Applied Biostatistics for Third MBBS Students

A practical guide to estimation, confidence intervals and hypothesis testing

Duration: Two sessions of approximately 2 hours each
Prerequisites: None assumed; essential concepts are revised.
Approach: Understand Identify Calculate Interpret Apply

Part I. Introduction and Learning Plan

1. Why do we need biostatistics?

Imagine that you want to know the average amount of time Third MBBS students spend studying every day. There are 150 students in the batch. You randomly select 25 students and find that their average study time is 3 hours per day. Can you conclude that all 150 students study for an average of exactly 3 hours daily? Not necessarily. Another random sample might produce a different average.

Inferential statistics helps us use sample data to draw conclusions about populations while accounting for uncertainty. Its two main applications here are **estimation** (including confidence intervals) and **hypothesis testing**.

2. Learning objectives

By the end of these sessions, students should be able to:

No.	Learning objective
1	Distinguish populations, samples, parameters and statistics.
2	Identify nominal, ordinal, interval and ratio scales of measurement.
3	Recall mean, standard deviation, z-score and normal distribution.

4	Calculate standard errors and confidence intervals using normal critical values.
5	Calculate and interpret intervals based on 1, 2 and 3 standard errors.
6	Explain null and alternative hypotheses, alpha, p-values and statistical significance.
7	Select and apply one-sample and independent two-sample z-tests when appropriate.
8	Calculate and interpret a chi-square test of independence.
9	Identify situations in which the tests taught are inappropriate.
10	Interpret findings without unsupported claims.

3. Suggested teaching schedule

Session	Component	Duration
1	Basic concepts, variables and scales	25 min
1	Mean, SD, normal distribution and z-scores	20 min
1	Standard error and confidence intervals	45 min
1	Hypotheses, alpha and p-values	30 min
2	Recap and test selection	15 min
2	One-sample and two-sample z-tests	40 min
2	Chi-square test	40 min
2	Integrated practice and formative assessment	25 min

How to use this guide: Each section begins with a recap, followed by explanations, worked examples and practice. Solutions and statistical reference tables are included.

Part II. Foundations of Biostatistics

4. Population, sample, parameter and statistic

Recap: What are we studying?

A college has 500 MBBS students. You randomly select 50 to estimate average daily screen time.

Term	Meaning	Example
Population	Entire group about which information is sought	All 500 students
Sample	Selected subset of the population	50 selected students
Parameter	Numerical characteristic of the population	True population mean screen time
Statistic	Numerical characteristic calculated from the sample	Mean screen time of 50 students

Symbols used:

- μ (mu): population mean.
- $\bar{x}$ (x-bar): sample mean.
- σ (sigma): population standard deviation.
- s: sample standard deviation.
- n: sample size.

Sampling error is the difference between a sample statistic and its population parameter arising from random sampling.

Practice 1

A researcher wants to estimate average monthly expenditure on food among 800 hostel residents. She randomly selects 100 students. Identify the population, sample, parameter and statistic.

Solution: Population: all 800 hostel residents; sample: 100 selected students; parameter: true mean monthly food expenditure of all 800; statistic: sample mean monthly food expenditure of the 100 students.

5. Variables and scales of measurement

Recap: What is a variable?

A **variable** is a characteristic that takes different values across observations. Variables can be categorical or numerical.

Scale	Definition	Examples
Nominal	Categories without inherent order	Blood group, smoking status
Ordinal	Ordered categories without necessarily equal differences	Mild/moderate/severe pain
Interval	Equal numerical intervals without an absolute zero	Temperature in °C
Ratio	Equal numerical intervals with a meaningful zero	Height, weight, age, income

A **continuous** variable can take values across a measurement range (e.g., height). A **discrete** variable has countable values (e.g., number of siblings). A numerical variable is not automatically suitable for a normal-theory test.

Why this matters: Means may be appropriate for quantitative measurements; nominal categories should be summarised as counts or percentages. Chi-square can examine association between categorical variables.

Practice 2

Identify the scale of measurement.

Variable	Scale (fill in)
Blood group	
Body weight	
Satisfaction: poor/fair/good/excellent	
Temperature in °C	
Number of siblings	
Smoking status	

Examination rank	
Height	

Answers (in order): Nominal; ratio; ordinal; interval; ratio (discrete); nominal; ordinal; ratio.

Part III. Understanding the Normal Distribution

6. Mean and standard deviation

Recap: Central tendency and dispersion

The **mean** is the arithmetic average:

$$\bar{x} = \frac{\sum x_i}{n}$$

Here x_i is an individual observation, $\sum$ means summation, and n is the number of observations.

Solved example 1: Daily study hours

Five students report 2, 3, 3, 4 and 8 hours of study per day.

$$\bar{x} = \frac{2+3+3+4+8}{5} = \frac{20}{5} = 4 \text{ hours}$$

The **standard deviation (SD)** measures dispersion around the mean. For a sample:

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

Here s is the sample SD, $x_i - \bar{x}$ is the deviation of observation i from the sample mean, and $n - 1$ is the denominator used for sample variance. A larger SD indicates greater variability. **SD is not the uncertainty of the estimated population mean**; that is described by standard error.

Worked calculation: Standard deviation of study hours

Use the five study times from Solved example 1: 2, 3, 3, 4 and 8 hours. Their sample mean is 4 hours.

Step 1. Subtract the mean from each observation. **Step 2.** Square each deviation. **Step 3.** Add the squared deviations.

Study hours (x_i)	Deviation ($x_i - \bar{x}$)	Squared deviation ($(x_i - \bar{x})^2$)
2	−2	4
3	−1	1
3	−1	1
4	0	0
8	4	16
Total	**0**	**22**

Step 4. Divide the sum of squared deviations by $n - 1$ to obtain the sample variance:

$$s^2 = \frac{22}{5 - 1} = \frac{22}{4} = 5.5 \text{ hours}^2$$

Step 5. Take the square root to obtain the sample SD:

$$s = \sqrt{5.5} \approx 2.35 \text{ hours}$$

Interpretation: Study times vary around their mean of 4 hours, with a sample SD of approximately 2.35 hours. The SD is expressed in hours, whereas the variance is expressed in squared hours. The unusually long study time of 8 hours contributes substantially to the SD.

7. The normal distribution

Recap: Distribution of observations

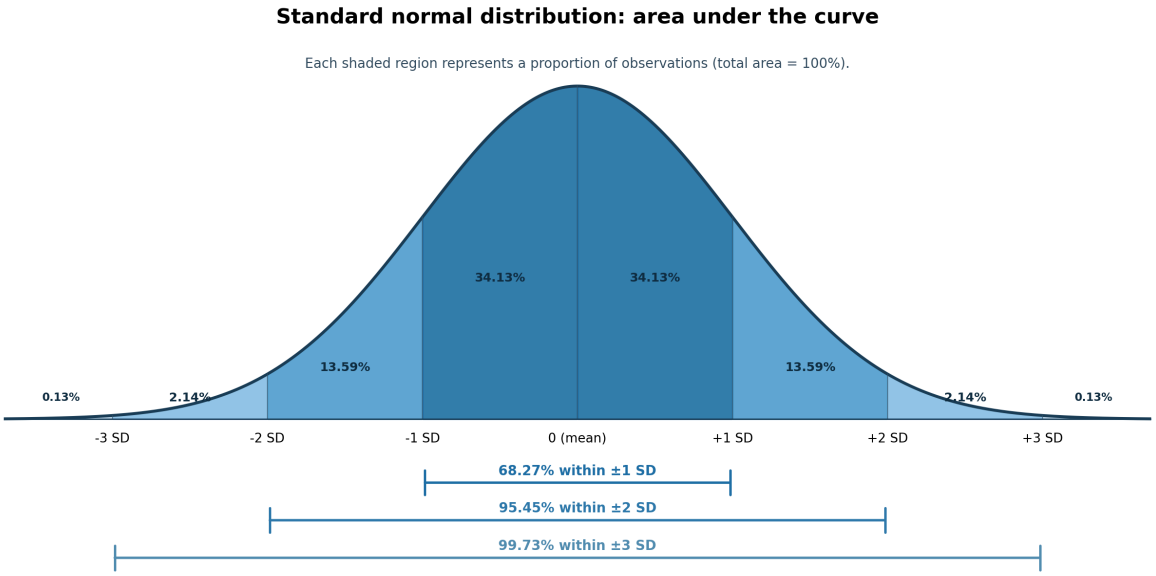
A **normal distribution** is a symmetrical, bell-shaped probability distribution. Its mean, median and mode coincide; its shape is determined by mean and SD; total area under the curve is 1.

For an approximately normal distribution:

Range	Approximate proportion of individual observations
Mean ± 1 SD	68.27%
Mean ± 2 SD	95.45%
Mean ± 3 SD	99.73%

Visual guide: Area under the standard normal curve

The graph divides the area under the normal curve into regions measured in standard deviations (SD) from the mean. Each half of the curve contains 50% of the total area. The values inside the shaded bands are percentages of **individual observations** under a normal distribution, not confidence intervals for a population mean.



Standard normal distribution with shaded areas between minus 3 and plus 3 SD

For example, the area from the mean to +1 SD is approximately 34.13%; the area between +1 SD and +2 SD is approximately 13.59%. Adding the areas symmetrically on both sides gives approximately 68.27% within ±1 SD, 95.45% within ±2 SD, and 99.73% within ±3 SD. Percentages are rounded.

Solved example 2: Examination marks

If marks are approximately normally distributed with mean 60 and SD 10, about 68.27% fall between 50 and 70; 95.45% between 40 and 80; and 99.73% between 30 and 90. These are ranges of **individual observations**, not confidence intervals for a mean.

8. z-scores

Recap: Mean, SD and the normal distribution

A **z-score** tells how many population SDs an observation is above or below the population mean:

$$z = \frac{x - \mu}{\sigma}$$

Here x is the observation, μ the population mean, and σ the population SD.

Solved example 3: Examination performance

Mean score = 60, SD = 10, student's score = 80.

$$z = \frac{80 - 60}{10} = 2$$

Interpretation: The student scored 2 SD above the mean. Under normality, approximately 97.72% score below 80 (Annexure 1).

Practice 3

Population mean height = 165 cm, SD = 10 cm. Calculate and interpret z for heights 175, 155 and 165 cm.

Solutions: $+1$ (one SD above); -1 (one SD below); 0 (at the mean).

Part IV. Standard Error and Confidence Intervals

9. Standard error

Recap: Sampling variability

Repeated random samples from the same population yield different sample means. Their distribution is called the **sampling distribution of the mean**. The standard deviation of this distribution is the **standard error (SE)**.

If population SD is known:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

If it is unknown, estimate SE with sample SD:

$$\widehat{\mathrm{SE}} = \frac{s}{\sqrt{n}}$$

Here $\sqrt{n}$ is the square root of sample size, σ is population SD, and s is sample SD.

Solved example 4: Daily screen time

A random sample of 100 students has mean screen time 5 hours/day and sample SD 2 hours/day.

$$\widehat{\mathrm{SE}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2 \text{ hours}$$

Interpretation: The estimated standard deviation of the sampling distribution of the sample mean is 0.2 hours, not the spread of individual screen times.

Sample size	Estimated SE when SD = 2
25	0.40
100	0.20
400	0.10

Increasing sample size decreases SE if the underlying variability remains unchanged.

10. Confidence intervals

Recap: Point estimate and uncertainty

A sample mean is a **point estimate** of the population mean. A **confidence interval (CI)** is calculated by a procedure designed to contain the fixed population parameter in a specified proportion of repeated samples under its assumptions.

For a normal-based interval:

$$\mathrm{CI} = \bar{x} \pm z^* \times \mathrm{SE}$$

- $\bar{x}$: sample mean.

- z^*: normal critical multiplier for the selected confidence level.
- SE: standard error of the sample mean.
- $z^* \times SE$: **margin of error**.
- $\pm$: calculate both lower and upper limits.

10.1 One, two and three standard errors

Interval	Formula	Approximate confidence level
Mean ± 1 SE	$\bar{x} \pm 1(SE)$	68.27%
Mean ± 2 SE	$\bar{x} \pm 2(SE)$	95.45%
Mean ± 3 SE	$\bar{x} \pm 3(SE)$	99.73%

Conventional confidence intervals use more precise multipliers:

Confidence level	z^*
68.27%	1.000
90%	1.645
95%	1.960
99%	2.576
99.73%	approximately 3.000

Important: Mean ± 2 SE is approximately a **95.45%** interval, not an exact 95% interval. The conventional 95% CI uses **1.96 SE**. Intervals based on SE concern uncertainty in the population mean; ranges based on SD concern variation in individual observations.

Solved example 5: Daily screen time

Mean = 5 hours/day; SD = 2 hours/day; $n = 100$.

Step 1: Calculate SE.

$$SE = \frac{2}{\sqrt{100}} = 0.2 \text{ hours}$$

Step 2: Calculate intervals.

Confidence	Calculation	Interval (hours/day)

68.27%	$5 \pm 1(0.2)$	4.8–5.2
95.45%	$5 \pm 2(0.2)$	4.6–5.4
99.73%	$5 \pm 3(0.2)$	4.4–5.6
Conventional 95%	$5 \pm 1.96(0.2)$	4.608–5.392

Interpretation: The estimated population mean daily screen time is 5 hours. The conventional 95% CI is approximately 4.61–5.39 hours. If this sampling and interval-construction procedure were repeated many times, approximately 95% of resulting intervals would contain the true population mean, assuming the model holds.

Do not say: 95% of students have screen times in this interval; there is a 95% probability that the fixed population mean lies in this particular observed interval; or 95% of future sample means will fall within this particular interval.

10.2 Why does a higher confidence level give a wider interval?

For the same data and SE, a larger critical multiplier gives a wider interval. A wider interval has higher long-run coverage but less precision.

Practice 4: Sleep duration

A random sample of 64 hostel students has mean daily sleep duration 7 hours and sample SD 1.6 hours. Calculate SE; intervals using 1, 2 and 3 SE; and the conventional 95% CI. Interpret the latter.

Solution:

$$\text{SE} = \frac{1.6}{\sqrt{64}} = \frac{1.6}{8} = 0.2$$

Interval	Calculation	Limits (hours)
68.27%	$7 \pm 1(0.2)$	6.8–7.2
95.45%	$7 \pm 2(0.2)$	6.6–7.4
99.73%	$7 \pm 3(0.2)$	6.4–7.6
95%	$7 \pm 1.96(0.2)$	6.608–7.392

The conventional 95% CI estimates the population mean sleep duration as approximately 6.61–7.39 hours, subject to sampling and modelling assumptions.

Practice 5: Calculate intervals

Calculate SE and intervals using 1, 2 and 3 SE.

Question	Context	Mean	SD	n
A	Daily mobile phone use (hours)	4	1.5	100
B	Daily travel time to college (minutes)	40	12	144
C	Daily food expenditure ( )	150	40	64

Solutions:

Question	SE	±1 SE	±2 SE	±3 SE
A	0.15	3.85–4.15	3.70–4.30	3.55–4.45
B	1.00	39–41	38–42	37–43
C	5.00	145–155	140–160	135–165

All are intervals for population means, not ranges of individual observations.

Assumptions: Normal-based CI calculations require independent observations and an approximately normal sampling distribution of the mean. Random sampling supports population inference. When population SD is unknown, the normal CI using sample SD is an approximation; a t-based CI is generally preferred, especially with small samples.

Part V. Introduction to Hypothesis Testing

11. What is a statistical hypothesis?

Recap: Sampling variation

Two samples from the same population can have different means. An observed difference does not automatically establish a population difference.

Everyday example: Biscuit packet weights

A manufacturer claims an average packet weight of 100 g. A random sample of 100 packets has mean 98 g. Does this establish the claim is false? Not immediately: sampling variation must be considered.

The **null hypothesis**, H_0, is the statement being tested:

$$\mathrm{H}_0 : \mu = 100 \text{ g}$$

The **alternative hypothesis**, H_1 or H_a, is the competing claim. For a two-sided test:

$$\mathrm{H}_1 : \mu \models 100 \text{ g}$$

A **two-sided** alternative permits a difference in either direction. A **one-sided** alternative specifies a direction, e.g. $\mathrm{H}_1 : \mu < 100$. All calculation exercises in this module use two-sided alternatives unless stated otherwise.

12. Significance level: alpha

Recap: Decisions under uncertainty

Even if the population mean really is 100 g, a random sample may give an unusually low mean. Rejecting a true H_0 would be a **Type I error**.

The **significance level**, denoted α (alpha), is the prespecified threshold that controls the probability of Type I error for a correctly calibrated test. For a test of exact size:

$$\mathrm{P}\left(\text{reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ true}\right) = \alpha$$

Often $\alpha = 0.05$. If the null hypothesis were true and we repeated the procedure many times, approximately 5% of tests would incorrectly reject H_0 for an exact 5% test. **Alpha is chosen before examining the data. It is not the probability that H_0 is true.**

A **Type II error** occurs when we fail to reject a false H_0. Its probability is β (beta). **Power** is $1 - \beta$: the probability of rejecting H_0 for a specified true alternative. Power calculations are outside this module.

Reality	**Reject H_0**	**Do not reject H_0**
H_0 true	Type I error	Correct decision
H_0 false	Correct decision	Type II error

13. Understanding p-values

Recap: Alpha versus observed evidence

Alpha is chosen before data collection or analysis. A p-value is calculated from the observed data and a statistical model.

A **p-value** is the probability, **assuming H_0 and the statistical model are true**, of obtaining a test statistic at least as extreme as the one observed, in the direction(s) specified by H_1.

If $\mathrm{p} = 0.03$, then under H_0 and the test assumptions, a result at least this extreme has probability 3%. **It does not mean that H_0 has a 3% probability of being true.**

At $\alpha = 0.05$:

p-value	Decision	Description
0.01	Reject H_0	Statistically significant
0.04	Reject H_0	Statistically significant
0.05	Reject H_0 under $\mathrm{p} \leq \alpha$	At threshold
0.08	Do not reject H_0	Not statistically significant
0.30	Do not reject H_0	Not statistically significant

A p-value does **not** directly tell us whether H_0 is true, how large the effect is, whether the effect is clinically important, whether the study is unbiased, or whether the association is causal.

Solved example 6: Revision methods

A study finds a mean score difference of 4 marks between two revision methods, with $\mathrm{p} = 0.02$. Since $0.02 < 0.05$, reject H_0 at the 5% level. **Appropriate:** There is statistically significant evidence of a difference in population mean scores. **Inappropriate:** There is a 98% probability that the new method is better. Educational importance requires considering magnitude, precision and design.

Part VI. z-Tests for Continuous Variables

14. One-sample z-test

Recap: z-scores, SE, hypotheses and alpha

A z-score measures distance of an individual observation from a population mean in SD units. A one-sample **z-test** measures distance of a sample mean from a hypothesised mean in SE units:

Feature	Individual z-score	One-sample z-test statistic
What is compared?	An individual observation with the population mean	A sample mean with a hypothesised population mean
Denominator	Population SD (σ)	Standard error ($\sigma/\sqrt{n}$)
Meaning	Number of SDs above or below the mean	Number of SEs above or below the hypothesised mean
Main purpose here	Describe the position of an observation	Test a hypothesis about a population mean

For the one-sample z-test, the formula is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- $\bar{x}$: observed sample mean.
- μ_0: hypothesised population mean under H_0.
- σ: known population SD.
- n: sample size.
- $\sigma/\sqrt{n}$: SE under the model.

For a two-sided test at $\alpha = 0.05$, reject H_0 if $z < -1.96$ or $z > 1.96$; equivalently, if the two-sided $p \leq 0.05$.

Solved example 7: Biscuit packet weights

Claimed mean = 100 g; known population SD = 8 g; random sample $n = 100$, mean = 98 g.

Step 1. Select test: One sample mean versus reference mean; population SD known; independent observations assumed.

Step 2. Hypotheses: $H_0 : \mu = 100$; $H_1 : \mu \neq 100$.

Step 3. Choose alpha: $\alpha = 0.05$.

Step 4. SE:

$$\mathrm{SE} = \frac{8}{\sqrt{100}} = 0.8 \text{ g}$$

Step 5. Test statistic:

$$\mathrm{z} = \frac{98 - 100}{0.8} = -2.5$$

Step 6. Read the z-table and calculate p. The standard normal table gives the area to the **left** of a positive z-value. Since the calculated z is negative, first look up its absolute value, $|\mathrm{z}| = 2.50$.

Local z-table extract (cumulative area to the left):

z row	.00	.01	.02
2.4	0.9918	0.9920	0.9922
2.5	0.9938	0.9940	0.9941

Use **row 2.5, column .00**: $\mathrm{P}\left(\mathrm{Z} \leq 2.50\right) = 0.9938$.

Because the observed z is -2.50, the area to its left is $1 - 0.9938 = 0.0062$. For a two-sided test, include both tails:

$$\mathrm{p} = 2(0.0062) = 0.0124$$

The complete table is in Annexure 1.

Step 7. Decision: $0.0124 < 0.05$, so reject H_0.

Step 8. Interpretation: There is statistically significant evidence that the population mean packet weight differs from 100 g. The observed mean is 2 g lower. The result does not establish why.

Practice 6: Salt consumption

Reference population mean daily salt consumption = 10 g; known population SD = 4 g. A random sample of 64 adults has mean 11 g. Test for a two-sided difference at $\alpha = 0.05$.

Solution:

$$\mathrm{H}_0 : \mu = 10, \qquad \mathrm{H}_1 : \mu \models 10$$

$$\mathrm{SE} = \frac{4}{\sqrt{64}} = 0.5, \qquad \mathrm{z} = \frac{11 - 10}{0.5} = 2$$

Local z-table extract:

z row	.00	.01
2.0	0.9772	0.9778

At $\mathrm{z} = 2.00$, the cumulative probability is 0.9772. The upper-tail probability is $1 - 0.9772 = 0.0228$. Therefore, two-sided $\mathrm{p} = 2(0.0228) = 0.0456$ (approximately 0.0455 using unrounded values).

Since $\mathrm{p} < 0.05$, reject H_0. There is statistically significant evidence that population mean salt consumption differs from 10 g; the observed sample mean is 1 g higher.

15. Independent two-sample z-test

Recap: Two independent groups

We may compare mean study time between hostel students and day scholars. If groups comprise different independently sampled individuals and both population SDs are known, an independent two-sample z-test may be appropriate.

$$\mathrm{z} = \frac{\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2}{\sqrt{\sigma_1^2/\mathrm{n}_1 + \sigma_2^2/\mathrm{n}_2}}$$

- $\bar{\mathrm{x}}_1, \bar{\mathrm{x}}_2$: sample means in groups 1 and 2.
- σ_1, σ_2: known population SDs.
- $\mathrm{n}_1, \mathrm{n}_2$: group sample sizes.
- Denominator: SE of the difference in sample means.

Hypotheses: $\mathrm{H}_0 : \mu_1 = \mu_2$ and $\mathrm{H}_1 : \mu_1 \neq \mu_2$.

Solved example 8: Study time

	Hostel students	Day scholars
n	100	100
Mean (hours/day)	4.0	3.5
Known population SD	1.5	1.5

Step 1. Hypotheses: $\mathrm{H}_0 : \mu_1 = \mu_2; \mathrm{H}_1 : \mu_1 \models \mu_2$.

Step 2. SE of difference:

$$\mathrm{SE}_{\text{diff}} = \sqrt{\frac{1.5^2}{100} + \frac{1.5^2}{100}} = \sqrt{0.045} \approx 0.2121$$

Step 3. z:

$$\mathrm{z} = \frac{4 - 3.5}{0.2121} \approx 2.36$$

Step 4. Read the z-table and calculate p. Look up $\mathrm{z} = 2.36$ using row **2.3** and column .06.

Local z-table extract (cumulative area to the left):

z row	.05	.06	.07
2.3	0.9906	0.9909	0.9911

The area above 2.36 is $1 - 0.9909 = 0.0091$. Since the test is two-sided:

$$\mathrm{p} = 2(0.0091) = 0.0182 \approx 0.018$$

Step 5. Decision: Reject H_0 at 5%.

Step 6. Interpretation: There is statistically significant evidence of a difference in population mean study time. The observed mean is 0.5 hours higher among hostel students; this does not establish causation.

Practice 7: Screen time

	Group A	Group B
n	100	100
Mean screen time (hours/day)	5.0	4.6
Known population SD	2.0	2.0

Calculate z, p and interpretation.

Solution:

$$\mathrm{SE}_{\text{diff}} = \sqrt{\frac{4}{100} + \frac{4}{100}} = \sqrt{0.08} \approx 0.2828$$

$$z = \frac{5 - 4.6}{0.2828} \approx 1.41$$

Local z-table extract:

z row	.00	.01	.02
1.4	0.9192	0.9207	0.9222

For $z \approx 1.41$, use row **1.4**, column **.01**. The area to the left is 0.9207; the upper-tail area is $1 - 0.9207 = 0.0793$. Thus two-sided $p \approx 2(0.0793) = 0.1586$. Using the unrounded statistic ($z \approx 1.4142$) gives $p \approx 0.1573$.

Since $p > 0.05$, do not reject H_0. There is insufficient evidence of a population mean difference at 5%. This does not prove equality.

Limitation: Exact z-tests for means require known population SDs and suitable assumptions. In real studies SDs are usually estimated, and t-tests are often preferable. A large-sample z approximation using sample SDs is possible but must be identified as an approximation.

Part VII. Chi-Square Test for Categorical Variables

16. Introduction to chi-square

Recap: Categories and frequencies

Suppose we examine whether regular breakfast eating is associated with difficulty concentrating. Both variables are categorical, so we compare counts rather than means.

The **Pearson chi-square test of independence** evaluates association between two categorical variables by comparing observed and expected frequencies under H_0 of independence.

Expected count for a cell:

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

Chi-square statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O: observed cell count.
- E: expected cell count under independence.
- $\sum$: add contributions from all cells.
- χ^2: nonnegative statistic measuring discrepancy between observed and expected counts.

Degrees of freedom for an $r \times c$ contingency table:

$$df = (r - 1)(c - 1)$$

Here r is number of rows and c is number of columns. For a 2×2 table, $df = 1$. At $\alpha = 0.05$, the chi-square critical value for $df = 1$ is approximately 3.841.

17. Solved example 9: Breakfast and concentration

	Difficulty concentrating	No difficulty	Total
Regular breakfast	20	80	100
Irregular breakfast	40	60	100
Total	60	140	200

Step 1. Select test: Two categorical variables, independent observations assumed, adequate expected counts; chi-square test of independence.

Step 2. Hypotheses: H_0: breakfast regularity and difficulty concentrating are independent. H_1: they are associated.

Step 3. Expected counts:

$$E_{\text{regular,difficulty}} = \frac{100 \times 60}{200} = 30$$

$$E_{\text{regular,no difficulty}} = \frac{100 \times 140}{200} = 70$$

The same expected counts apply to the second row.

	Difficulty	No difficulty
Regular	30	70
Irregular	30	70

Step 4. Chi-square:

$$\chi^2 = \frac{(20-30)^2}{30} + \frac{(80-70)^2}{70} + \frac{(40-30)^2}{30} + \frac{(60-70)^2}{70}$$

$$= 3.333 + 1.429 + 3.333 + 1.429 \approx 9.52$$

Step 5. Degrees of freedom: $\mathrm{df} = (2-1)(2-1) = 1$.

Step 6. Use the chi-square table. With $\mathrm{df} = 1$, compare the calculated statistic with the upper-tail critical values.

Local chi-square table extract:

df	alpha=.05	alpha=.01	alpha=.001
1	3.841	6.635	10.828
2	5.991	9.210	13.816

Since $9.52 > 3.841$, reject H_0 at the 5% level. In fact, $9.52 > 6.635$ but $9.52 < 10.828$, so $0.001 < \mathrm{p} < 0.01$. A calculator or more detailed table gives $\mathrm{p} \approx 0.002$. The critical-value table alone does not give that exact p-value.

The complete critical-value table is in Annexure 2.

Step 7. Interpretation: There is statistically significant evidence of association between breakfast regularity and difficulty concentrating. Difficulty was reported by 20% of regular breakfast eaters and 40% of irregular breakfast eaters. This association does not establish causation.

18. Practice 8: Exercise and sleep quality

	Good sleep	Poor sleep	Total
Regular exercise	70	30	100
No regular exercise	50	50	100
Total	120	80	200

Calculate expected counts, chi-square, df, significance and interpretation.

Solution: Expected counts in each row: 60 good sleep, 40 poor sleep.

$$\chi^2 = \frac{(70-60)^2}{60} + \frac{(30-40)^2}{40} + \frac{(50-60)^2}{60} + \frac{(50-40)^2}{40}$$

$$= 1.667 + 2.5 + 1.667 + 2.5 \approx 8.33$$

Local chi-square table extract:

df	alpha=.05	alpha=.01	alpha=.001
1	3.841	6.635	10.828

Here $\mathrm{df} = 1$ and $\chi^2 \approx 8.33$. Because $8.33 > 6.635$ but $8.33 < 10.828$, the table shows $0.001 < \mathrm{p} < 0.01$. A more detailed calculation gives $\mathrm{p} \approx 0.0039$. Reject H_0 at 5%. There is statistically significant evidence of association. Good sleep was reported by 70% of regular exercisers and 50% of non-exercisers. Causation is not established.

18.1 When is ordinary chi-square inappropriate?

The Pearson chi-square approximation may be unreliable with very small expected counts; paired observations also require another method. Enter **counts**, not percentages. For simple introductory 2×2 exercises, a conservative rule is to require all expected counts to be at least 5. Fisher's exact test may be appropriate with small expected counts.

Part VIII. Selecting the Appropriate Statistical Method

19. Quick selection guide

Recap: Identify the outcome and question

Ask: Is the outcome numerical or categorical? Are we estimating or testing? One group or two? Independent or paired? Are assumptions satisfied?

Research question	Method taught here
Estimate population mean screen time	Confidence interval for a mean
Compare one sample mean with a reference mean, known population SD	One-sample z-test

Compare two independent population means, known population SDs	Two-sample z-test
Examine association between two categorical variables	Chi-square test of independence if assumptions hold
Compare before/after measurements in same people	Not the independent two-sample z-test
Compare single ordinal ratings between two groups	Usually another method
Categorical association with very small expected counts	Consider an exact test

20. Integrated practice

For each situation, identify the most appropriate statistical method and explain why.

No.	Situation
1	A researcher wants to estimate the mean daily water consumption of hostel students using a random sample.
2	A biscuit manufacturer claims a mean packet weight of 100 g. A sample is tested against this claim; the population SD is known.
3	A study compares mean examination scores between two independent groups; population SDs are known.
4	A study examines whether smoking status is associated with respiratory symptoms.
5	A study compares blood pressure before and after treatment in the same patients.
6	A study examines whether blood group is associated with a categorical disease outcome, with adequate expected frequencies.

Answers

No.	Appropriate approach	Reason

1	Confidence interval	Estimating a population mean
2	One-sample z-test	Comparing one mean with a reference value
3	Two-sample z-test	Comparing two independent means
4	Chi-square test	Examining association between categorical variables
5	Not an independent two-sample z-test	The observations are paired
6	Chi-square test	Examining association between categorical variables

---

Part IX. Self- Assessment

Suggested duration: 30 minutes
Total: 20 marks

Question 1. Confidence intervals (5 marks)

A random sample of 100 students has a mean daily mobile phone use of 5 hours and an SD of 2 hours.

Calculate the standard error (SE) and confidence intervals using 1, 2 and 3 SE. Interpret the interval using 2 SE.

Question 2. One-sample z-test (5 marks)

The reference mean weight of a packaged food product is 200 g, with a known population SD of 20 g.

A random sample of 100 packets has a mean weight of 195 g.

Test whether the population mean differs from 200 g at $\alpha = 0.05$. State the hypotheses, calculate z, determine the two-sided p-value using the table below, and interpret.

Extract from the standard normal table

The table gives the cumulative probability $\mathrm{P}\left(\mathrm{Z} \leq \mathrm{z}\right)$.

z	.00	.01	.02	.03	.04	.05
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960

For a two-sided test, use:

$$\mathrm{p} = 2[1 - \mathrm{P}\left(\mathrm{Z} \leq |\mathrm{z}|\right)]$$

Question 3. Chi-square test (6 marks)

A study investigates whether regular physical activity is associated with reported stress.

Physical activity	Stress present	Stress absent
Regular physical activity	30	70
No regular physical activity	50	50

Calculate the expected frequencies, chi-square statistic and degrees of freedom. Determine significance at the 5% level and interpret.

Extract from the chi-square critical-value table

The values represent upper-tail critical values.

Degrees of freedom	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.001$
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266

Question 4. Interpretation (4 marks)

State whether each statement is true or false and briefly explain.

No.	Statement
1	A p-value of 0.01 means there is a 1% probability that H_0 is true.

2	A 95% confidence interval contains 95% of individual observations.
3	A statistically nonsignificant result proves that H_0 is true.
4	A statistically significant association does not necessarily imply causation.

Part X. Self- Assessment Answer Key

Answer 1. Confidence intervals (5 marks)

Given: sample mean = 5 hours; SD = 2 hours; $n = 100$.

Step 1. Calculate SE.

$$SE = \frac{s}{\sqrt{n}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2 \text{ hours}$$

Step 2. Calculate the intervals.

Interval	Calculation	Result
Mean ± 1 SE	5 ± 0.2	4.8–5.2 hours
Mean ± 2 SE	5 ± 0.4	4.6–5.4 hours
Mean ± 3 SE	5 ± 0.6	4.4–5.6 hours

Interpretation: The interval using 2 SE has approximately 95.45% confidence under the normal approximation. It estimates the population mean daily mobile phone use, not the spread of individual observations.

Answer 2. One-sample z-test (5 marks)

Step 1. State the hypotheses.

$H_0 : \mu = 200 \text{ g}; H_1 : \mu \models 200 \text{ g}.$

Step 2. Calculate SE.

$$SE = \frac{20}{\sqrt{100}} = 2 \text{ g}$$

Step 3. Calculate z.

$$z = \frac{195 - 200}{2} = -2.5$$

Step 4. Obtain the p-value from the z-table.

z	.00	.01	.02
2.4	0.9918	0.9920	0.9922
2.5	**0.9938**	0.9940	0.9941
2.6	0.9953	0.9955	0.9956

Since $z = -2.50$, use the absolute value 2.50 and the symmetry of the normal distribution. Row 2.5, column .00 gives $P(Z \leq 2.50) = 0.9938$.

Upper-tail probability: $1 - 0.9938 = 0.0062$. Include both tails for the two-sided test:

$$p = 2(0.0062) = 0.0124$$

Step 5. Decide and interpret. Since $p = 0.0124 < \alpha = 0.05$, reject H_0. There is statistically significant evidence that the population mean packet weight differs from 200 g. The observed sample mean is 5 g lower than the reference value.

Answer 3. Chi-square test (6 marks)

Step 1. State the hypotheses.

H_0: Physical activity and reported stress are independent in the population.
H_1: Physical activity and reported stress are associated.

Step 2. Calculate row and column totals.

Physical activity	Stress present	Stress absent	Row total
Regular physical activity	30	70	100
No regular physical activity	50	50	100
Column total	80	120	200

Step 3. Calculate expected frequencies.

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

For stress present: $\mathrm{E} = (100 \times 80)/200 = 40$. For stress absent: $\mathrm{E} = (100 \times 120)/200 = 60$.

Physical activity	Stress present (expected)	Stress absent (expected)
Regular physical activity	40	60
No regular physical activity	40	60

Step 4. Calculate chi-square.

$$\chi^2 = \sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

$$\chi^2 = \frac{(30-40)^2}{40} + \frac{(70-60)^2}{60} + \frac{(50-40)^2}{40} + \frac{(50-60)^2}{60}$$

$$= 2.5 + 1.667 + 2.5 + 1.667 \approx 8.33$$

Step 5. Calculate degrees of freedom.

$$\mathrm{df} = (\mathrm{r}-1)(\mathrm{c}-1) = (2-1)(2-1) = 1$$

Step 6. Use the chi-square table.

df	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
1	3.841	6.635	10.828

The calculated statistic is 8.33. It exceeds 3.841 and 6.635, but not 10.828. Therefore, $0.001 < \mathrm{p} < 0.01$. A more precise calculation gives $\mathrm{p} \approx 0.0039$; the critical-value table alone does not give that exact value.

Step 7. Decide and interpret. Reject H_0 at the 5% significance level. There is statistically significant evidence of an association between regular physical activity and reported stress. Stress was reported by 30% of students engaging in regular physical activity and 50% of those not engaging in regular physical activity. This does not establish that physical activity causes a reduction in stress.

Answer 4. Interpretation (4 marks)

No.	Answer	Explanation

1	False	A p-value is calculated assuming H_0 and the statistical model are true. It does not indicate the probability that H_0 is true.
2	False	A 95% confidence interval for a population mean describes uncertainty in the estimated mean. It does not mean that 95% of individual observations lie within the interval.
3	False	A statistically nonsignificant result indicates insufficient evidence to reject H_0 at the chosen significance level. It does not prove that H_0 is true.
4	True	Statistical association does not necessarily establish causation. Other explanations, including confounding and bias, must be considered.

Annexure 1. Standard Normal Distribution Table

This table gives cumulative probability $\Phi(\mathrm{z}) = \mathrm{P}\left(\mathrm{Z} \leq \mathrm{z}\right)$ for a standard normal variable Z. For negative z-values use $\Phi(-\mathrm{z}) = 1 - \Phi(\mathrm{z})$. For a two-sided z-test, $\mathrm{p} = 2[1 - \Phi(|\mathrm{z}|)]$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389

1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Example: For $z = 2.40$, $\Phi(2.40) = 0.9918$. Upper tail $= 1 - 0.9918 = 0.0082$; two-sided $p = 2(0.0082) = 0.0164$.

Annexure 2. Chi-Square Critical Values

Values are **upper-tail** critical values: reject H_0 when the calculated χ^2 exceeds the relevant value.

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.001$
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.124
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588

Example: If $\chi^2 = 8.33$, $df = 1$, $\alpha = 0.05$, the critical value is 3.841. Since $8.33 > 3.841$, reject H_0.

Annexure 3. Essential Formula Sheet

Purpose	Formula
Sample mean	$\bar{x} = \sum x_i / n$
Sample SD	$s = \sqrt{\sum (x_i - \bar{x})^2 / (n-1)}$
Individual z-score	$z = (x - \mu)/\sigma$
Estimated SE of mean	$\widehat{SE} = s/\sqrt{n}$
Normal-based CI for mean	$\bar{x} \pm z^* SE$

One-sample z-test	$z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$
Independent two-sample z-test	$z = (\bar{x}_1 - \bar{x}_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
Expected frequency	$E = (\text{row total} \times \text{column total})/N$
Chi-square statistic	$\chi^2 = \sum(O - E)^2/E$
Chi-square df	$(r - 1)(c - 1)$

Annexure 4. Glossary of Important Terms

Term	Definition and correct interpretation
Population	Entire group about which conclusions are sought.
Sample	Subset of a population selected for study.
Parameter	Numerical characteristic of a population, such as its true mean.
Statistic	Numerical quantity calculated from sample observations.
Variable	Characteristic that takes different values among observations.
Nominal variable	Categorical variable without inherent order.
Ordinal variable	Categorical variable with ordered categories.
Interval scale	Numerical scale with equal intervals but no meaningful absolute zero.
Ratio scale	Numerical scale with equal intervals and meaningful zero.
Mean	Arithmetic average of observations.
Standard deviation (SD)	Measure of dispersion of observations around their mean.
Normal distribution	Symmetrical probability distribution characterised by mean and SD.
z-score	Number of population SDs an observation is above or below a reference population mean.

Sampling distribution	Distribution of a statistic over repeated samples under a specified sampling process.
Sampling error	Difference between a sample statistic and population parameter due to random sampling.
Standard error (SE)	SD of the sampling distribution of a statistic; not individual variability.
Point estimate	Single numerical estimate of a population parameter.
Confidence interval (CI)	Interval calculated using a procedure that contains the true parameter in a specified proportion of repeated samples under its assumptions.
Confidence level	Long-run coverage probability of the CI procedure, not a probability assigned to the fixed parameter after observing one interval.
Margin of error	Quantity added to/subtracted from a point estimate; for normal-based CIs, z^*SE.
Null hypothesis (H_0)	Statistical hypothesis being tested, often no difference or no association.
Alternative hypothesis (H_1)	Competing hypothesis describing departure from H_0.
Significance level (α)	Prespecified Type I error probability threshold for a correctly calibrated test.
Type I error	Rejecting H_0 when it is true.
Type II error	Failing to reject H_0 when it is false.
Statistical power	Probability of rejecting H_0 for a specified true alternative; $1-\beta$.
Test statistic	Quantity calculated from data and compared with a reference distribution under H_0.
p-value	Probability, assuming H_0 and the model, of a test statistic at least as extreme as observed; not the probability H_0 is true.

Statistical significance	Meeting a prespecified p-value threshold; does not necessarily imply clinical importance.
Critical value	Boundary defining a rejection region under the reference distribution.
One-sided test	Alternative specifies a direction of departure.
Two-sided test	Alternative allows departure in either direction.
z-test	Test using standard normal reference distribution; the versions taught here test means with known population SDs.
Chi-square test of independence	Test for association between categorical variables using observed and expected counts.
Observed frequency	Actual count recorded in a category or cell.
Expected frequency	Count expected under H_0 for a category or cell.
Degrees of freedom (df)	Number of independent quantities free to vary after constraints.
Association	Statistical relationship between variables; does not necessarily imply causation.
Confounding	Distortion of an observed exposure–outcome relationship due to other related factors.
Clinical significance	Practical importance for health or care, distinct from statistical significance.

Final Quick-Review Checklist

Before answering any biostatistics question, ask:

No.	Question
1	What is the population and what is the sample?
2	What variable is being studied, and what is its scale of measurement?
3	Am I estimating a population characteristic or testing a hypothesis?

4	Which statistical method is appropriate, and are its assumptions satisfied?
5	What are the relevant formula and its components?
6	Have I substituted the values and calculated correctly?
7	If testing a hypothesis, what are H_0, H_1 and α?
8	What does the p-value or confidence interval actually mean?
9	What conclusion is supported by the data?
10	Am I avoiding unsupported claims about equality, probability, clinical importance or causation?

The central objective of this module is not merely to perform calculations. It is to learn how to select appropriate statistical methods, use them correctly and interpret their results in a scientifically defensible manner.

A correct calculation with an incorrect interpretation is not a correct statistical conclusion.