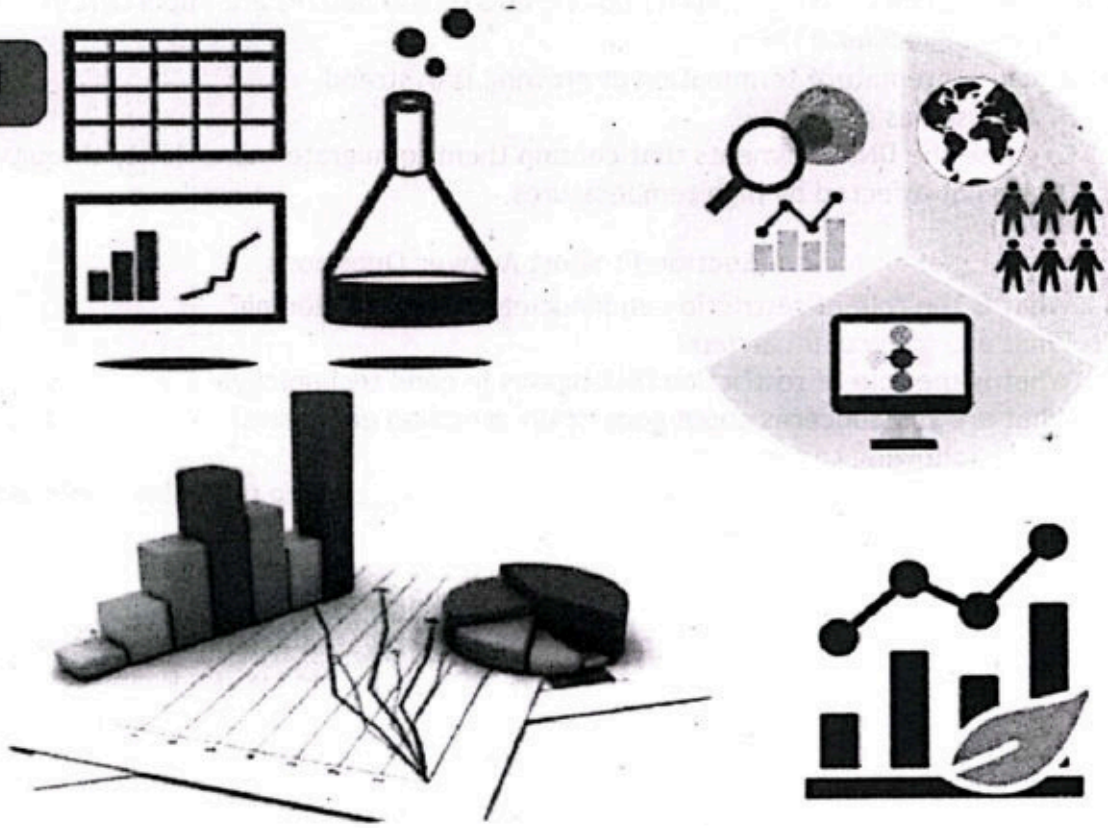
Chapter 11



BIOSTATISTICS AND DATA HANDLING

Students' learning outcomes
After studying this chapter, students will be able to:

1. [B-12-K-01] Define biostatistics and its use.
2. [B-12-K-02] Define mean, median, mode, standard deviation, range, percentile.
3. [B-12-K-03] Calculate mean, median, mode, standard deviation, range, percentile from a given set of data.
4. [B-12-K-04] Sketch a bar chart for a given set of data.
5. [B-12-K-05] Sketch error bars based off of range or standard deviation for a given set of data on a bar chart.
6. [B-12-K-06] Evaluate the appropriate type of figure or chart for a given set of data and/or experiment (bar chart, pie chart, xy axis data figure etc).
7. [B-12-K-07] Make the appropriate chart with proper title, labeled axes, legend, axes units.
8. [B-12-K-08] Design an appropriate experiment with control group and dependent, independent and control variables.

11.1 BIOSTATISTICS AND ITS USES

11.1.1 Biostatistics

Biostatistics is the application of statistical principles and techniques to analyse, interpret and present the findings of research in the fields of biology, medicine and health sciences. Biostatistics mainly deals with the data associated with living organisms and then presentation of findings in an appropriate way for possible plans or solutions.

Biostatistics is the application of statistical principles and techniques to analyze, interpret, and present research findings in the fields of biological, medical, and health sciences. It primarily deals with data associated with living organisms, presenting findings to support effective plans or solutions. Biostatistics is applicable to various fields, including medicine, public health, and agriculture. It helps in calculating and assessing agricultural, poultry, and dairy farming outcomes, such as crop yields and livestock growth. Biostatistics is an essential part of modern health sciences and is frequently used in areas like medicine, epidemiology, and clinical research. It plays a key role in understanding disease outbreaks, health trends, possible preventive measures, and the effectiveness of specific disease treatments.

Data

In biostatistics, data refers to the facts, measurements or observations related to living organisms collected to analyze and draw conclusions about biological phenomena. It includes the values such as heights, weights, enzyme activity levels, population counts or rate of any disease.

Data set

Data set is a collection of related data values, usually organized systematically for analysis. It usually includes multiple measurements or observations, organized in a table. Each row typically represents a subject or experimental unit and each column represents a variable. Each entry in a data set represents an individual sample and each variable describes a particular characteristic being measured.

Example: A table recording the heights and weights of 100 plants grown under different availability of light or moisture conditions is a data set.

Components of Biostatistics studies

For effective application of biostatistics, following components (steps) are followed:

Identification of problem: It is first ever requirement to solve any problem affecting the living organisms especially human.

Collecting and analyzing data: Data collected before experiments and then outcome of the experiments is analyzed thoroughly.

Designing experiments: For the solution of any problem associated to living organisms, biologist use biostatistics to plan biological experiments specifically the requirements and durations.

Interpreting results: After experiments, the results are used to interpret the reasons and then possible future effects of biological problem are estimated.

Developing new tools: Interpretations of experimental results lead to creating new predictive tools and plans for effectively dealing the problem e.g. health related issues.

It is highly recommended for every biologist to have basic learning or understanding of biostatistics for effective dealing to biological problems.

11.1.2 Uses of Biostatistics

1. Need Assessments of Agriculture and Cattle Farming for Growing Population

Biostatistics is involved in the assessment of demands of agriculture and dairy farming according

to the rate of population growth. Furthermore it also help the government to assess the need for import or export of food items as per growing population.

2. Evaluation of Efficacy of Medical Instruments and Pharmacological Drugs

Biostatistics helps to design clinical trials and controlled studies to assess the efficacy and safety of newly designed drugs, treatment plan or medical instrument. Findings of medical research guide whether treatment resulted to improvements or side effects. Assessment of the effectiveness and safety of new drugs and deciding optimal doses for any new pharmacological drug.

3. Monitoring of Epidemiological Studies and Policy Development

Biostatistics help in monitoring and analysis of data collected from population about the spread of any epidemic disease. Researchers can identify the risk factors, pattern and rate of disease. Epidemiological studies by using biostatic tools help to control and prevent future outbreaks (e.g., prevalence of Hepatitis and Polio, spread of COVID-19 etc.). Biostatistics evidence helps governments and public organizations to make decisions about population management, healthcare plans and funding sources.

4. Management of Public Health and Population Growth

Each country conduct biostatics studies to estimate health trends within its population. Routine studies include data about birth rates, death rates and prevalence of different disease. It guides the government in planning health initiatives and allocation of resources for public. Biostatistics is useful to monitors and optimize the hospital performance by calculating patient number, availability of doctors and medicine and effectiveness of treatment.

5. Genetic Diseases

Biostatistics is also useful to analyse the data about inheritance patterns of genetic diseases in any population. For example, number of thalassemia and muscular dystrophy patients. It helps in guessing risk factors and behaviour of genetic disorders.

6. Pollution Indicators and Environmental Protection

Biostatistics is used to analyse the pollution level, its causes and impact on population health. Different biostatic tools are designed to monitor the pollution level and associated potential health risks. It helps in designing policies to reduce health risks from environmental hazards. On the basis of findings, multiple steps are taken to protect the environment. For example, causes of smog, identification of affected areas, possible solution like plantation derives in Punjab especially in Lahore.

7. Survival Analysis

Biostatistics is used to predict and then note the survival rate of patients after a particular treatment in any specific disease. It helps to estimate life expectancy and the chances of success for medical treatments. For example, five years survival after the treatment of cancer is considered a successful treatment. Data of survival rate indicate the effectiveness of health plans.

11.2 DEFINITION AND CALCULATION OF MEAN, MEDIAN, MODE, STANDARD DEVIATION, RANGE AND PERCENTILE

In our daily life routine, we commonly hear the statements like:

i. Heartbeat of human is 72 beats per minute

ii. Death rate due to cancer in Pakistan is 25 per thousand patients.

iii. Production of wheat in Punjab is 2000 Kg per acer.

iv. The price of the bananas in the market is Rs. 120 per dozen.

v. Food consumption of a human is 1kg per day.

vi. The rain fall in Islamabad is 1500mm per year.

If we think on about above mentioned statements, none of them is found exactly correct. In statement no. "i", our heart may not be same 72 in each minute of the day. During running it may be 90 per minute and during sleep it may be 60 per minute. Our heartbeat is approximately 72 in each minute. As given in statement "ii" It is quite possible that in one hospital only 5 cancer patient died out of 1000 visited for treatment and in other hospital 60 patients died out of 1000 patients gone through treatment. The death rate of cancer patients is about 25 per 1000, may not be same all the year and in every hospital. The production of wheat in one district of Punjab may be 1500 Kg from an acer and it may be 2400 Kg from one acer in another district. Although above statements are not exactly true but still they are very important. Actually, these are approximate statements in specific situations. In terms of statistics, we call such statements as average statements.

In our daily conversation, we make many statements which have some meaning only on average basis. In different fields of life like agriculture, health, poultry, the idea of average is very important. Many experts at national and international level discuss the findings of studies in averages. The average is also called measure of central tendency.

Calculation of Average

Average is a single value which is calculated to represent the whole data. It may be calculated for a data sample of patients or a population of migrating birds etc. The average is a value which expresses the central idea of the observations. There are different ways to represent average of a data in different situations. For representation of a specific data, proper type of average is used by the expert who is calculating the average.

Types of averages

The following types of averages are commonly used:

(i) Arithmetic Mean (ii) Median (iii) Mode (iv) Geometric mean (v) Harmonic mean

Here in this chapter we will study only first three types of averages in detail.

11.2.1 Arithmetic Mean or simply Mean:

Definition

Mean is the sum of all the values of data set divided by the total number of values in the data set. It is the single value which is calculated to represent the whole set of data. The symbol "x□" (read as "x bar") represent the sample mean. The bar above the letter x represents the mean of a set of values.

$$\text{Mean } (\bar{x}) = \frac{\textit{Some of values}}{\textit{Total Number of values}} = \frac{(\Sigma X)}{(n)}$$

Formula

Direct method for un-group data	Direct method for group data
$\bar{X} = \frac{\Sigma x}{n}$ where, x = *Value of data set*	$\bar{X} = \frac{\Sigma fx}{\Sigma f}$ (where, x = *Value of data set and* is calculated as $x = \frac{\textit{lower limit + upper limit}}{2}$ and f= frequency distribution)

In biostatistics, data can be categorized as **grouped data** or **ungrouped data**, depending on how it's organized:

Ungrouped Data: Raw data in the form of individual values that has not been organized into groups or categories. It is suitable for small datasets.

Example: Individual blood pressure readings from 10 patients:

120, 125, 130, 122, 118, 135, 128, 124, 129, 121

Grouped Data: Data that has been organized into classes or intervals with their frequencies. It makes large datasets easier to summarize and analyze.

Example: Blood pressure ranges with number of patients:

110-119 mm/Hg: 2 patients 120-129 mm/Hg: 5 patients 130-139 mm/Hg: 3 patients

Frequency distribution: It is a way of organizing and summarizing data in the form of table or graph to display the number of occurrences (frequencies) of different outcomes or data values in a dataset. A frequency distribution makes it easier to understand and interpret the data by showing how often values occur. It is useful to summarize large data sets, Identify patterns, constructing histograms and other statistical charts

There are two types of Frequency Distributions:

1. Ungrouped Frequency Distribution: Lists each individual value and its frequency.
2. Grouped Frequency Distribution: Groups data into intervals (classes) and shows the frequency for each class.

Example 1(un-group data)

Dataset: Team of world health organization (WHO) planned to assess the prevalence of polio disease in Pakistan. Study to record the polio cases in Pakistan continued for consecutive two years. Number of confirmed Polio patients detected each month of 2022 and 2023 are given in table below. Calculate the mean polio patients per month in each year. Also compare the prevalence of Polio disease in both years.

Sr. no.	Month of sampling	Number of Polio patients in year 2022	Number of Polio patients in year 2023
1	January	9	7
2	February	13	9
3	March	15	13
4	April	19	17
5	May	22	18
6	June	25	17
7	July	20	15
8	August	22	14
9	September	18	11
10	October	13	9
11	November	10	8
12	December	6	6
Total	**12 months**	**192**	**144**

Mean number of Polio patients per month in 2022 = $\frac{\text{Sum of Polio cases in complete year}}{\text{Total number of Months}} = \frac{192}{12} = 16$

Mean number of Polio patients per month in 2023 = $\frac{\text{Sum of Polio cases in complete year}}{\text{Total number of Months}} = \frac{144}{12} = 12$

Difference of Polio patients per month detected in 2022 and 2023= 16 - 12 = 04

Prevalence of Polio disease decreased in 2023 compared to 2022 by 04 patients per month.

Example 2 (Group data)

Data set: United Nation-World Food Programme (UN-WFP) conducted a research to assess the impact of food quality on body mass of 100 students of different age groups. Find the arithmetic mean for the Frequency distribution of body mass of 100 students as shown in table given below.

Mass of student groups before the start of experiment				
Sr. No.	Mass (Lbs)	Frequency (f)	$X = \frac{\text{Lower limit + Upper limit}}{2}$	Frequency x mean of limits (fx)
1	35-39	13	$\frac{35 + 39}{2} = 37$	13 X 37 = 481
2	40-44	15	$\frac{40 + 44}{2} = 42$	15 X 42 = 630
3	45-49	28	$\frac{45 + 49}{2} = 47$	28 X 47 = 1316
4	50-54	17	$\frac{50 + 54}{2} = 52$	17 X 52 = 884
5	55-59	12	$\frac{55 + 59}{2} = 57$	12 X 57 = 684

6	60- 64	10	$\frac{60 + 64}{2} = 62$	10 X 62 = 620
7	65- 69	05	$\frac{65 + 69}{2}$ = 67	05 X 67 = 335
	Total	100		4950

Calculations $= \frac{\sum fx}{\sum f} = \frac{4950}{100}$ = 49.5 Lbs

49.5 Lbs is the mean mass of 100 students of different groups before the start of experiment

11.2.2. Median

Definition

The median is the middle value in a dataset that are arranged in ascending or descending order of magnitude. Median divides the data set into two equal parts. One part of the dataset have the values less than the middle value and the other part of the dataset have values greater than the middle item. It is denoted by $\tilde{x}$, (read as "x tilde" or "x snake"). It is necessary to arrange the values of dataset in ascending or descending order.

The methods of calculating the median are simple and the value of median is not affected by change in extreme values of dataset.

Median for un-group data

- If the number of values is odd, the median is the middle value.

  Median ($\tilde{x}$) = The value of $(\frac{n+1}{2})^{th}$ item in the dataset (where n = number of items or data)
- If the number of values is even, the median is the average of the two middle values.

Example 1: (For the dataset with odd number of values)

Dataset: In an experiment to check the effect of specific fertilizer on the growth of the plants, heights of the plants in the experimental dataset was recorded. The heights (in cm) of 13 plants are given below.

58, 67, 65, 70, 68, 62, 63, 46, 64, 49, 55, 66, 72

Find out the Median value of height

Solution: The median is calculated as follows

The measured heights of plants after arranging in ascending order is as given below:

46, 49, 55, 58, 62, 63, **64**, 65, 66, 67, 68, 70, 72

Number of values in the dataset of observed heights= 13

Median ($\tilde{x}$) = $\frac{n+1}{2} = \frac{13+1}{2} = \frac{14}{2}$= 7 (i.e. value of 7th value in datasheet)

Median ($\tilde{x}$) = 64

Example 1: (For the dataset with odd number of values)

Median ($\tilde{x}$) = $\frac{\left(\frac{n}{2}th\ value\right) + \left(\frac{n}{2}+1\right)}{2}$

Dataset: The number of fruits produced on each of the 16 plants of experimental group are recorded as below:

3, 5, 18, 21, 15, 10, 8, 12, 13, 7, 11, 14, 9, 16, 20, 24

Find out the median value of fruits produced per plant.

Solution

The numbers of fruits produced on 16 plants after arranging the values of dataset in ascending order are given below:

3, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 21, 24

Total number of values in the above given dataset n = 16 (even number)

$$\text{Median } (\bar{x}) = \frac{\left(\frac{n}{2}th\ value\right) + \left(\frac{n}{2}+1\right)}{2} \qquad = \frac{\left(\frac{16}{2}\right) + \left(\frac{16}{2}+1\right)}{2}$$

$$\text{Median } (\bar{x}) = \frac{(8th\ value) + (8th\ value+1)}{2} \qquad = \frac{(12) + (12+1)}{2}$$

$$\text{Median } (\bar{x}) = \frac{12+13}{2} = \frac{25}{2} = 12.5 \text{ fruits per plant}$$

Activity: (to be solved by the students)

Calculate the median of the following values of data set obtained by measuring the heights of 22 plants in an experimental group to assess the effects of less water availability:

8, 9, 10, 18, 10, 8, 9, 7, 8, 11, 23, 19, 17, 15, 7, 12, 14, 12, 11, 14, 15, 13.

2. Median for grouped data

(i) Discontinuous or Unconnected values in dataset

To calculate the value of median for discontinuous grouped data, first of all the cumulative frequency (cf) of complete dataset is obtained.

The value of data against n+1/2 the cumulative frequency will be the median for odd number of data.

- Median for dataset with odd number of values will be:

  Median $(\bar{x})$ = *value of data against* $\left(\frac{n+1}{2}th\right)$ *cumulative frequency*
- Median for dataset with even number of values will be:

  $$\text{Median } (\bar{x}) = \left(\frac{Class\ values\ against\ \frac{n}{2}^{th} + \frac{n}{2}+1^{th}\ \ cumulative\ frequencies}{2}\right)$$

Example for odd number dataset:

Calculate the median of the following dataset obtained by counting the number of flowers on 21 rose plants.

Class (no. of flowers/plant)	1	2	3	4	5
Frequency (no. of plants)	3	5	6	4	3

Calculations:

Class (no. of flowers/plant)	Frequency (no. of plants = n)	Cumulative frequency
1	3	3
2	5	8
3	6	14
4	4	18
5	3	21

Number of plants (n) = 21

Median $(\tilde{x})$ = $(\frac{n+1}{2}th)$ cumulative frequency

Median $(\tilde{x})$ = $\left(\frac{21+1}{2}\right) = \frac{22}{2}$ = 11

Since cumulative frequency 11 is included in 13 which represents the class value 3, therefore, for cumulative frequency 11, the class value will be 3 which is the median.

Example for even number dataset

Calculate the median for the following data recorded for height (in cm) of 80 plants.

Class (plant height in cm)	119	120	121	122	123	124	125
Frequency (no. of plants)	4	9	14	18	15	13	7

Calculations

Calculate the median for the following data recorded for height (in cm) of 80 plants.

Class (plant height in cm)	Frequency (no. of plants = n)	Cumulative frequency
119	4	4
120	9	13
121	14	27
122	18	45
123	15	60
124	13	73
125	7	80

Number of plants (n) = 80

Median $(\tilde{x})$ = $\left(\frac{Class\ values\ against\ \frac{n}{2}th + \frac{n}{2}+1th\ \text{cumulative frequencies}}{2}\right)$

Median $(\bar{x})$ = $\left(\frac{\frac{80}{2}th+\frac{80}{2}+1th \text{ cumulative frequencies}}{2}\right)$

Median $(\bar{x})$ = $\left(\frac{\text{Class values against cumulative frequencies 40 \& 41}}{2}\right)$

The class values for cumulative frequencies 40 and 41 are included in the class value of cumulative frequency 45 which is 122. Therefore, **Median $(\bar{x})$ = 122+122/2 = 122.**

Advantages of Median

1. It is easy to calculate and is located exactly.
2. It is not affected by abnormally large and small values in dataset.
3. It can be used in quantitative measurements.

Disadvantages of Median

1. The median of two or more datasets cannot be calculated by using/adding the median of the individual datasets.
2. Median value may not be necessarily represented in central data.
3. It cannot be used where specific values e.g. weight & age used.

11.2.3. Mode

Definition: The mode is the value in a dataset that appears maximum number of times. A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal) or no mode at all (amodal) i.e. all values occur in same number of times. It may be denoted by "$\hat{X}$ "(read as x-hat or x cap).

Mode = the most frequent or repeated value occurring in a given dataset.

Formula for calculating Mode for Group data is given below

Mode $(\hat{X})$ = $l+\left[\frac{(fm-f1)\,x\,h}{(fm-f1)+(fm-f2)}\right]$

Where, l= lowercase boundary of model group

f_m= maximum frequency of a group

f_1= previous frequency of fm

f_2= next frequency of fm

h= class interval of model group

Class interval and Modal Group (Modal Class)

1. **Class Interval:** A class interval is a range of values into which data is grouped in a frequency distribution table. Each interval includes a lower and an upper boundary (e.g., 120-129). All intervals are typically of equal width.

Example:

Blood Pressure (mm/Hg)	Frequency (f)
110-119	2
120-129	5
130-139	3

Here, 110-119, 120-129, and 130-139 are class intervals.

2. **Modal Group (Modal Class)** The modal group is the class interval with the highest frequency in a grouped frequency distribution. It indicates the most common range of values in the dataset. Using the table above: The class interval 120-129 has the highest frequency (5). So, 120-129 is the modal group.

Examples for calculating Mode

Example #01: Find the mode from the word "BIOLOGY".

Mode ($\hat{X}$) = the most repeated value of the data

Mode ($\hat{X}$) = "O"

Example #02: Find mode from the following data; 3, 5, 6, 6, 9, 10, 9, 6, 6

Mode ($\hat{X}$) = the most repeated value of the data

Mode ($\hat{X}$) =6

Example #03: Find the mode from the following data; 3, 5, 5, 6, 9, 9, 10, 12

Mode ($\hat{X}$) = the most repeated value of the data

Mode ($\hat{X}$) =5, 9

Example #04: Find the mode from the following frequency distribution;

Groups	Frequency (f)	Class Boundary (CB)	
10-19	5	9.5-19.5	
20-29	25 (f_1)	19.5-29.5	
30-39	40 (f_m)	29.5-39.5	(model group)
40-49	20 (f_2)	39.5-49.5	
50-59	10	49.5-59.5	

Mode ($\hat{X}$) = $l + \left[\frac{(fm-f1) \times h}{(fm-f1)+(fm-f2)}\right]$

l=29.5, f_m=40, f_1=25, f_2=20, h=10

Mode ($\hat{X}$) = $29.5 + \frac{(40-25) \times 10}{(40-25)+(40-20)}$

Mode ($\hat{X}$) = 33.786

11.2.4. Standard deviation (SD)

Standard deviation (SD) is a fundamental concept in statistics that measures the dispersion of data points. It states the extent to which data values in a dataset deviate from the mean. It provides a clear sense of the variability within the data. It is used to find the deviation of any data value from the mean value in a data set. **"Standard deviation is a measure of the spread of data values around the mean in a normal curve. It measures that how much a set of data varies from its mean value."**

Standard deviation is sometimes referred to as the "mean of the mean". For any given situation, the standard deviation measures how close the data values are to the mean. If most of the data values are clustered around the mean, then the standard deviation is small. Thus, a lower value of standard deviation means the data values are tightly grouped around the mean. Conversely, if most of the data values are widely spread and are not grouped around the mean, then the standard deviation is large. Thus a higher value of standard deviation means the data values are more spread out. In other words, the more the data values differ from the mean, the greater the standard deviation, and vice-versa. Examples of data values for a biologist may be height of each individual in the sample population or number of hepatitis infected patients in each city of Pakistan or number of fruits on each plant in a garden being investigated for fruit production etc.

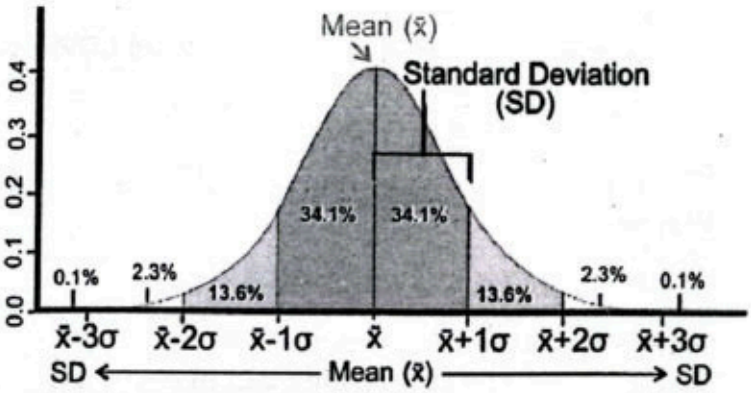


Fig. 11.1: Normal Curve showing Mean (x) and Standard Deviation (SD)

The standard deviation is represented by the lowercase Greek letter sigma (σ) for a population and the Latin letter (s) for a sample. Thus, sigma (σ) value describes how far a specific data value is from its mean. A higher sigma value means the data point is further from the mean. SD and variance are related, because standard deviation is the square root of the variance. Standard deviation is easier to interpret because it is expressed in the same units as the data. The variance is sigma square $(\sigma)^2$.

To clarify the concept of SD, let's consider the quiz results of biology subject in a class. Each of the 40 students scored 77 marks and same grade, so the mean score is also 77. As, all of the scores are equal to the mean, so there is no arithmetic difference between the scores and the mean. Therefore, the standard deviation in this scenario would be zero. But, if a few students have scored 75 marks, then standard deviation would not be zero, and it would be very small and much less than one sigma (σ).

Calculation of standard deviation

The calculation of standard deviation is quite simple, but there are two slightly different ways to do it depending on the context. First, consider the steps below:

i. Determine the mean (arithmetic average)

ii. Subtract the mean from each score

iii. Square the result for each score

iv. Add the results together

v. Divide this result by either the number of scores (biased) or the number of scores minus 1 (unbiased), as explained below.

vi. Determine the square root of this number which is what we call the biased standard deviation

Let's calculate the SD for prevalence of COVID in 2021. Number of confirmed COVID cases in four major cities of Pakistan were recorded:

Cities	Number of confirmed COVID cases	Mean of COVID cases
Karachi	90	60
Lahore	80	
Peshawar	40	
Quetta	30	

1. **Determine the mean:** The mean is 60.
2. **Subtract the mean cases from cases of each city:** Karachi = (90-60) = 30;
   Lahore = (80-60) = 20; Peshawar = (40-60) = -20; Quetta = (30-60) = -30
3. **Square the result for each city:**
   Karachi = (30)2 = 900; **Lahore** = (20)2 = 400; **Peshawar** = (-20)2 = 400; **Quetta** = (-30) = 900
4. **Add the results of all cities together:** 900 + 400 + 400 + 900 = 2600
5. **Divide this result by the number of scores minus 1 (unbiased),** because we are interested in considering these cities as a sample from the entire Pakistan: 2600/3= 866.7
6. **Determine the square root of this number which is the standard deviation (SD):**

The square root of 866.7 = 29.45, Hence, Standard deviation (SD): 29.45

As a general rule large standard deviation means that data is well dispersed away from the mean. A small standard deviation indicates a tight cluster of data points near the mean.

Standard deviation (SD) of a population or sample can be calculated by using following formulas:

Population	Sample
$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$	$S = \sqrt{\frac{\Sigma(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population mean N - Total number of observations	X - The value in the data distribution $\bar{x}$ - The sample mean n - Total number of observations

As calculations of SD are complex and there's a high risk of error, so statisticians don't usually calculate standard deviation by hand. We can use a calculator or computer software to find the SD.

11.2.5. Range

In biostatistics, the range helps researchers to understand the variability within data obtained from some experiment or variation in a population. Range is defined as **"the difference between the highest and lowest values in a set of data"**. The range can also be defined as **"the difference between the highest observation and lowest observation"**.

The obtained result is called the **"range of data values"** or **"range of observation"**. It is the measure of spread or variability of a dataset or observations obtained from experiment.

Calculation of Range

To find or calculate the range first step is to arrange the given values of data set or set of

observations in an ascending order. This means, firstly we have to arrange the values or observations from the lowest to the highest sequence. Then, we can use the formula to find the range of values or observations.

Formula for Range

The formula of the range in statistics can simply be given as follows:

Range = Highest or Maximum Value - Lowest or Minimum Value

Or

Range = Highest observation - Lowest observation

During biological studies, we can use range to find the variation in population or findings of biological research.

Example: Imagine a student studying the heights of a sample of plants in a garden after the treatment of fertilizers. Student collected the following data set of plant heights (in centimetres) from sampled plant population:

1. **Raw data Set:**

10cm, 15cm, 18cm, 21cm, and 25cm, 32cm, 41cm, 28cm, 54cm, 35cm, 26cm, 23cm, 33cm, 38cm, 40cm, 19cm, 33cm, 43cm, 08cm, 13cm

2. **Arranged data set (ascending order):**

   08cm, 10cm, 13cm, 15cm, 18cm, 19cm , 21cm, 23cm, 25cm, 26cm, 28cm, 32cm, 33cm, 33cm, 35cm, 38cm, 40cm, 41cm, 43cm, 54cm

3. **Identify the highest value:** 54cm
4. **Identify the lowest value:** 08cm
5. **Calculate the range:** Range = Highest or Maximum Value - Lowest or Minimum Value

   Range = 54cm - 08cm = 46cm

   The range here is 46cm, meaning there is a 46 cm difference between the smallest and tallest plants in the sample.

Importance of Range in Biology

In biology, the range helps scientists understand the variability within a population. This can be useful in studies of genetics, ecology and physiology where variability may indicate environmental influences or adaptations or organisms in a given population.

A small range would mean that the organisms in sampled population responding to the treatment in about the same level. A small range indicates that the individuals are more similar. A small range is generally good news for the researcher because it indicates that the treatment is almost equally effective for all organisms. A small range is best when the low range is also at a high effectivity level.

The middle range is often the most difficult for the researcher because no clear-cut information can be derived from the data. The researcher must re-plan to proceed the experiment when intermediate range of results are observed.

A large range suggests a high level of variation among individuals.

Limitations of Range

Before planning for the future study on the basis of range, the researcher must consider another factor regarding range. The range statistic may be deceiving. The range only accounts for the

highest and lowest values. For example, if the almost all the plants responded well to the fertilizer treatment with greater productivity and only one or two plants showed lowest productivity. In this situation we can conclude that looking at the range alone may be misleading.

11.2.6. Percentile

A percentile is a value that indicates the percentage of data in a set that is below a certain value. For example, the 25th percentile is the value below which 25% of the data points lie. Percentiles are used to interpret and understand data and are often used to report test scores and biometric measurements.

Percentiles divide the ordered data into 100 separate units. For instance, it is possible to have a data point at any number between 0 and 100 such as the 99 percentile, 44 percentile or 9th percentile. A percentile is the value or score on the ordered array of data that indicates the percentage of the scores that fall at or below that value.

For example, the 25th percentile of the number of fruits on the plants in the garden is number of 39 fruits on a plant. This would mean that 25% of the plants from that garden produced fruits below 39.

The 25th percentile is also known as the first quartile; the 50th percentile is also the median. The desired 99th percentile is that location point where no scores are above it. A plant which produced number of fruits in the 99th percentile has produced very well in comparison to the rest of the plants who were included in the pool total fruits produced in the garden. For research and biometry purposes, percentile ranks are used to interpret an individual's raw value by comparing it with the value of other individual in the study group. For example, if a student's height measured 77cm and that height is translated to a percentile rank of 89, then 89% of the students in that population have a height of 77cm or less.

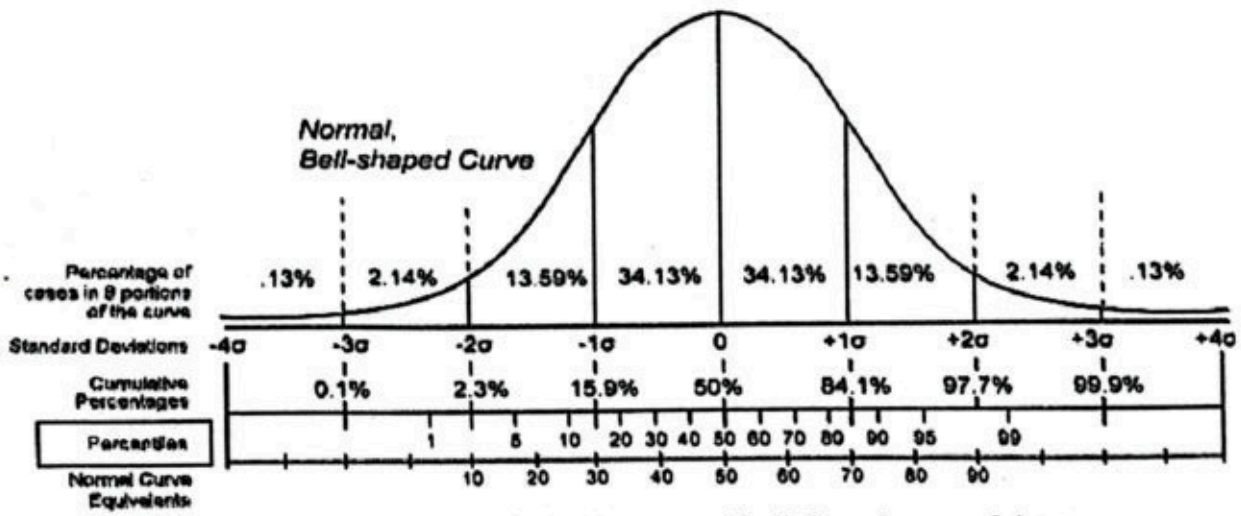


Fig. 11.2: Percentile calculation in a normal bell shaped curve of data

Further helping material to understand Percentile:

Quartiles

A quartile is a special type of percentile. A quartile is any of the three values which divide the sorted data set into four equal parts so that each part represents one-fourth of the population. Each quartile defines a specific section of the ordered data distribution.

The first quartile (Q.) or lower quartile is that point on the ordered array where 75% of the values are above it and 25% of the values are below it. The range of the first quartile accounts for the bottom one-fourth of the data, or the lowest 25% of whatever is being measured. Thus, the lowest quartile is also known as the 25th percentile.

The second quartile (Q.) would identify the midpoint of the data where 50% of the values are above and 50% are below. The second quartile is also called the median and the 50th percentile.

The third quartile (Q) establishes that point where 25% of the values are located above it and 75% below it.

The interquartile range (IQR) is the difference between the upper and lower quartiles. The interquartile range is often used to characterize the bulk of the population.

The quartiles labelled in the data given below are collected as the number of fruits on the orange trees in a small garden.

Quartile	Number of fruits on each plant
	11
	13
	14
First Quartile	15
	18
	18
	19
Second Quartile= (21)	19
	23
	23
	23
	24
Third Quartile	24
	25
	25
	26

Quartiles divide the data so the researcher can analyze the results by an ordered grouping rather than by focusing on the entire data pool.

Steps to calculate percentile

1. Arrange Data: Sort the data values in ascending order.
2. Calculate Position: Use the formula for the position of the P^{th} percentile:

Position = $\frac{P}{100} X(N+1)$

P is the desired percentile (e.g., 25 for the 25th percentile).
N is the number of observations.
Percentile can be calculated by using the data by formula:

$P = \frac{n}{N} X\ 100\%$

Where P is the percentile,
n is the number of data points below the data point of interest, and
N is the total number of data points in the data set.

11.3 SKETCHING A BAR CHART FOR A GIVEN SET OF BIOLOGICAL DATA

11.3.1 Graphic presentation of scientific data: Need and its importance

Biostatistical values (data) are documented in the form figures and are usually recorded in the form of tables. But numerical or arithmetical figures are usually not attractive. Large number of figures in tables are not easily understandable. It seems difficult to get a clear picture from the numerical data. Diagrams and graphs are comparatively more attractive way of presenting the clear and complete picture of data. Moreover, data information presented in the form of diagrams, pictures, charts and graphs leave long lasting impression on the mind of the observer. Multiple types of graphs, pictures and diagrams are used in representing data including bar-chart, rectangles and pie-chart.

11.3.2 Simple bar diagram or simple bar chart

Simple bar chart have bars which are drawn to represent a single data without further classification of the characteristics. Bar chart represent only one characteristic of the data. One bar represents only one value or figure and there may be multiple bars representing many values or figures. In a simple bar chart, bars of equal width are drawn, but their lengths are variable. Heights of length of the bars represent the magnitude of a quantity or degree of value in the data.

11.4. SKETCHING OR CONSTRUCTION OF BAR CHARTS AND PIE CHARTS WITH PROPER TITLE, LABELLED AXES, LEGEND, AXES UNITS

Some basic steps in sketching of bar charts are:

1. Take a graph paper with suitable sizes and scales as per requirement and type of data.
2. Draw two lines on the graph, perpendicular to each other and intersecting at zero.
3. Label the axis of the graph, horizontal line is X-axis and vertical line is Y-axis.
4. Choose uniform width of bars and uniform gap between the bars along the horizontal axis (X-axis).
5. Label X-axis with the names of the data items whose values are to be presented.
6. Choose a suitable scale to determine and specify the heights of the bars for the given values along the vertical axis (Y-axis).
7. Label the Y-axis scale with specific unit of data value to grade the length of bars.
8. Calculate the heights of the bars according to the scale chosen and mark as dots.
9. Draw the bars carefully and recheck all the entries.

Bars in the graph may be drawn vertically or horizontally but normally vertical bars are used.

Some basic steps to make a pie chart manually, follow these steps

1. **Gather Your Data:** You'll need a list of categories and their values.
   Example: Category A: 20, Category B: 30, Category C: 50
2. **Calculate Total and Percentages:** Add up the values to get the total. Then, for each category, divide its value by the total and multiply by 100 to get its percentage.

Example: Total = 20 + 30 + 50 = 100, Category A: (20 / 100) × 100 = 20%, Category B: (30 / 100) × 100 = 30%, Category C: (50 / 100) × 100 = 50%

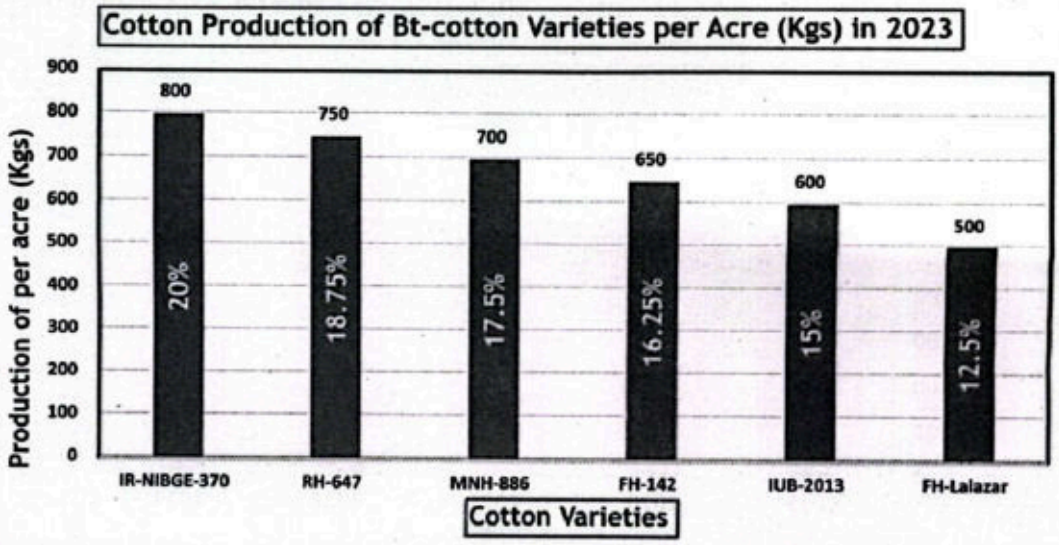
3. **Convert Percentages to Degrees:** Since a circle is 360 degrees, multiply each percentage by 3.6 (because 100% × 3.6 = 360°) to get each angle.
   Example: Category A: 20% × 3.6 = 72°, Category B: 30% × 3.6 = 108°, Category C: 50% × 3.6 = 180°
4. **Draw a Circle:** Use a compass to draw a circle on your paper. Mark the center.
5. **Measure and Draw Each Segment:** Start at the top of the circle (12 o'clock position).Use a protractor to measure each angle from the center, marking each section. Draw lines from the center to the edge of the circle at each angle to divide the circle into sections.
6. **Label Each Segment:** Write each category name and percentage inside or near each section.
7. **Add Color (Optional):** You can color each segment differently for clarity.

Bar chart and pie chart can be constructed by using computer software.

Example 1: Draw simple bar graph/chart to represent the cotton (weight in Kgs) of produced per acer of land from five Bt cotton varieties (IR-NIBGE-370, RH-647, MNH-886, FH-142, IUB-2013 and FH-Lalazar) in year 2023.

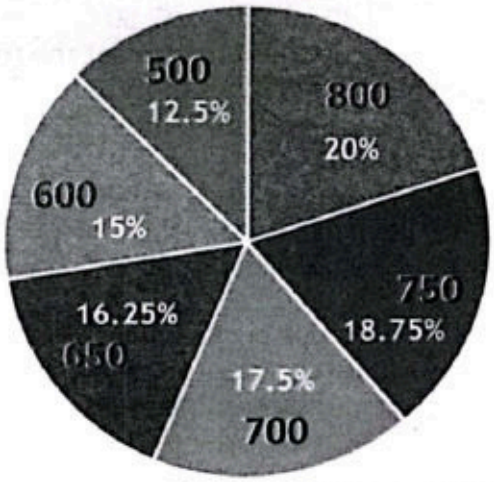
Cotton Varieties	IR-NIBGE-370	RH-647	MNH-886	FH-142	IUB-2013	FH-Lalazar
Production per acre (Kgs)	800	750	700	650	600	500
Production per acre (%age)	20%	18.75%	17.5%	16.25%	15%	12.5%

Simple bar chart using above data



Simple pie chart using above data

Pie chart to show Cotton Production of Bt-cotton Varieties per Acre (Kgs) in 2023



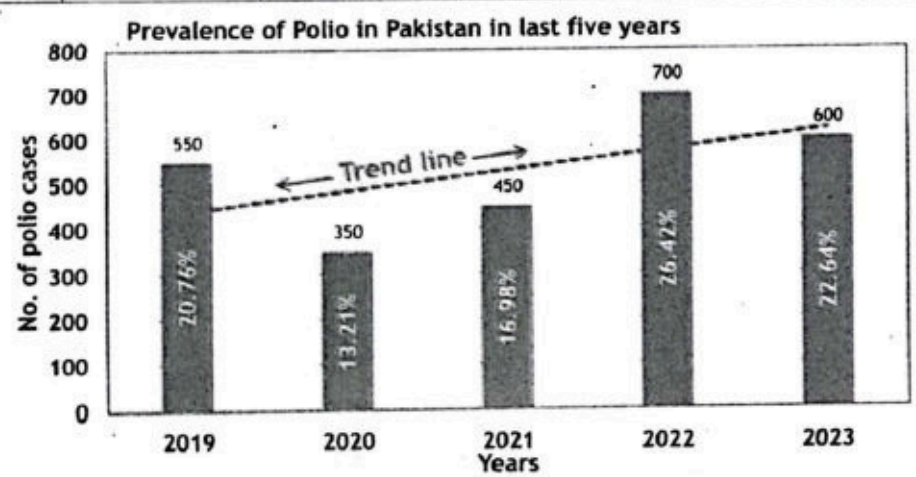
Trend Line and it's purpose

A trend line shows the overall direction or pattern in the data over time or across categories. In bar graphs, trend lines are usually drawn as a dotted line across the tops of the bars to show increasing, decreasing or constant trends. Trend line explains that

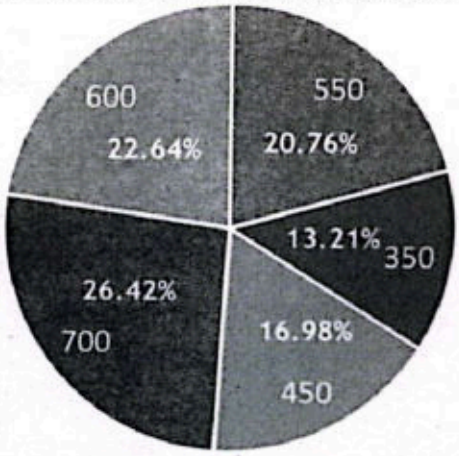
"There is an increasing trend...
""The trend line shows a gradual decline...
""The trend remains fairly constant...
""There is a sharp rise/drop between..." etc.

Example 2: Draw simple bar graph/chart to represent the prevalence of polio, showing number of confirm polio cases in Pakistan in last five years (2019-2023)

Years	2019	2020	2021	2022	2023
No. of polio cases	550	350	450	700	600
%age of polio cases	20.76%	13.21%	16.98%	26.42%	22.64%



Pie chart for Prevalence of Polio in Pakistan in last five years



11.5 SKETCHING ERROR BARS BASED ON RANGE OR STANDARD DEVIATION ON A BAR CHART

To sketch error bars on a bar chart, representing variability such as range or standard deviation, follow these steps:

1. Calculate Error Values

Range-Based Error Bars: For each data point or category, first determine the minimum and maximum values to get the range. The error bar's length will be half of the range, extending above and below the mean or median value.

Standard Deviation (SD) Error Bars: For each data point or category, calculate the standard deviation. Error bars will extend above and below the mean by one standard deviation.

2. Draw the Bar Chart

Create a bar chart with the values representing the mean (or other central tendency) of each category.

3. Add Error Bars

Vertical Error Bars: Draw a line centered on each bar, extending upward and downward by the calculated range or standard deviation. Use caps on the ends of each line for a clean representation.

Error bar Length:

For range, set the error bar to $\frac{(\text{Max value} - \text{Min value})}{2}$

For standard deviation, set it equal to the standard deviation.

Calculation for Standard Deviation

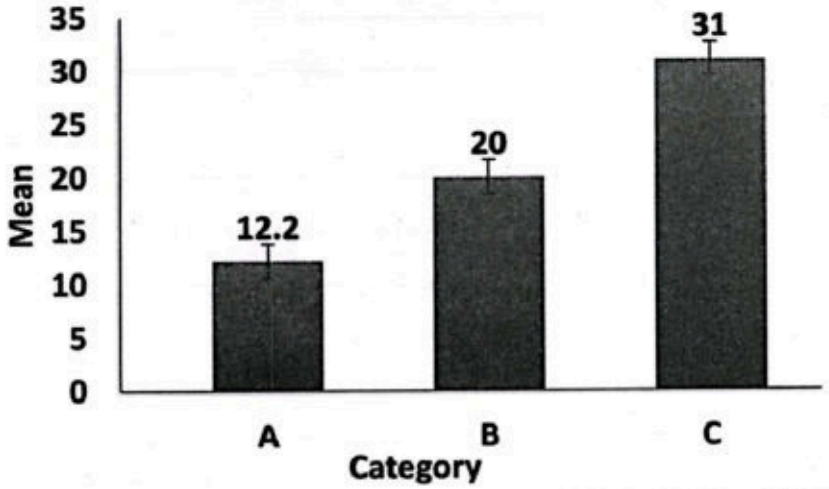
Suppose we have the following data points for three categories in a biological experiment:

1. Calculate the mean and standard deviation for each category.
2. For each bar representing the mean of each category, add error bars with lengths based on the calculated standard deviations.

Here are the data points, mean values, ranges and standard deviation (SD) and error bar lengths for each category:

Category	Data Points	Mean	Range $\frac{(\text{Max value} - \text{Min})}{2}$	Standard deviation (SD)	lengths of error bars (±SD)
A	(10,12,15,11,13)	12.2	2.5	1.92	±1.92
B	(20, 21, 22, 19, 18)	20.0	2	1.58	±1.58
C	(30, 33, 32, 31,29)	31.0	2	1.58	±1.58

Graph showing error bars



11.6. EXPERIMENTAL DESIGN WITH CONTROL GROUP AND DEPENDENT, INDEPENDENT AND CONTROL VARIABLES

While designing experiments, two main types of groups are used to get better and clear outcome of any experiment. Control and experimental groups are used in experiments to compare results and evaluate acceptance of the findings of experiments.

11.6.1. Experimental Group

This group is to test and experience a change in a specific variable that is being tested. The experimental group is also known as the treatment group because in this group, scientist apply a specific change or treatment to observe the effects of that treatment. For example, if students of class 12 are testing the effect of a fertilizer on the plant growth, plants in the experimental group would receive the fertilizer.

11.6.2. Control Group

This group is kept under unchanged normal conditions without any specific treatment as applied in the experimental group. It serves as a standard or reference to compare the effects of treatment in experimental group. In example of effects of fertilizer on plant growth, the control group would consist of plants grown without any added fertilizer.

Independent, Dependent and Controlled Variables

There are three main types of variables in scientific investigations: independent, dependent, and controlled variables.

11.6.3. Independent Variables

The variable that the experimenter (researcher) changes or manipulates in the experiment. This is the variable that scientist think, will cause an effect during experiment. That is why, they are also known as the **"input variables"** or the **"cause variables"** because they are the factors that cause changes in the dependent variable. In case of fertilizer experiment, the independent variable would be the presence or absence of fertilizer.

11.6.4. Dependent variables

These variables depend on the independent variable or affected by the independent variable in an experiment. The dependent variables are what researcher measure or observe to determine the effect of the independent variable that is why they are also known as the **outcome variables** or the **effect variables.** In the plant fertilizer experiment, the dependent variable might be plant height, number of leaves or overall growth rate.

11.6.5. Controlled variables

These are factors that must be kept constant during an experiment across both the experimental and control groups to ensure that they do not influence the results. These variables are also known as **constant variables** or the **controlled factors.** The purpose of controlling these variables is to ensure that any changes observed in the dependent variable are due to changes in the independent variable and not due to other factors. These elements are essential for designing a fair and effective experiment, allowing you to clearly see the impact of the independent variable on the dependent variable. For example, control variables in the fertilizer experiment might include the amount of sunlight, temperature, water, type of plant and soil quality.

Further examples of independent, dependent and control variables:

Independent variable: If a researcher is investigating the effect of temperature on the rate of photosynthesis in plants, temperature would be the independent variable. Researcher would manipulate the temperature to see how it affects the rate of photosynthesis. It is essential to note that an experiment should have only one independent variable at a time. This is because if researcher changes more than one variable in the same experiment, he would not know which variable caused the change in the dependent variable. Therefore, by controlling the independent variable, researcher can determine the effect of that variable on the dependent variable.

Dependent variable: If a researcher is experimenting to know the effect of temperature on photosynthesis, the dependent variable would be the rate of photosynthesis, which is affected by changes in temperature. It is crucial to keep the record of dependent variables continuously during an experiment to ensure the clear observation of any effect as a result of changes in the independent variable. Additionally, the dependent variables should be measurable quantitatively so that it can be expressed and compared in numerical values.

Controlled variable: If a researcher is investigating the effect of temperature on the rate of photosynthesis in plants, the controlled variables would include factors such as the type of plant, the duration and intensity of light and the amount of CO_2. By keeping these variables constant, researcher can ensure that any changes in the rate of photosynthesis are surly due to changes in temperature and not due to any other relevant factors.

11.6.6. Experimental design

Let's consider a scenario where we want to investigate the effect of light intensity on plant growth and effect of different amounts of water on plant growth. For this purpose we design two separate experiments:

Experiment 1: Effect of Light Intensity on Plant Growth

Objective: To investigate how different light intensities affect the growth of a specific plant (e.g., bean plants).

Hypothesis: Plants exposed to higher light intensities will grow more compared to plants under lower light intensities.

Materials required: 20 potted bean plants (similar age and size), light source with adjustable intensity, ruler, water and watering can, soil with same composition, thermometers

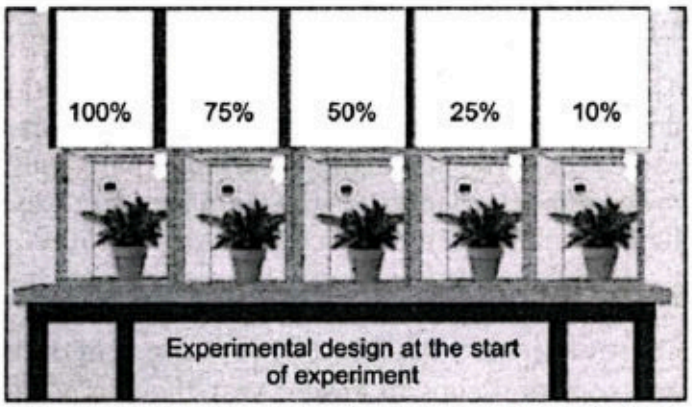


Fig. 11.3: Experimental design to check the effect of light intensity on plant growth

Procedure:

1. Divide the 20 potted plants into five groups of four plants each.
2. Place each group under different light intensities (e.g., 100%, 75%, 50% and 25%).
3. Ensure all groups receive the same amount of water and are kept at a constant temperature.
4. Measure the height of the plants every three days continuously for four weeks.
5. Record and analyze the growth data.

Variables:

Independent Variable: Light intensity (100%, 75%, 50%, 25% and 10%)

Dependent Variable: Plant growth; which we could be measured by tracking the height of the plants (in cm) and number of leaves of the plants.

Control Variables: Type of plant, type of soil, amount of water, temperature, duration of light, amount of CO_2, pot size and initial plant size.

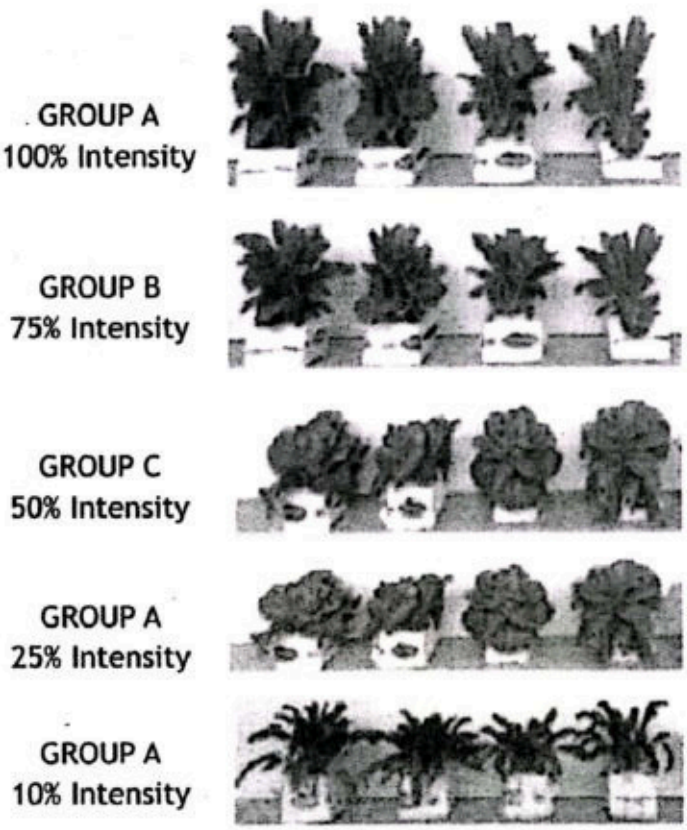


Fig. 11.4: Experimental results showing the effect of light intensity on plant growth

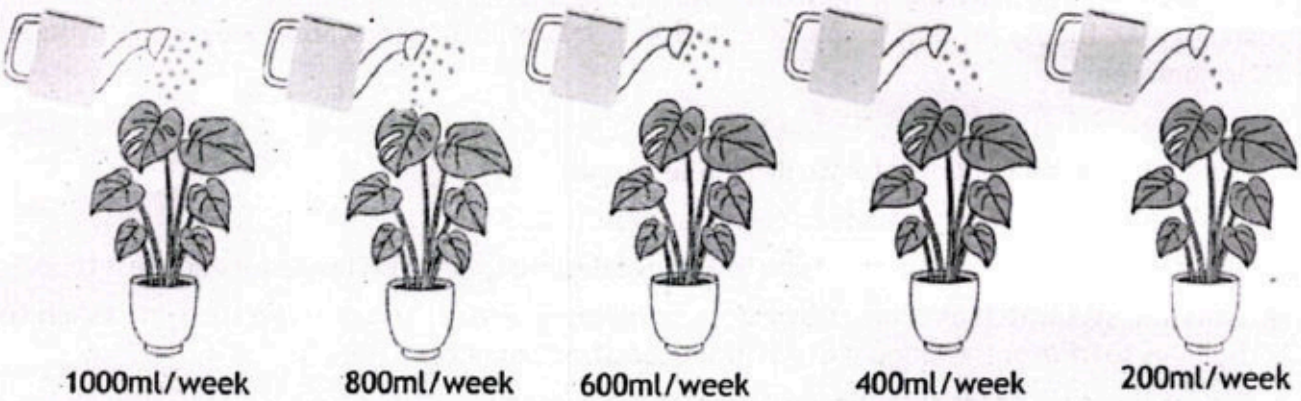
Control Group: Plants grown under the normal light intensity (e.g., 100%)

Experiment 2: Effect of different amounts of water on plant growth

Objective: To investigate how different amounts of watering, affects the growth of a specific plant (e.g., potato plants).

Hypothesis: Plants given more amount of water (1000ml per week) will grow more compared to plants given lower amount of water.

Materials required: 20 potato plants (similar age and size) in lager pots, water and graduated watering can, Light source with same intensity for all plants, ruler, soil with same composition, Thermometer.



Experimental Design at the start of experiment

Fig. 11.5: Experimental design to check the effect of amount of water on plant growth

Procedure:

1. Divide the 20 potato plants potted in larger but same size pots into five groups of four plants each.
2. Place all five groups under light of same higher intensity (100%) and at a constant temperature (30°C).
3. Give different amount of water per week to each group of plants (1000ml, 800ml, 600ml, 400ml and 200ml).
4. Measure the height (in cm) of the plants and number of leaves on each group of plants after every three days continuously for six weeks.
5. Record and analyze the data of plant growth.

Fig. 11.6: Experimental results showing the effect of amount of water on plant growth

Variables:

Independent Variable: Amount of water given per week to each group of plants (1000ml, 800ml, 600ml, 400ml and 200ml)

Dependent Variable: Plant growth; which we could be measured by tracking the height of the plants (in cm) and number of leaves of the plants.

Control Variables: Type and species of plant, type of soil, duration and intensity of light,

temperature, amount of CO_2, pot size and initial plant size

Control Group: Plants watered 1000ml per week

Experiment 3: Effect of Different pH Levels on Enzyme Activity

Objective: To examine how pH levels affect the activity of the enzyme catalase in breaking down hydrogen peroxide.

Hypothesis: Catalase will have the highest activity at a neutral pH and will be less active in acidic or alkaline conditions.

Materials required: Hydrogen peroxide solution, catalase (can use potato or liver tissue as a source), Test tubes, pH buffer solutions (pH 4, pH 7, pH 10), Stopwatch, Measuring cylinder, Thermometer

Procedure:

1. Place an equal amount of catalase in each test tube.
2. Add 10 mL of hydrogen peroxide to each test tube.
3. Add different pH buffers to each tube to achieve desired specific pH levels for each test tube.
4. Start the stopwatch as soon as hydrogen peroxide is added, and observe the time taken for bubbles to form or the amount of gas produced after 1 minute.
5. Repeat three times for each pH and calculate the average.

Variables:

Independent Variable: pH level (pH 4, pH 7, pH 10)

Dependent Variable: Enzyme activity (rate of reaction measured by gas produced or reaction time)

Control Variables: Amount of hydrogen peroxide, amount of catalase, temperature, and test tube size

Control Group: Reaction at neutral pH (pH 7)

Above mentioned experiments, will give a hands-on and easily observable understanding to students of scientific methods and the effects of variables on biological processes.

EXERCISE

Section I: Multiple Choice Questions
Select the correct answer:

1. In experiment examining the effects of pH levels on the activity of the enzyme, independent variable will be
   A. pH B. Temperature C. Substrate D. Reaction time
2. The variable that the experimenter changes in the experiment to check its effect during experiment is called.
   A. Dependant variable B. Independent variable
   C. Control variable D. Experimental variable

3. Variables that researcher measure to determine the effect of the independent variable is also known as:
   A. Outcome variables B. Input variables C. Cause variables D. Control variable
4. To which of the following fields of life sciences, biostatistics is not applicable?
   A. Health B. Chemical engineering C. Agriculture D. Medicine
5. 25th percentile is the value ---------- of the data points lie.
   A. Above which 25% B. Below which 25% C. ¾ th D. 25
6. Biostatistics is not involved in:
   A. Epidemiological studies B. Agriculture C. Petroleum industry D. Public Health
7. The factors that must be kept constant during an experiment across both the experimental and control groups are known as
   A. Control variables B. Outcome variables C. Input variables D. Cause variables
8. All of the followings are types of averages except:
   A. Arithmetic Mean B. Median C. Mode D. Product of data values
9. The symbol for mean is:
   A. "$\bar{x}$" (read as "x bar") B. $\tilde{x}$ (read as "x tilde")
   C. "$\hat{X}$" (read as x-hat) D. "h" ("read as highly stable")
10. Formula of "Mean" for un-group data is:
    A. $\frac{\Sigma x}{n}$ B. $\frac{\Sigma fx}{\Sigma f}$ C. $\frac{\Sigma f1}{\Sigma f2}$ D. $\frac{n+1}{2}$
11. Find the mode from this dataset given 46, 49, 55, 58, 62, 66, 64, 65, 66, 67, 68, 70, 72:
    A. 46 B. 62 C. 66 D. 72
12. The first step in genetic engineering is:
    A. Isolation of the gene of interest
    B. Insertion of the gene into a vector
    C. Transfer of recombinant DNA into the host organism
    D. Expression of the gene
13. Biostatistics is similar to maths because of:
    A. Calculations B. Counting birds
    C. Used of symbol pie D. It is also useful in engineering studies
14. If the numbers of diseased persons is greater in 2nd year of study than 1st year then
    A. Height of the bar will be greater in 1st year
    B. Width of the bar will be greater in 1st year
    C. Height of the bar will be smaller in 1st year
    D. Width of the bar will be greater in 2nd year
15. In a bar chart of number of deaths due to hepatitis in Pakistan in last 10 years what will we write at the x-axes of graph?
    A. Years B. Number of deaths
    C. Any of the value D. Causes of hepatitis

Section II: Short Answer Questions

1. How Standard Deviation is different from Range?
2. What are the demerits of relying on range to decide future plans in scientific experiments?
3. Why control group is important for scientific experiments?
4. Define mean, median and mode.
5. Write down the formulae of median and mode.
6. What is importance of mean calculation?
7. Why value of mean is not completely reliable to get the correct picture?
8. Write the formula for calculation of Mode.
9. How bar chart is different from pie-chart?
10. Write one similarity and one difference between rectangle graph and pie-graph.
11. How graphical representation of data is important that tabular representation?
12. Find out the standard deviation of the values in following dataset.
    33, 55, 66, 84, 91, 72, 68, 87, 78, 66, 82, 46, 71, 31, 38
13. Calculate the median of the values given in dataset of above question 13.

Section III: Extensive Answer Questions

1. What are advantages and disadvantages of Median?
2. Describe detailed procedure to add error bars in the graph.
3. Complete the following table by filling the values of Cumulative frequency.

Class (no. of flowers/plant)	Frequency (no. of plants = n)	Cumulative frequency
1	30	
2	50	
3	60	
4	40	
5	30	

4. Give details steps to construct the Bar-chart.
5. For what the letters l, fm, f1, f2 and h stands for in the given formula? For which calculation this formula is used? $l + \left[\frac{(fm-f1) \times h}{(fm-f1)+(fm-f2)}\right]$
6. Sketch and label the Pie-chart by using the data given in the table below about the deaths of vulture birds due to pollution in last five years? Also label the data values in the pie-chart.

Years	2019	2020	2021	2022	2023
No. of deaths	250	400	275	350	175

7. Draw a table and then sketch and label the bar-chart by using the data of height (in cm) collected from 200 students of 12th class of your college.
8. Compare and contrast control variables, dependant variables and independant variables.
9. Describe the usage of percentile along with its demerits.
10. Describe stepwise details to make the appropriate chart with proper title, labeled axes, legend, axes units from a fictitious data set.