Chapter 22

BIOSTATISTICS

STUDENTS'LEARNING OUTCOMES

After studying this chapter, the students will be able to:

- Define biostatistics and its use.
- Define mean, median, mode, standard deviation, range and percentile.
- Calculate mean, median, mode, standard deviation, range and percentile from a given set of data.
- Sketch a bar chart from a given set of data.
- Sketch error bars based off of range or standard deviation for a given set of data on a bar chart.
- Evaluate the appropriate type of figure or chart for a given set of data and/or experiment (bar chart, pie chart, x-axis data figure etc.).
- Make the appropriate chart with proper title, labelled axis, legend, axis units.
- Design an appropriate experiment with a control group and dependent, independent and control variables.

22.1 BIOSTATISTICS

Biostatistics is a scientific discipline that applies statistical methods and techniques specifically to biological, medical, and health-related data. It involves the collection, presentation, analysis, and interpretation of data to draw meaningful conclusions. Biostatistical Analysis allows researchers and health professionals to make informed decisions about populations using sample data, rather than testing entire populations which may be impractical or costly. For example, by studying a representative subset of workers in a population, biostatistics helps estimate the prevalence of a health condition among all workers.

22.1.1 Components of Biostatistics

The key components of biostatistics comprehensively encompass several critical areas that collectively enable the proper collection, analysis, interpretation, and application of biological or health-related data. These components include:

1- Study Design and Experimental Planning

This involves formulating the research question, determining objectives, and structuring experiments or observational studies to test hypothesis. It includes selecting the population, defining variables, randomization, controlling confounding factors, and calculating adequate sample size. Proper study design ensures validity, reliability, and unbiased results.

2- Data Collection

Collecting accurate and relevant data is fundamental. This can be done through various methods:

Primary Data Collection: Direct data from experiments, surveys, interviews, observations, and clinical trials.

Secondary Data Collection: Using existing data from published research, medical records, government reports, and databases.

3- Data Management and Processing

Organizing the collected data systematically and ensuring accuracy before analysis. It includes data cleaning, coding, and storage, maintaining confidentiality and integrity.

4- Data Analysis

Application of statistical techniques to summarize, describe, and draw inferences from data. **Descriptive statistics** describe characteristics of the data (mean, median, mode, standard deviation etc.) **Inferential statistics**, including hypothesis testing, regression analysis, and modelling, allow conclusions about populations from sample data.

5- Interpretation of Results

Translating statistical findings into meaningful biological or health conclusions. Consideration of context, limitations, and assumptions of the analysis is important.

6- Presentation and Communication

Effectively communicating results using visualizations (charts, graphs) and clear language tailored to scientific, medical, and public audiences. Transparency in reporting methods and findings is essential for reproducibility and trust.

22.1.2 Applications of Biostatistics

1. Clinical trials and medical research

Biostatistics plays a crucial role in clinical trials and medical research by providing the framework for designing studies and analyzing data to ensure valid and reliable outcomes. Through statistical methods, biostatisticians evaluate the safety and effectiveness of new drugs, treatments, and medical devices. This process enables healthcare professionals to make evidence-based decisions that improve patient care and advance medical knowledge.

2. Epidemiology

In epidemiology, biostatistics is used to investigate the distribution and determinants of diseases within populations. It helps public health experts understand how diseases spread, identify risk factors, and evaluate the impact of preventive measures. By modelling disease trends and assessing intervention

outcomes, biostatistics supports efforts to control epidemics and improve community health.

3. Genetics and genomics

The field of genetics and genomics relies heavily on biostatistics to analyze complex genetic data. This includes uncovering relationships between genes and diseases, understanding inheritance patterns, and exploring genetic variations. Biostatistical methods are essential in personalized medicine, where treatments are tailored to an individual's genetic profile to enhance effectiveness and reduce adverse effects.

4. Public health policy

Biostatistics is instrumental in shaping public health policy by analyzing population health data and evaluating health programs. It identifies vulnerable groups and health disparities, providing policymakers with evidence to design and implement interventions that promote equity and improve overall health outcomes on a large scale.

5. Environmental health

Environmental health benefits from biostatistics through the assessment of how environmental factors like pollution and toxins affect human health. Statistical analysis helps identify harmful exposures and guides regulations and preventive strategies to protect communities from environmental hazards.

6. Bioinformatics

In bioinformatics, biostatistics supports the analysis of vast biological datasets from genomics and proteomics. By developing algorithms and models, biostatisticians enable the interpretation of complex data essential for drug discovery, disease diagnosis, and the development of personalized therapies.

7. Healthcare quality control

Healthcare quality control also relies on biostatistics to monitor and improve healthcare services. By applying statistical tools, healthcare providers can evaluate diagnostic test accuracies, track patient outcomes, and ensure consistency in medical procedures, ultimately enhancing the quality of patient care.

8. Health programs and population interventions

Biostatistics evaluates the effectiveness of health programs and population-based interventions. It helps public health officials understand the impact of immunization campaigns, health education efforts, and training programs for healthcare workers. This evaluation is vital for optimizing resource allocation and maximizing the benefits of public health initiatives.

9. Controlling epidemics

During epidemics, biostatistics is critical in tracking the spread of disease, estimating mortality rates, and identifying at-risk populations. This real-time statistical analysis informs decision-making and helps implement targeted strategies to contain outbreaks and mitigate their effects.

10. Identifying barriers to care

Biostatistics also uncovers barriers to healthcare access by analyzing survey data and health service usage patterns. This insight helps healthcare providers and policymakers develop strategies to improve access and reduce health disparities, making healthcare systems more efficient and equitable.

11. Health risk assessment and demography

Finally, biostatistics plays an essential role in health risk assessment and demography by analyzing trends in births, deaths, disease prevalence, and other vital statistics. These analysis provide crucial information for government agencies and health organizations to plan services, allocate resources, and address public health challenges effectively. Together, these multifaceted uses illustrate the indispensable nature of biostatistics in improving health outcomes worldwide.

22.2 KEY STATISTICAL MEASURES AND THEIR CALCULATION

In the field of statistics, understanding and calculating key measures is essential for summarizing and interpreting data. These measures provide valuable insights into the central tendency, variability, and distribution of data, helping to identify patterns and make informed decisions. The mean represents the average value, the median indicates the middle point, and the mode highlights the most frequent value of dataset. Standard deviation and range measure the spread and dispersion of data, while percentiles break the data into specific intervals for a deeper analysis.

22.2.1 Mean

The mean, also known as the average, is a measure of central tendency that gives an idea of the central value in a dataset. From a biological perspective, it can be used to summarize large datasets, providing a central value that represents typical characteristics, such as average size or growth rate.

Formula

For ungrouped data

$$\text{Mean} \quad \frac{x}{n}$$

Where,

x = sum of all the values.

n = number of values in the dataset.

For grouped data

$$\text{Mean} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Where:

x_i = midpoint of the ith class interval.

f_i = frequency of the ith class.

f_i = sum of all the frequencies.

Steps to calculate the Mean (for ungrouped data)

A biologist measured the average weight of a specific species of bird in a park. He recorded the weights (g) of five birds from the same species. The bird's weights are 400g, 420g, 470g, 390g and 460g.

Step 1: Add all the weights together (X)

$$X = 400+420+470+390+460 = 2140 \text{ g}$$

Step 2: Count the total number of birds (n)

$$n = 5 \text{ birds}$$

Step 3: Divide the sum of bird's weights by the number of birds

$$\text{Mean} = \frac{X}{n} = \frac{2140}{5} = 428g$$

The average weight of a specific species of birds is 428g.

Steps to calculate the Mean (for grouped data)

A biologist studied the growth of a certain plant species under different soil conditions. The plant heights (cm) are grouped into class intervals based on measurements from different plots of land with varying soil types.

Step 1: Organize the data set into class intervals.

Step 2: Calculate the midpoint of each class interval.

Step 3: Multiply each midpoint by its corresponding frequency.

Step 4: Sum the products of midpoints and frequencies.

Step 5: Divide the result of the step 4 by the total frequency.

Table: Mean plant height (cm) calculation using grouped data across soil types.

Soil Type	Class Interval (Height in cm)	Frequency (f_i)	Midpoint (x_i = lower limit + upper limit / 2)	$f_i x_i$
Sandy Soil	0 – 10	5	5	25
Loamy Soil	10 – 20	8	15	120
Clayey Soil	20 – 30	7	25	175
Peaty Soil	30 – 40	4	35	140
Median		$f_i = 24$		$f_i x_i = 4601$

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{460}{24} = 19.16667 \approx 19.17(cm)$$

The average growth of a certain plant species under different soil conditions is 19.17cm.

22.2.2 Median

The median is a measure of central tendency that represents the middle value in a dataset when the data is arranged in order. If the number of values in data set is odd, the median is the middle values. If the number of values in data set is even, the median is the average of the two middle numbers. From a biological perspective, it provides a better representation of the central tendency in cases like population size, disease prevalence, or body measurements, where data might not be normally distributed.

Formula

There is no single formula to calculate the median but there are steps to calculate it.

Steps to calculate the Median

Step 1: Arrange the dataset in ascending or descending order.

Step 2: Count the total number of values in the dataset.

Step 3: If the number of values in the dataset is odd, the median is the value at the position $\left(\frac{n+1}{2}\right)^{th}$.

Step 4: If the number of values in the dataset is even, the median is the average of values at the positions $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$.

Example 1 (odd number of values)

A biologist studied the average height (cm) of five plants in a garden. The heights of the plants are 50cm, 65cm, 60cm, 75cm, and 80cm.

Step 1: 50cm, 60cm, 65cm, 75cm and 80cm.

Step 2: n =5 plants

Step 3: n =5 is odd, the median is the value at the position $\left(\frac{5+1}{2}\right)=\left(\frac{6}{2}\right)=3$.

Example 1 (even number of values)

A biologist studied the average weight (kg) of six different animals in a zoo. The weights of the animals are 300kg, 400kg, 330kg, 315kg, 450kg and 500kg.

Step 1: 300kg, 315kg, 330kg, 400kg, 450kg and 500kg.

Step 2: n =6 animals

Step 3: n = 6 is even, the median is the average of value at the position $\left(\frac{6}{2}\right)=3$ and $\left(\frac{6+2}{2}\right)=\left(\frac{8}{2}\right)=4$.

Median= 330 + 400/2=365kg

22.2.3- Mode

The mode is the value or values that appear most frequently in a dataset. From a biological perspective, it can be used to understand the most common characteristics or traits in a biological sample, such as the most frequent number of flowers on a plant, the most common size of seeds in a population, or the most frequently occurring weight in a sample of animals.

Formula

There is no specific formula to calculate mode in **ungrouped data**. The mode is simply the value or values that appear more frequently in a data set.

For **grouped data** the mode can be calculated by using the given formula:

$$\text{Mode} = l + \frac{(fm - f_1)}{(fm - f_1) + (fm - f_2)} \times h$$

Where:

l = lower boundary of the modal class.

fm = frequency of the model class.

F_1 = frequency of the class before the modal class.

f_2 = frequency of the class after the modal class.

h = class width (the difference between upper and lower limits of any class interval).

Steps to calculate the Mode for ungrouped data with example

A biologist studied the height (cm) of ten plants in a garden. The heights of plants are 50cm, 60 cm, 50 cm, 55 cm, 50 cm, 65 cm, 60 cm, 70 cm, 55 cm and 50 cm.

Step 1: Arrange the data in ascending order.

50 cm, 50 cm, 50 cm, 50 cm, 55 cm, 55 cm, 60 cm, 60 cm, 65 cm, 70 cm.

Step 2: Count the frequency of each value.

50 cm appears 4 times
55 cm appears 2 times
60 cm appears 2 times
65 cm appears 1 time
70 cm appears 1 time

Step 3: Identify the mode.

The number 50 appears the most frequently i.e. 4 times. Therefore, the mode for this dataset is 50 cm.

Steps to calculate the Mode for grouped data with example

Consider we are studying the weight (kg) distribution of animals in a zoo.

Class	Class interval Weight (kg)	Frequency
I	10-20	5
II	20-30	12
III	30-40	18
IV	40-50	10
V	50-60	3

Step 1: Identify the modal class.

The class interval (30-40) Class III is the modal class with the highest frequency (18).

Step 2: Put the values in the formula to calculate mode for grouped data.

l =30 (lower boundary of the modal class)
fm =18 (frequency of the modal class)
f_1 =12 (frequency of the class before the modal class)
f_2 =10 (frequency of the class after the modal class)
h =10 (class width)

$$\text{Mode} = l + \frac{(fm - f_1)}{(fm - f_1) + (fm - f_2)} \times h$$

$$= 30 + \frac{(18 - 12)}{(18 - 12) + (18 - 10)} \times 10$$

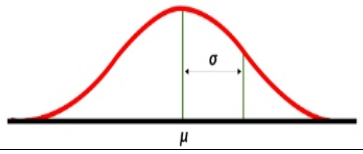
$$= 30 + \frac{6}{6 + 8} \times 10 = 30 + \frac{6}{14} \times 10$$

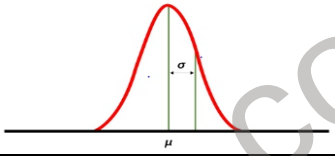
$$= 34.29\, kg$$

The mode of the givens dataset is 34.29 kg.

22.2.4- Standard deviation

A standard deviation is a measure of how dispersed the data is in relation to the mean. Low, or small, standard deviation indicates data are clustered tightly around the mean, and high, or large, standard deviation indicates data are more spread out. A standard deviation close to zero indicates that data points are very close to the mean, whereas a larger standard deviation indicates data points are spread further away from the mean.





a) **The curve is more spread out and has a high standard deviation**	b) **The curve is more clustered around the mean and has a lower standard deviation**

In biology, standard deviation is commonly used to assess the variability in measurements, such as the growth rates of plants, the concentration of enzymes in a sample, or any other biological data.

Formula

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Where:

σ = standard deviation

x_i = each individual data point in the set

μ = mean

N = total number of data points.

Steps to calculate Standard Deviation with example

A biologist studied the heights of 9 different plants of the same species. The heights of the plants are 56cm, 65cm, 74cm, 75cm, 76cm, 77cm, 80cm, 81cm and 91cm.

Step1: Find the mean.

Step2: Subtract the mean from each data point to find the deviation of each point.

Step3: Square the result of each deviation which eliminates negative values and emphasizes larger deviations.

Step4: Find the average of the squared deviations.

Step5: Take the square root of the result from step 4. This is the standard deviation.

Table: Calculation of standard deviation for plant heights

Height in (cm) x_i	Mean μ	Subtract mean from each data point $(x_i - \mu)$	Square the result of each deviation $(x_i - \mu)^2$	Sum of squared deviations $\sum(x_i - \mu)^2$	Variance (Average of the squared deviations) $\sum(x_i - \mu)^2/N$	Standard deviation $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
56	75	56 – 75 = -19	361	784	87.11	9.33
65		65 – 75 = -10	100			
74		74 – 75 = -1	1			
75		75 – 75 = 0	0			
76		76 – 75 = 1	1			
77		77 – 75 = 2	4			
80		80 – 75 = 5	25			
81		81 – 75 = 6	36			
91		91 – 75 = 16	256			

22.2.5 Range

The range is a measure of the spread or dispersion of a set of data. It is the simplest form of variability and is used to understand the extent of variation in the dataset. The range represents the difference between the maximum and minimum values in a dataset. A larger range suggests a greater spread of data, meaning the data points are more dispersed from each other while a smaller range indicates that the data points are closely clustered around a central value.

Formula

$$\text{Range} = X_{\max} - X_{\min}$$

Where:

X_{max} = maximum value in the data set.

X_{min} = minimum value in the data set.

Steps to calculate the Range with example

Consider a dataset representing the ages of a group of people: 25, 20, 22, 30, 42, 45, 50.

Step 1: Arrange the data in ascending order.

20, 22, 25, 30, 42, 45, 50

Step 2: Identify the greatest value in the dataset.

50

Step 3: Identify the smallest value in the dataset.

20

Step 4: Subtract the maximum value from the minimum value to calculate the range.

$$\text{Range} = X_{\max} - X_{\min}$$

$$\text{Range} = 50\text{-}20 = 30$$

22.2.6 Percentile

A percentile is a measure that indicates the value in a dataset below which a given percentage of observations in a group of data falls. For instance, the 25th percentile, also known as the first quartile (Q1), is the value below which 25% of the data points lie. Similarly, the 50th percentile, or the median, is the value below which 50% of the data points lie, and the 75th percentile, or the third quartile (Q3), is the value below which 75% of the data points lie. Percentiles are particularly useful in biostatistics as they help in understanding the distribution of biological data which is skewed or normally distributed.

Formula

$$R = R = \frac{P}{100} \quad (n + 1)$$

Where

R = rank

P = desired percentile

n = number of observations in the data set

If R is an integer, the value at the Rth position in the ordered dataset is the percentile. If R is not an integer (e.g., 4.4), interpolation between the two nearest ranks is used to estimate the percentile value. For example, if it lies between 4th and 5th data points, percentile value=

Value at 4th + 0.4 X (value at 5th - value at 4th)

Steps to calculate percentile with example

Suppose we have the following dataset representing the heights (cm) of 10 individuals: 160, 162, 168, 165, 170, 175, 172, 178, 180, 182.

To calculate the 40th percentile:

Step 1: Arrange the data in ascending order.

160, 162, 165, 168, 170, 172, 175, 178, 180, 182.

Step 2: Determine the total number of observations.

n = 10

Step 3: Calculate the rank.

$R = P/100 \text{ X } (n+1)$

R = 40/100 X (10+1) = 4.4

Step 4: Locate the rank in the dataset and interpolate if necessary.

The rank is 4.4, which is not an integer. This means that the 40th percentile lies between 4th (168) and 5th (170) data points.

Interpolate between the 4th and 5th values:

Percentile value = value at 4th + 0.4 X (value at 5th - value at 4th)

Percentile value = 168 + 0.4 X (170-168) = 168 + 0.4 X 2 = 168 + 0.8 = 168.8

Step 5: Interpret the result.

The 40th percentile value is 168.8. This means that 40% of the observations fall below or equal to 168.8 in this data set.

22.3 CHART

22.3.1 Bar Chart

A bar chart (or bar graph) is a graphical representation used to display categorical data with rectangular bars, where the length or height of each bar is proportional to the value it represents. Bar charts are used to compare quantities across distinct categories or groups. In biology, bar charts are valuable for showing data like the counts of different species, experimental groups, or measurement results from biological samples. The key characteristics of a bar chart are:

i) Categories are represented by the x-axis (horizontal axis).

ii) Values or frequencies are represented by the y-axis (vertical axis).

iii) Bars can be oriented vertically or horizontally.

Steps to create a bar chart with example

A team of research students conducted an insect diversity study in a local forest and counted the number of individuals for five common insect species over a month.

Table: Relative abundance of the insect species

Insect species	Number of individuals counted	Percentage (%)
Ladybug	45	16
Ant	120	41
Butterfly	30	10
Bee	75	26
Dragonfly	20	7

Step 1: Purpose and data.

Purpose: To visually compare the abundance of five different insect species.

Categorical variable: Insect species (ladybug, ant, butterfly, bee, dragonfly).

Numerical variable: Number of individuals counted.

Step 2: Determine the axis.

X-axis (Categorical): Insect species
Y-axis (Numerical): Number of individuals counted.
Step 3: Scale the numerical axis.
Greatest value is 120 (Ants). Scale the y-axis from 0 to 130 or 140. An increment of 10 or 20 is used for good readability. Let's choose 0 to 140 with increment of 20.
Step 4: Draw the bars.
Draw a bar for ladybug up to 45 on the y-axis.
Draw a bar for ant up to 120 on the y-axis.
Draw a bar for butterfly up to 30 on the y-axis.
Draw a bar for bee up to 75 on the y-axis.
Draw a bar for dragonfly up to 20 on the y-axis.
Ensure bars are of equal width and evenly spaced.
Step 5: Add labels and title.
Chart title: Insect species abundance in local forest
X-axis label: Insect species
Y-axis label: Number of individuals

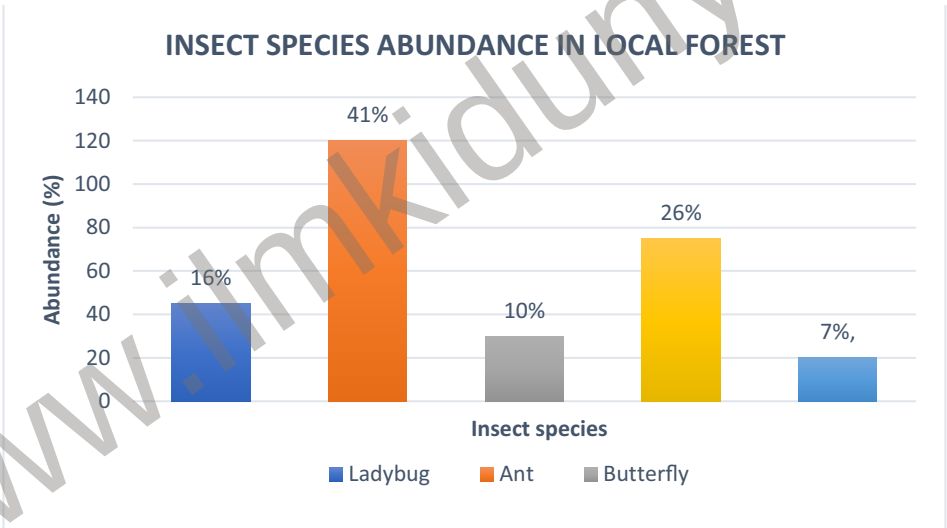


Fig. Bar chart showing relative abundance of insect species in a local forest

A **legend** is a key or guide on a chart, graph, or map that explains the meaning of symbols, colours, patterns, or labels used in the visual. It helps the viewer understand what different elements represent, making the data or information easier to interpret and analyse. For example, in a bar chart with different colours for each category, the legend identifies what each colour corresponds to.

22.3.2 Pie Chart

A pie chart is a circular statistical graph divided into sectors (slices), each representing a proportion of the whole dataset. The entire circle corresponds to 100% (or 360°), and each sector's angle is proportional to its category's share in the dataset. Pie charts are ideal for visualizing categorical data as portions of a whole, making it intuitive to compare different group's contributions in biology, such as types of cells, species distributions, or survey responses.

Steps to create a pie chart with example

A biology class surveys 100 students to find their blood types. The results are as follows:

Table: Surevy results of student blood types	
Blood type	**Number of students (frequency)**
A	30
B	20
AB	10
O	40

Step 1: Organize data

Arrange the categorical data in a table, noting the frequency or total for each category as shown in table.....

Step 2: Add up all frequency values to get the total sum

Total number of students= 30+20+10+40 = 100

Sum of frequency	Blood Type	No. of Student	Sector angle = Sum of values
30	A	30	$\frac{30}{100} \times 360^\circ = 108^\circ$
20+30=50	B	20	$\frac{20}{100} \times 360^\circ = 72^\circ$
10+50=60	AB	10	$\frac{10}{100} \times 360^\circ = 36^\circ$
40+60=100	O	40	$\frac{40}{100} \times 360^\circ = 144^\circ$
	Total	n =100	

Step 3: Calculate the sector angles (Sector angle = Category value/ total value X 100)

A= 30/100 X 360° = 108° %

B= 20/100 X 360° = 72° %

AB= 10/100 X 360° = 36° %

O= 40/100 X 360° = 144° %

Step 4: Draw the pie chart

Draw a circle by using a protractor, draw each sector angle in sequence (clock wise or counter clockwise) using the calculated angles in step 3. Start with a reference line (radius). Use different colours/shades to differentiate each category.

Step 5: Label and title

Clearly label each sector and provide a legend if needed. Add a descriptive title to the chart.

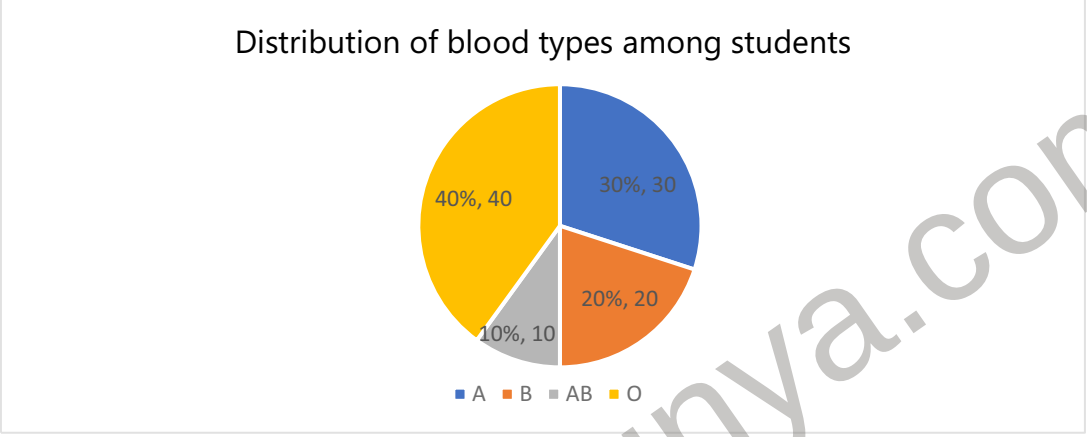


Fig. Pie chart showing distribution of blood types among students

X-axis data figure

The x-axis refers to the horizontal line on a Cartesian coordinate system or data figure such as a graph, chart, or plot. It is most commonly used to represent the independent variable. The x-axis provides the baseline against which other variables (typically shown on the y-axis) are measured. It is always drawn horizontally, running left to right, and often starts at the origin (0, 0), extending in the positive direction, though it can also include negative values when needed. In biological studies, the *x*-axis is often used to represent variables like time intervals, different species, or experimental conditions. It provides a **baseline for comparison** and helps visualize **trends or relationships** between variables in biological data.

Steps to create an x-axis data figure with example

Suppose we are measuring the activity of enzyme. The activity is measured at 5°C, 10°C, 20°C, 30°C and 40°C.

Step 1: Identify variables

Independent variable (x-axis): Temperature, measured in degrees Celsius (°C).

Dependent variable (y-axis): Rate of enzyme activity

Step 2: Determine scale and range

Decide the minimum and maximum values, and the intervals or categories for the x-axis. Temperatures at which enzyme activity is measured i.e. 5°C, 10°C, 20°C, 30°C, and 40°C.

Step 3: Label the axis

X-axis is labeled as Temperature (°C) and Y-axis is labeled as Enzyme Activity (µmol/min).

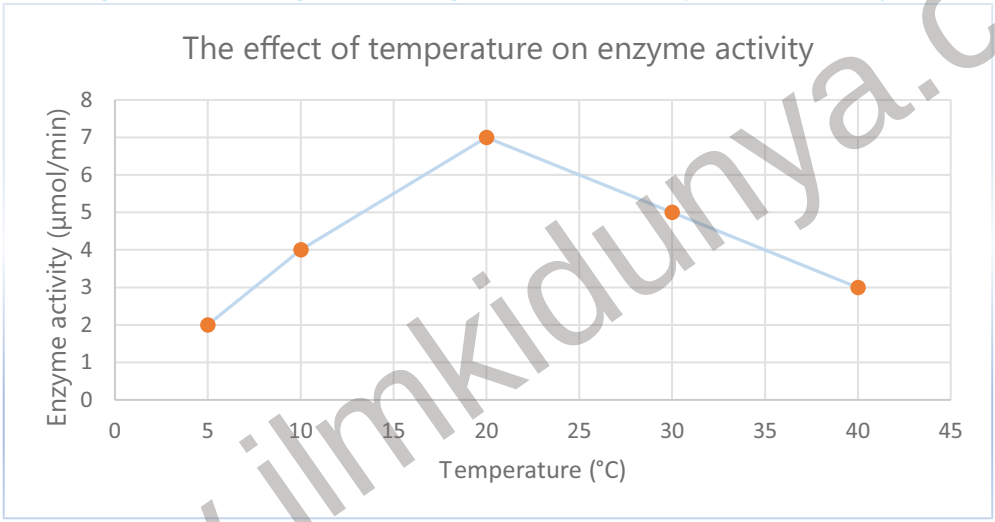
Step 4: Plot the data points

For each measurement, find its position on the x-axis and plot it against the corresponding y-axis value (the dependent variable). We measured these pairs: (5°C, 2 µmol/min), (10°C, 4 µmol/min), (20°C, 7 µmol/min), (30°C, 5 µmol/min), (40°C, 3 µmol/min).

Step 5: Adjust formatting

Make sure tick marks are evenly placed for this data. Temperature values are evenly spaced along the x-axis since the interval between points is similar. Connect the data points with a line to show the trend.

Fig. X-axis data figure showing the effect of temperature on enzyme activity



Error Bars

Error bars are graphical elements placed on bar charts (and other types of plots) to visually represent the variability or uncertainty in data. They help visualize how much the individual data points typically deviate from the central tendency (e.g., mean) represented by the bar.

Types of error bars

The two most common ways to sketch errors bars for a bar chart are:

Range error bars

Range error bars use the smallest (minimum) and greatest (maximum) data points in a set for each bar. The bar extends from the minimum to the maximum value, illustrating the full spread of the data.

Standard deviation (SD) error bars

Standard deviation error bars show the average distance of each data point from the mean. On a bar chart, each bar represents a group mean. **Standard deviation error bars** extend **above and below the mean** value of a dataset by one or more standard deviations (±SD).

Steps to sketch error bars on a bar chart

The researchers are measuring the average plant height (cm) in three different light conditions. They grow bean plants in three groups:

Group A: Low light
Group B: Medium light
Group C: High light

After four weeks they measure the heights of the plants.

	Group A	Group B	Group C
Plant 1	10	14	18
Plant 2	11	16	20
Plant 3	9	15	19

Step 1: Calculate the mean of each group
Group A Mean of = (10 + 11 + 9)/3 = 10 cm
Group B Mean of= (14 + 16 + 15)/3 = 15 cm
Group C Mean of= (18 + 20 + 19)/3 = 19 cm
The bars on the bar chart are drawn at heights 10, 15, and 19 for Groups A, B, and C, respectively.

Step 2: Calculate the spread

A) For SD error bars:
Group A: SD ≈ 1 (Data: 10, 11, 9)
Group B: SD ≈ 1 (Data: 14, 16, 15)
Group C: SD ≈ 1 (Data: 18, 20, 19)

B) For range error bars:
Group A: min = 9, max = 11
Group B: min = 14, max = 16
Group C: min = 18, max = 20

Step 3: Draw the error bars

SD error bars: Start at the mean (top of each bar), extend up and down by 1 cm (Mean ±1 SD).
Group A: from 9 (10-1) to 11 (10+1)
Group B: from 14 (15-1) to 16 (15+1)
Group C: from 18 (19-1) to 20 (19+1)

Range error bars: Extend from the minimum to maximum value for each group. Add a horizontal cap at each end of the vertical error bar for clarity.

Table: Statistical summary of plant measurements

Group	Plant 1	Plant 2	Plant 3	Mean	Range (max-min)	Range error (±)= Range/2	Standard deviation error(±SD)
A	10	11	9	10	2	1	1
B	14	16	15	15	2	1	1
C	18	20	19	19	2	1	1

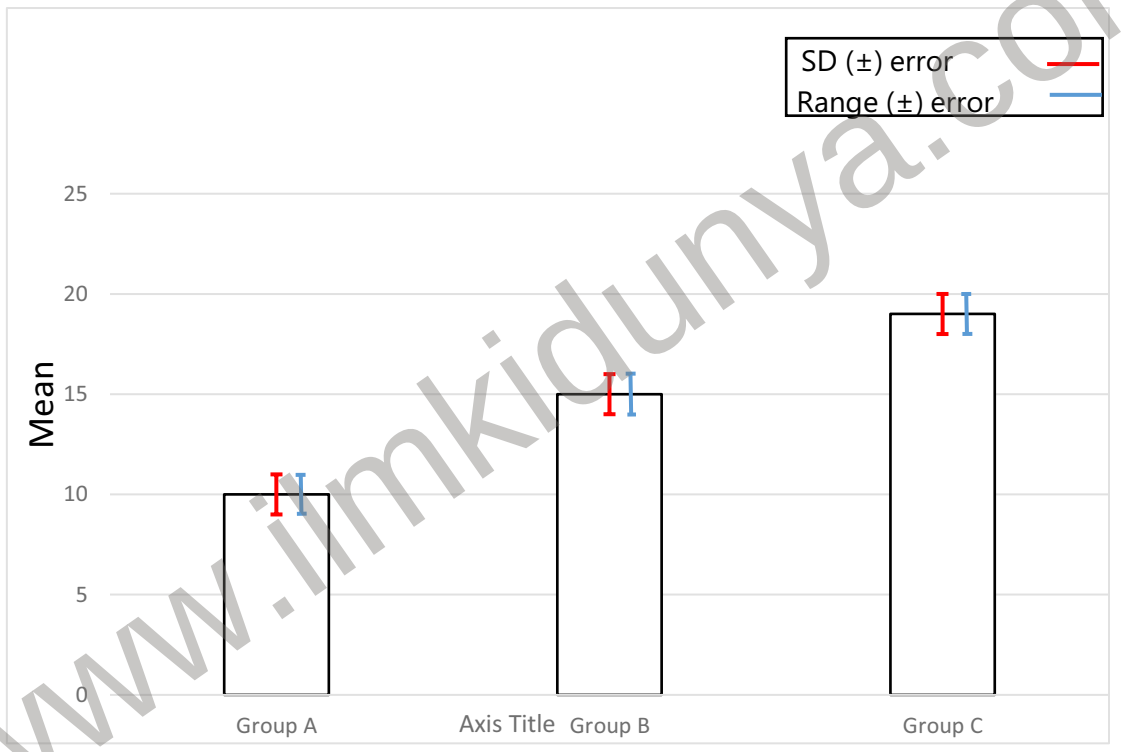


Fig. Graph showing average plant height in different light conditions with error bars

22.3 EXPERIMENTAL DESIGN

When scientists want to test an idea, they plan an experiment carefully. This planning is called **experimental design**. They decide subjects or participants and control the conditions so results are fair. Then, they change one thing (the independent variable) and see how it affects another thing (the dependent variable). This help them find out if their hypothesis is right or wrong by looking at how different variables are connected.

To do this effectively, scientists identify different types of variables. The **independent variable** is the factor that the experimenter changes or manipulates to observe its effect. The **dependent variable** is what is measured or observed; it changes in response to the independent variable. **Control variables** (or constants) are all the other factors that must be kept the same throughout the experiment to ensure a fair test.

A **control group** is a group in the experiment that does not receive the treatment or change in the independent variable. It serves as a baseline to compare the results of the experimental group, helping to show whether the independent variable actually causes any effect. An **experimental group** is a group that is exposed to the specific treatment, condition, or environmental change being tested in an experiment. Their responses are measured and compared to a control group to assess the effect of that treatment on biological processes or characteristics.

Steps for experimental design:

1. State the question of your experiment first. This clearly defines what you want to find out.
2. Formulate a hypothesis based on your research e.g., what you think will happen, based on existing knowledge.
3. Clearly identify variables to understand what you will change (independent variable), measure (dependent variable) and keep constant (control variable).
4. Set up groups e.g., control group (does not receive the treatment or change in independent variable) and experimental group (receives the treatment or change).
5. Plan the procedure and outline detailed steps on how experiments will be conducted e.g., how subjects are selected and assigned to groups, how each group treated, the duration and frequency of measurement and how data will be collected.
6. Use appropriate methods (such as averaging, graphing or statistical tests) to compare outcomes between groups and assess whether there is a significant effect
7. Draw conclusions in relation to the original hypothesis and research question. Accept or reject hypothesis on the basis of results.

Experiment: Effect of high-fat diet on weight gain in albino mice

1.	**Question:**	Does a high-fat diet cause greater weight gain in mice compared to normal diet?
2.	**Hypothesis:**	Mice given a high-fat diet will gain more weight than those on a normal diet.
3.	**Identify variables:**	**Independent variable:** Type of diet **Dependent variable:** Weight gain **Control variables:** Mouse age, strain, sex, cage, environment (light, temperature) and feeding schedule etc
4.	**Set up groups:**	**Control group:** Mice on normal diet **Experimental group:** Mice on a high-fat diet
5.	**Procedure outline:**	1. Select albino mice that are matched by age, strain, and sex. 2. Randomly assign albino mice to control and experimental groups. 3. House all albino mice under identical conditions (same cages, light, temperature, humidity). 4. Administer the assigned diet to each group for a set period (e.g., 8 weeks). 5. Measure and record each albino mice weight weekly.
6.	**Analyse results:**	Use statistical tests or graphs to compare the average weight gain between groups at the end of the experiment.
7.	**Draw conclusions:**	The experiment will determine whether a high-fat diet leads to greater weight gain in mice compared to a normal diet. If the high-fat diet group shows a statistically significant increase in average weight gain, the hypothesis will be supported, suggesting that diet type influences weight gain. However, if there is no significant difference between groups, or if the normal diet group gains more weight, the hypothesis will be rejected, indicating that factors other than diet type may play a greater role in influencing weight gain under the given experimental conditions.

EXERCISE

SECTION 1: MULTIPLE CHOICE QUESTIONS

1. Which of the following is a primary data collection method?
   (a) Government reports (b) Published articles
   (c) Experiment and surveys (d) Online databases
2. What does mode represent in a dataset?
   (a) The average of values (b) The value occurring most frequently
   (c) The difference between two values (d) None of these
3. Which statistical measure is also called the 50^{th} percentile?
   (a) Mean (b) Median
   (c) Mode (d) Range
4. If the maximum value in a dataset is 50 and the minimum value is 20, what is the range?
   (a) 30 (b) 25
   (c) 20 (d) 115
5. What does it mean if the 40^{th} percentile value is 148cm?
   (a) 40% of values are above 148cm
   (b) 40% of values are below or equal to 148cm
   (c) All the values are equal to 148cm
   (d) 60% of values are above 148cm
6. In a pie chart, what does the full circle represent?
   (a) 180° (b) 280° (c) 380° (d) 360°
7. What do error bars represents in a chart?
   (a) Average (b) Variability or uncertainty in data
   (c) Both of these (d) None of these
8. In an experiment, what does the dependent variable represent?
   (a) The factor kept constant (b) The factor being changed
   (c) The factor ignored (d) All of these
9. What is the main purpose of including a control group in an experiment?
   (a) To test the dependent variable (b) To provide a baseline for comparison
   (c) To collect random data (d) All of these
10. Which group in an experiment receives the treatment?
    (a) Control group (b) Dependent group
    (c) Experimental group (d) Independent group

SECTION 2: SHORT QUESTIONS

1. Explain why biostatistics is important in medical research?
2. Describe the difference between descriptive and inferential statistics?

3. What is the relationship between percentiles and quartiles?
4. How are independent and dependent variables related in experimental design?
5. The height of six children are 110, 115, 118, 120, 122, and 125. What is the median height of the children?
6. What does a legend in a chart present?
7. What are error bars?

SECTION 3: LONG QUESTIONS

1. Explain the concepts of mean, median, and mode as measures of central tendency. Illustrate with examples when each measure would be most appropriate.
2. Describe standard deviation, range, and percentiles as measures of variability. Compare their advantages and limitations in interpreting biological data?
3. What are error bars? Explain their types (range error bars, standard deviation error bars) and significance in experimental biology. Illustrate your answer with a hypothetical dataset?
4. Compare and construct bar charts and pie charts. Which situations favour one over the other? Support your answer with real-world biological or health data examples.
5. Discuss different types of data presentation tools in biostatistics (bar chart, pie chart and error bars). For each, explain an appropriate biological example where it would be the best choice.

INQUISITIVE QUESTIONS

1. A pharmaceutical company claims their drug improves recovery time from flu. Propose an experimental design using biostatistics principles, and explain how you would ensure validity and reliability of results?
2. Evaluate the usefulness of range versus standard deviation in representing variability. Which one would you trust more in medical research and why?
3. During an epidemic, mortality rate estimates vary depending on whether the mean, median, or percentile is used. Discuss how policymakers might misinterpret such statistics and how biostatistics can prevent this.