5. 48576

Chapter : 7

1. $n = 4$
2. $\frac{7! \times 2!}{5! \times 4!}$
3. $n = 7$
4. 180

Chapter : 8

1. $\frac{20!}{9!}$
2. 36
3. $21! \times \frac{22!}{17!}$
4. 2520
5. 720
6. $n = 1$

Chapter : 9

1. 5
2. 1120
3. 14400
4. ${}^{20}C_3$