STA 201: BIOSTATISTICS

Comprehensive Lecture Notes with Worked Examples, Tables, Charts, Graphs and Class Work

200 Level • First Semester • 2026/2027

Prepared from the STA 201 Course Outline and Work Schedule
College of Pure and Applied Sciences (COPAS)
Mathematics and Physics

Academic Session 2026/2027

Preface

These notes follow the twelve content areas listed in the STA 201 Biostatistics course outline and the fifteen-week work schedule. They are written for students in Medical Laboratory Science and related biomedical programmes. Each chapter combines explanations, formulas, worked examples, tables or graphs where useful, and class-work questions. Numerical datasets are hypothetical and are intended for learning; they are not clinical reference limits or real patient records.

Course Information

Course code	STA 201
Course title	Biostatistics
Level / semester	200 Level / First Semester
Credit units	2 (30 lecture hours; no practical hours stated in the outline)
Department	Medical Laboratory Science (MLS)
Academic session	2026/2027
Lecturers named in outline	Prof. Sunday Adewale and Dr. John Fatokun

Course Outcomes

By the end of the course, students should be able to:

1. Explain the role of biostatistics in medical laboratory science, clinical research and preventive medicine.
2. Distinguish populations, samples, parameters, statistics and biomedical variable types.
3. Construct frequency distributions and choose suitable tables, charts and graphs.
4. Calculate and interpret selected vital statistics.
5. Calculate and interpret mean, median and mode.
6. Calculate range, variance, standard deviation and coefficient of variation.
7. Apply elementary probability and probability distributions to biomedical situations.
8. Formulate hypotheses, perform introductory tests and interpret p-values.
9. Explain and apply one-way analysis of variance (ANOVA).
10. Calculate and interpret correlation and simple linear regression.
11. Explain randomisation, replication, controls and treatment allocation in experimental design.

Contents

1 Introduction to Biostatistics **1**

1.1 Meaning and scope . . . 1
1.2 Why biostatistics matters . . . 1
1.3 Core terms . . . 1
1.4 Four worked examples . . . 1

2 Biomedical Data: Populations, Samples and Variables **3**

2.1 Population and sample . . . 3
2.2 Types of variables . . . 3
2.3 Sampling and bias . . . 4
2.4 Four worked examples . . . 4

3 Frequency Distributions and Data Presentation **5**

3.1 Organising data . . . 5
3.2 Worked example: haemoglobin readings . . . 5
3.3 Charts and graphs . . . 5
3.3.1 Graph 1: Bar chart of haemoglobin values . . . 6
3.3.2 Graph 2: Pie chart . . . 6
3.3.3 Graph 3: Line graph . . . 6
3.3.4 Graph 4: Cumulative frequency curve (ogive) . . . 6
3.4 Four worked examples . . . 7

4 Vital Statistics and Health Indicators **8**

4.1 Meaning . . . 8
4.2 Common formulas . . . 8
4.3 Worked examples . . . 8

5 Descriptive Statistics and Central Tendency **10**

5.1 Arithmetic mean . . . 10
5.2 Median and mode . . . 10
5.3 Worked examples . . . 10
5.4 Comparison table . . . 11
5.5 Tabular worked examples: mean, median and mode . . . 11
5.5.1 Arithmetic mean for ungrouped data . . . 11

5.5.2 Mean from a frequency table . . . 12
5.5.3 Median in tabular form: ungrouped data . . . 12
5.5.4 Mode in tabular form . . . 13
5.5.5 Grouped-data mean using class midpoints . . . 13
5.5.6 Grouped-data median . . . 13
5.5.7 Grouped-data mode . . . 14
5.5.8 Worked frequency distribution from the handout . . . 14
5.6 Extended worked example from the course handout: household feeding expenditure 16
5.6.1 Step 1: Construct seven classes . . . 16
5.6.2 Step 2: Arithmetic mean . . . 16
5.6.3 Step 3: Harmonic mean using the raw observations . . . 17
5.6.4 Step 4: Geometric mean using the raw observations . . . 18
5.7 Additional grouped-data practice: central tendency . . . 19
5.8 Choosing and reporting a measure of central tendency . . . 20

6 Measures of Dispersion 21
6.1 Range . . . 21
6.2 Variance and standard deviation . . . 21
6.3 Coefficient of variation . . . 21
6.4 Worked examples . . . 21

7 Elements of Probability 23
7.1 Basic ideas . . . 23
7.2 Worked examples . . . 23

8 Probability Distributions 25
8.1 Discrete distributions . . . 25
8.1.1 Binomial distribution . . . 25
8.1.2 Poisson distribution . . . 25
8.1.3 Normal distribution . . . 25
8.2 Worked examples . . . 25

9 Hypothesis Testing and Statistical Significance 27
9.1 The logic of a test . . . 27
9.2 One-sample t test . . . 27
9.3 Worked examples . . . 27
9.4 Basic decision table . . . 28

10 Analysis of Variance (ANOVA) 29
10.1 Purpose . . . 29
10.2 ANOVA calculations . . . 29
10.3 Worked example: three laboratory methods . . . 29
10.4 Four worked examples . . . 30

11 Correlation Analysis **31**
11.1 Meaning . . . 31
11.2 Scatter plot . . . 31
11.3 Four worked examples . . . 31

12 Simple Linear Regression **33**
12.1 Purpose and equation . . . 33
12.2 Four worked examples . . . 33

13 Experimental Design in Clinical and Laboratory Research **35**
13.1 Key principles . . . 35
13.2 Common designs . . . 35
13.3 Four worked examples . . . 35

14 Integrated Biomedical Applications and Revision **37**
14.1 Choosing an appropriate method . . . 37
14.2 Integrated worked example . . . 37
14.3 Revision questions . . . 38

Selected Answers to Class Work **39**

A Advanced Worked Examples by Subtopic **40**
A.1 Applications and interpretation of biostatistics . . . 40
A.2 Populations, samples, parameters and sampling . . . 43
A.3 Frequency distributions and grouped data . . . 45
A.4 Charts, pie charts, histograms, ogives and graphs . . . 47
A.5 Vital statistics and health indicators . . . 49
A.6 Mean and weighted/frequency mean . . . 51
A.7 Median and quartiles . . . 53
A.8 Mode and shape of distributions . . . 55
A.9 Range, variance and standard deviation . . . 57
A.10 Coefficient of variation and precision . . . 59
A.11 Probability rules and conditional probability . . . 61
A.12 Binomial distribution . . . 63
A.13 Poisson distribution . . . 65
A.14 Normal distribution and standard scores . . . 67
A.15 Hypothesis testing and one-sample t test . . . 69
A.16 ANOVA and comparison of means . . . 71
A.17 Correlation analysis . . . 73
A.18 Simple linear regression . . . 75
A.19 Experimental design and applications . . . 77

Suggested 15-Week Teaching Sequence **79**

Assessment Structure **80**

Recommended Textbooks **81**

B Supplementary Rigorous Worked Examples from the Course Handout **82**

B.1 Extended tabular calculations for variance and standard deviation 82

B.2 Additional worked examples from the handout: frequency distributions and central tendency 84

B.3 Additional probability-distribution examples from the handout 87

B.4 University-standard integrated class work 88

Chapter 1

Introduction to Biostatistics

1.1 Meaning and scope

Biostatistics is the application of statistical ideas and methods to biology, medicine, laboratory science and public health. Statistics helps researchers collect, organise, summarise, analyse and interpret observations while recognising uncertainty.

1.2 Why biostatistics matters

- **Clinical medicine:** summarising outcomes and comparing treatment groups.
- **Medical laboratory science:** evaluating precision, agreement and performance of laboratory methods.
- **Preventive medicine:** monitoring disease occurrence and assessing prevention programmes.
- **Public health:** describing births, deaths, outbreaks and population health indicators.
- **Research:** planning sample sizes, testing hypotheses and estimating relationships.

1.3 Core terms

Data Recorded observations, such as a patient's haemoglobin measurement.

Descriptive statistics Methods that summarise the data collected.

Inferential statistics Methods that use sample information to draw conclusions about a population.

Uncertainty The fact that sample results vary and may not equal the true population value.

1.4 Four worked examples

Worked Example

Example 1: Summarising laboratory results. Ten hypothetical haemoglobin readings are recorded. Which branch of statistics is used to calculate their average? **Solution.** Calculating the average of the ten observed readings is descriptive statistics because it summarises the data already collected.

Worked Example

Example 2: Comparing two methods. A laboratory compares the average result from Method A with Method B using samples from patients. **Solution.** Descriptive statistics first summarise each method. Inferential statistics may then be used to estimate whether the observed difference is compatible with random sampling variation.

Worked Example

Example 3: Monitoring a prevention programme. A health team records monthly malaria cases before and after a programme. **Solution.** A table and line graph can describe the trend. A suitable inferential method may assess whether the change is larger than expected by chance, while considering other possible explanations.

Worked Example

Example 4: Estimating prevalence. Researchers test 200 randomly selected adults from a town and find 30 positive results. **Solution.** Sample proportion $= 30/200 = 0.15 = 15\%$. The 15% is a sample estimate of the town's prevalence, not automatically the exact population prevalence.

Class Work:

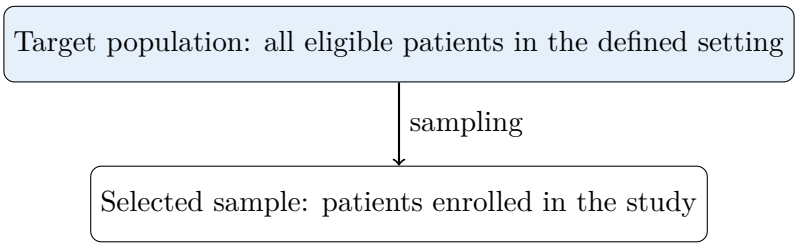
1. Give five applications of biostatistics in medical laboratory science.
2. Distinguish descriptive from inferential statistics using one biomedical example each.
3. Explain why a sample result may differ from the true population value.
4. State two ways statistics can improve public-health decisions.

Chapter 2

Biomedical Data: Populations, Samples and Variables

2.1 Population and sample

A **population** is the complete set of people, specimens or events about which a study asks a question. A **sample** is the subset actually observed. A **parameter** describes a population; a **statistic** describes a sample.



Measure the sample; use the result to estimate population characteristics

2.2 Types of variables

Type	Meaning	Biomedical examples
Nominal categorical	Categories without natural order	Blood group, organism species, test positive/negative
Ordinal categorical	Ordered categories	Pain rating mild/moderate/severe; disease stage
Discrete quantitative	Countable numbers	Number of admissions, colonies counted
Continuous quantitative	Measurement on a scale	Temperature, weight, glucose concentration

2.3 Sampling and bias

Simple random sampling gives eligible units a known chance of selection. Systematic sampling selects every kth unit after a random start. Stratified sampling selects within important subgroups. Convenience sampling uses readily available participants and may be biased. Selection bias occurs when the sample differs systematically from the target population.

2.4 Four worked examples

Worked Example

Example 1: Define the study population. A researcher studies glucose levels among all first-year students at a university but measures 120 students. **Solution.** Population: all first-year students at that university who meet the study criteria. Sample: the 120 measured students. Their mean glucose is a statistic; the true mean for all eligible first-year students is a parameter.

Worked Example

Example 2: Classify variables. Classify blood group, number of clinic visits, body temperature and pain category. **Solution.** Blood group: nominal; clinic visits: discrete quantitative; temperature: continuous quantitative; pain category: ordinal.

Worked Example

Example 3: Parameter or statistic? The mean platelet count from 80 sampled patients is 245 units. **Solution.** This is a statistic because it was calculated from a sample. If 245 were the mean for every eligible patient in the population, it would be a parameter.

Worked Example

Example 4: Identify sampling method. From a list of 1,000 records, a computer randomly selects 100. **Solution.** This is simple random sampling, assuming each record has the intended chance of selection and the list adequately represents the target population.

Class Work:

1. Classify: age, blood group, number of positive cultures, and disease severity (low/medium/high).
2. Give one population, one sample, one parameter and one statistic for a laboratory study.
3. Explain one advantage and one limitation of convenience sampling.
4. A clinic has 600 records and selects every 10th record after a random start. Name the sampling method.

Chapter 3

Frequency Distributions and Data Presentation

3.1 Organising data

A frequency distribution lists values or class intervals and the number of observations in each. Relative frequency is f/n; percentage frequency is $100f/n$. Cumulative frequency is the running total of frequencies.

3.2 Worked example: haemoglobin readings

Use the hypothetical readings (g/dL): $11, 12, 13, 12, 14, 15, 12, 13, 11, 14$.

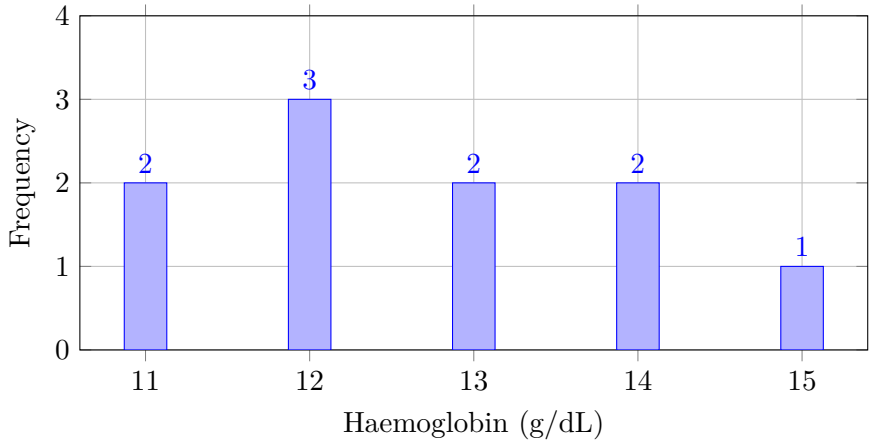
Value x	Frequency f	Relative frequency $f/10$	Cumulative frequency
11	2	0.20	2
12	3	0.30	5
13	2	0.20	7
14	2	0.20	9
15	1	0.10	10
Total	10	1.00	

The value 12 occurs three times, so its percentage frequency is $3/10 \times 100 = 30\%$.

3.3 Charts and graphs

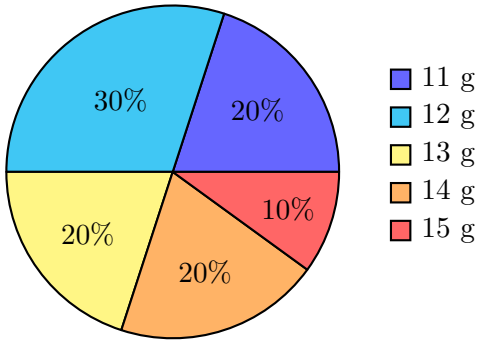
- **Bar chart:** compares categories; bars have gaps.
- **Histogram:** displays a continuous variable grouped into intervals; bars touch.
- **Pie chart:** shows parts of a whole. Sector angle $= (f/n) \times 360^\circ$.
- **Line graph:** displays change over time.
- **Frequency polygon:** joins class-midpoint frequency points.
- **Ogive:** plots cumulative frequency against upper class boundaries.
- **Scatter plot:** displays paired quantitative observations.

3.3.1 Graph 1: Bar chart of haemoglobin values

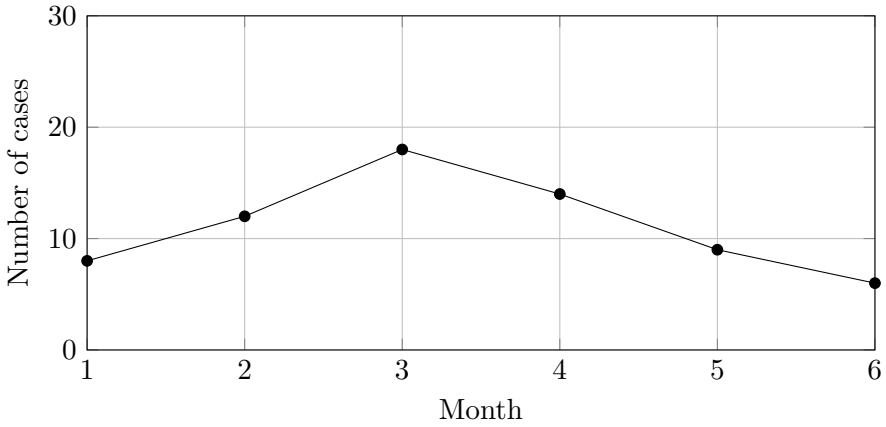


3.3.2 Graph 2: Pie chart

For the same data, the sector angle for value 12 is $(3/10) \times 360^\circ = 108^\circ$.

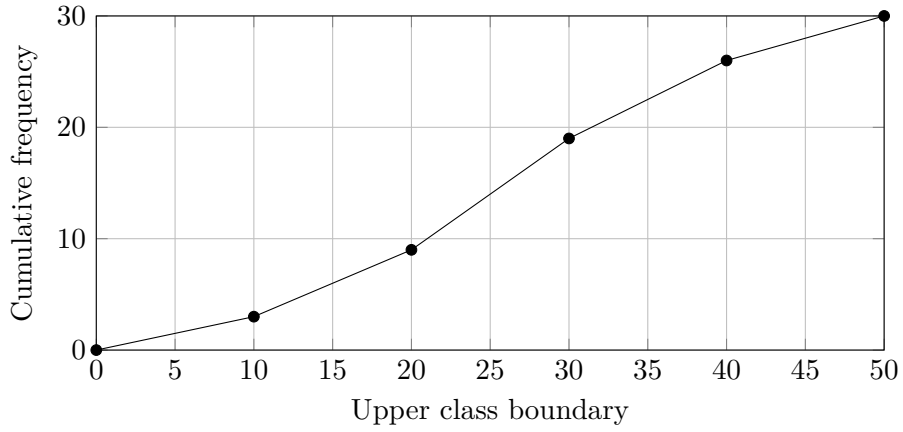


3.3.3 Graph 3: Line graph



3.3.4 Graph 4: Cumulative frequency curve (ogive)

Hypothetical grouped data are shown below. The ogive uses the upper class boundary and cumulative frequency.



3.4 Four worked examples

Worked Example

Example 1: Relative frequency. In 40 specimens, 8 are positive. **Solution.** Relative frequency $= 8/40 = 0.20$. Percentage $= 0.20 \times 100 = 20\%$.

Worked Example

Example 2: Pie-chart angle. A category contains 15 of 60 observations. **Solution.** Angle $= (15/60) \times 360^\circ = 90^\circ$.

Worked Example

Example 3: Class width and midpoint. For class 20–29, assuming integer measurements, find the class midpoint and real boundaries. **Solution.** Midpoint $= (20 + 29)/2 = 24.5$. Real boundaries are 19.5 and 29.5, giving width $29.5 - 19.5 = 10$.

Worked Example

Example 4: Cumulative frequency. Frequencies are 3, 6, 10 and 5. Find cumulative frequencies. **Solution.** Running totals are 3, $3 + 6 = 9$, $9 + 10 = 19$, and $19 + 5 = 24$. The total number of observations is 24.

Class Work:

1. Construct a frequency table for $4, 5, 4, 6, 7, 5, 4, 6, 5, 8$.
2. A test is positive in 18 of 90 samples. Find the percentage and pie-chart angle.
3. State two differences between a bar chart and a histogram.
4. For frequencies 2, 5, 7, 4, calculate cumulative frequencies and describe how to draw an ogive.

Chapter 4

Vital Statistics and Health Indicators

4.1 Meaning

Vital statistics concern events such as births and deaths in a population. Rates require a clearly defined numerator, denominator, period and multiplier. In practice, use the official definition required by the relevant health authority.

4.2 Common formulas

$$\text{Crude birth rate} = \frac{\text{live births during period}}{\text{mid-year population}} \times 1000$$

$$\text{Crude death rate} = \frac{\text{deaths during period}}{\text{mid-year population}} \times 1000$$

$$\text{Case-fatality proportion} = \frac{\text{deaths among identified cases}}{\text{identified cases}} \times 100\%$$

$$\text{Incidence proportion} = \frac{\text{new cases during period}}{\text{population at risk at start}} \times 100\%$$

$$\text{Prevalence} = \frac{\text{all existing cases at a point/period}}{\text{population assessed}} \times 100\%$$

4.3 Worked examples

Worked Example

Example 1: Crude birth rate. A town of 25,000 people records 500 live births in a year. **Solution.** $CBR = (500/25,000) \times 1000 = 20$ live births per 1,000 population per year.

Worked Example

Example 2: Crude death rate. A district of 80,000 people records 640 deaths. **Solution.** $CDR = (640/80,000) \times 1000 = 8$ deaths per 1,000 population for the period.

Worked Example

Example 3: Case-fatality proportion. Of 240 identified cases, 12 die from the disease. **Solution.** $(12/240) \times 100 = 5\%$. This describes the proportion of identified cases who died; interpretation depends on case definition and follow-up.

Worked Example

Example 4: Incidence proportion. At the start of follow-up, 1,200 people are disease-free and at risk. During follow-up, 36 develop the disease. **Solution.** $(36/1,200) \times 100 = 3\%$ over the stated follow-up period.

Worked Example

Example 5: Prevalence. A survey finds 75 people with a condition among 1,500 examined. **Solution.** Prevalence $= (75/1,500) \times 100 = 5\%$ at the survey time.

Class Work:

1. A population of 60,000 has 900 live births. Calculate the crude birth rate per 1,000.
2. There are 150 deaths in a population of 30,000. Calculate the crude death rate per 1,000.
3. Of 400 cases, 16 die. Calculate the case-fatality proportion.
4. Explain the difference between incidence and prevalence.

Chapter 5

Descriptive Statistics and Central Tendency

5.1 Arithmetic mean

For raw observations $x_1, \ldots, x_n$, the sample mean is

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

For a frequency table, $\bar{x} = \sum fx / \sum f$. For grouped data, use class midpoints as approximations.

5.2 Median and mode

The median is the middle ordered value. If n is even, it is the average of the two middle values. The mode is the most frequent value; a dataset can have no mode, one mode or more than one mode.

5.3 Worked examples

Worked Example

Example 1: Mean. Measurements are 4, 6, 7, 8, 10. **Solution.** Sum = 4+6+7+8+10 = 35, $n = 5$, so $\bar{x} = 35/5 = 7$.

Worked Example

Example 2: Median with odd n. Data are 3, 9, 4, 7, 5. **Solution.** Ordered data: 3, 4, 5, 7, 9. The middle value is 5, so median = 5.

Worked Example

Example 3: Median with even n. Data are 10, 4, 8, 6. **Solution.** Ordered data: 4, 6, 8, 10. Median = $(6 + 8)/2 = 7$.

Worked Example

Example 4: Mode. Results are 2, 3, 3, 4, 5, 3, 6. **Solution.** The value 3 occurs three times, more than any other value; mode = 3.

Worked Example

Example 5: Frequency mean. Values 1, 2, 3 have frequencies 2, 4, 2. **Solution.** $\sum fx = (1)(2) + (2)(4) + (3)(2) = 16$; $\sum f = 8$. Mean $= 16/8 = 2$.

5.4 Comparison table

Measure	Strength	Limitation
Mean	Uses all observations; useful for further calculations	Sensitive to extreme values
Median	Resistant to extreme values; useful for skewed data	Does not use distances of all values from the centre
Mode	Applies to categories and identifies most common value	May be absent or not unique

Class Work:

1. Calculate mean, median and mode for 6, 8, 8, 9, 10, 13.
2. Find the mean of values 2, 4, 6 with frequencies 3, 2, 5.
3. Explain why median may be preferred for strongly skewed income data.
4. Give an example of a categorical variable for which a mode is useful.

5.5 Tabular worked examples: mean, median and mode

The following tables show each stage of the calculation. In biomedical work, retain the units and state whether a grouped-data answer is exact or approximate.

5.5.1 Arithmetic mean for ungrouped data

Worked Example

Example 6: Ages of students (adapted from the course handout). The handout lists the ages $13, 18, 18, 19, 17, 19.5, 19, 17, 18.5, 18, 17$. The printed solution says there are 10 observations and gives 18.1; however, the displayed list contains 11 observations, so the calculation below uses all 11 listed values.

Observation	Age x (years)	Running total
1	13	13
2	18	31
3	18	49
4	19	68
5	17	85
6	19.5	104.5
7	19	123.5
8	17	140.5
9	18.5	159
10	18	177
11	17	194
Total	$\sum x = 194$	$n = 11$

Solution. $\bar{x} = \sum x/n = 194/11 = 17.636$ years (approximately 17.64 years). Always check that the number of observations used agrees with the raw list.

5.5.2 Mean from a frequency table

Worked Example

Example 7: Values and frequencies. Suppose the number of colonies counted on plates is recorded as follows.

Colonies, x	Frequency, f	fx
1	2	2
2	4	8
3	2	6
Total	$\sum f = 8$	$\sum fx = 16$

Solution. $\bar{x} = \sum fx / \sum f = 16/8 = 2$ colonies.

5.5.3 Median in tabular form: ungrouped data

Worked Example

Example 8: Odd number of observations. Find the median of measurements $3, 9, 4, 7, 5$.

Position after sorting	Measurement
1	3
2	4
3	5
4	7
5	9

Solution. The ordered list has $n = 5$, so the median position is $(n + 1)/2 = 3$. The median is 5.

Worked Example

Example 9: Even number of observations. Find the median of $10, 4, 8, 6$.

Position after sorting	Measurement
1	4
2	6
3	8
4	10

Solution. The two middle values are 6 and 8. Median $= (6 + 8)/2 = 7$.

5.5.4 Mode in tabular form

Worked Example

Example 10: Most frequent result. The observed results are $2, 3, 3, 4, 5, 3, 6$.

Result x	Frequency f
2	1
3	3
4	1
5	1
6	1

Solution. The highest frequency is 3, belonging to value 3. Therefore, mode = 3.

5.5.5 Grouped-data mean using class midpoints

Worked Example

Example 11: Weight distribution (from the course handout). Estimate the mean weight from the grouped frequency table. The class midpoint is $x = (\text{lower limit} + \text{upper limit})/2$.

Weight class	Frequency f	Midpoint x	fx
52–56	3	54	162
56–60	6	58	348
60–64	10	62	620
64–68	4	66	264
68–72	8	70	560
72–76	2	74	148
76–80	1	78	78
Total	34		2180

Solution. $\bar{x} \approx \sum fx / \sum f = 2180/34 = 64.12$ (weight units). This is an estimate because individual observations have been represented by class midpoints.

5.5.6 Grouped-data median

For continuous grouped data, an estimate of the median is

$$\text{Median} = L + \left(\frac{N/2 - CF}{f_m} \right) h,$$

where L is the lower class boundary of the median class, N is the total frequency, CF is the cumulative frequency before the median class, f_m is the frequency in the median class, and h is class width.

Worked Example

Example 12: Grouped clinical waiting times. Use this illustrative frequency table to estimate the median waiting time.

Waiting time (minutes)	Frequency f	Cumulative frequency
0–10	4	4
10–20	7	11
20–30	12	23
30–40	5	28
40–50	2	30
Total	30	

Solution. $N/2 = 15$, so the median class is 20–30. Using $L = 20$, $CF = 11$, $f_m = 12$, and $h = 10$:

$$\text{Median} = 20 + \left(\frac{15 - 11}{12}\right) 10 = 20 + 3.333 = 23.33 \text{ minutes (approximately).}$$

The result is an estimate based on the grouped table, not the exact median of the original waiting times.

5.5.7 Grouped-data mode

For grouped continuous data, an estimated mode is

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) h,$$

where L is the lower boundary of the modal class, f_1 is its frequency, f_0 is the frequency of the preceding class, f_2 is the frequency of the following class, and h is class width.

Worked Example

Example 13: Modal waiting-time class. Use the grouped waiting-time table above. The modal class is 20–30 minutes because its frequency, 12, is the largest. Thus $L = 20$, $f_1 = 12$, $f_0 = 7$, $f_2 = 5$, and $h = 10$. **Solution.**

$$\text{Mode} = 20 + \left(\frac{12 - 7}{2(12) - 7 - 5}\right) 10 = 20 + \frac{5}{12}(10) = 24.17 \text{ minutes (approximately).}$$

This is an estimated mode; the modal class itself (20–30 minutes) is often the safest direct description.

5.5.8 Worked frequency distribution from the handout

The course handout gives 55 measurements ranging from 4.5 to 52.0 and illustrates seven classes. The following table retains its stated class intervals and frequencies.

Measurement interval	Frequency
4.5–11.5	22
11.5–18.5	21
18.5–25.5	5
25.5–32.5	4
32.5–39.5	1
39.5–46.5	1
46.5–53.5	1
Total	55

Working: Range $= 52.0 - 4.5 = 47.5$. Sturges' rule gives $k = 1 + 3.322 \log_{10}(55) \approx 6.78$, or about 7 classes. Approximate class width $= 47.5/7 \approx 6.79$, conventionally rounded to about 7. When constructing such tables from raw data, use a consistent boundary convention so observations on a class boundary are counted exactly once.

5.6 Extended worked example from the course handout: household feeding expenditure

The handout gives feeding expenditure (in thousands of naira) for 40 households in Ibadan in March 2023 and asks for a seven-class frequency distribution and arithmetic, harmonic and geometric means. The raw observations are reproduced here for practice:

$$32, 22, 19, 18, 43, 42, 40, 43, 18, 21, 31, 26, 22, 25, 47, 40, 26, 32, 34, 22,$$

$$35, 38, 34, 41, 36, 25, 22, 45, 48, 26, 28, 38, 35, 19, 47, 28, 26, 35, 38, 35.$$

5.6.1 Step 1: Construct seven classes

Range $= 48 - 18 = 30$. For seven classes, a convenient class width is approximately $30/7 = 4.29$, so use intervals of width 5.

Expenditure (N'000)	Midpoint x	Frequency f	fx	Cumulative f
18–22	20	9	180	9
23–27	25	6	150	15
28–32	30	5	150	20
33–37	35	7	245	27
38–42	40	7	280	34
43–47	45	5	225	39
48–52	50	1	50	40
Total		40	1280	

Check: Frequencies sum to $9 + 6 + 5 + 7 + 7 + 5 + 1 = 40$, matching the 40 households.

5.6.2 Step 2: Arithmetic mean

For the grouped table,

$$\bar{x} \approx \frac{\sum fx}{\sum f} = \frac{1280}{40} = 32.$$

Thus the grouped-data estimate is N32,000 per household. The exact arithmetic mean from the 40 raw observations is

$$\bar{x} = \frac{1282}{40} = 32.05$$

thousand naira, or N32,050. The small difference occurs because grouping replaces each observation with its class midpoint.

5.6.3 Step 3: Harmonic mean using the raw observations

For positive observations, the harmonic mean is

$$H = \frac{n}{\sum(1/x_i)}.$$

For a frequency table, $H = \sum f / \sum(f/x)$. The raw-data calculation is shown by listing each distinct expenditure and its frequency.

Value x	Frequency f	f/x	$f \ln x$
18	2	0.11111	5.7807
19	2	0.10526	5.8889
21	1	0.04762	3.0445
22	4	0.18182	12.3636
25	2	0.08000	6.4378
26	4	0.15385	13.0334
28	2	0.07143	6.6644
31	1	0.03226	3.4340
32	2	0.06250	6.9315
34	2	0.05882	7.0551
35	4	0.11429	14.2228
36	1	0.02778	3.5835
38	3	0.07895	10.9310
40	2	0.05000	7.3778
41	1	0.02439	3.7136
42	1	0.02381	3.7377
43	2	0.04651	7.5229
45	1	0.02222	3.8067
47	2	0.04255	7.7017
48	1	0.02083	3.8712
Total	40	≈ 1.3560	≈ 137.0783

Working: Using the raw observations directly, $\sum(1/x_i) \approx 1.3560$, hence

$$H = \frac{40}{1.3560} \approx 29.50$$

thousand naira. Small differences in the last decimal may occur from rounding the reciprocal column. Use unrounded reciprocals for the final answer.

5.6.4 Step 4: Geometric mean using the raw observations

For positive observations, the geometric mean is

$$G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \exp\left(\frac{1}{n}\sum \ln x_i\right).$$

Using the distinct-value frequency table on the preceding page, the sum of the log values over all observations is approximately 137.0783 (use full-precision logarithms in the calculator). Therefore,

$$G = \exp(137.0783/40) \approx 30.78$$

thousand naira. The arithmetic mean (N32.05 thousand) is greater than the geometric mean (about N30.78 thousand), which is greater than the harmonic mean (about N29.50 thousand), as expected for positive, unequal observations:

$$\text{AM} \geq \text{GM} \geq \text{HM}.$$

These means answer slightly different questions. The arithmetic mean describes the average amount; the geometric mean is useful for multiplicative changes and ratios; the harmonic mean is useful for certain rate or ratio averages. Do not choose a mean solely because its numerical value is convenient.

Class Work:

1. Verify the grouped frequency table from the 40 raw values.
2. Recalculate the arithmetic mean directly from the raw data and compare it with the grouped estimate.
3. Use a calculator with full precision to recalculate the harmonic and geometric means.
4. Explain why the three means are not identical.

5.7 Additional grouped-data practice: central tendency

Worked Example

Example 14: Estimate all three measures. A laboratory records the turnaround time for 50 specimens as grouped below.

Time (minutes)	Frequency f	Midpoint x	Cumulative frequency
0–9	5	4.5	5
10–19	9	14.5	14
20–29	16	24.5	30
30–39	12	34.5	42
40–49	8	44.5	50
Total	50		

Mean:

$$\sum fx = 5(4.5) + 9(14.5) + 16(24.5) + 12(34.5) + 8(44.5) = 1305.$$

$$\bar{x} \approx 1305/50 = 26.1 \text{ minutes.}$$

Median: $N/2 = 25$, so the median class is 20–29. Using class boundary $L = 19.5$, $CF = 14$, $f_m = 16$, and $h = 10$:

$$\text{Median} = 19.5 + \left(\frac{25 - 14}{16}\right) 10 = 26.375 \text{ minutes.}$$

Mode: The modal class is 20–29, with $f_1 = 16$, preceding frequency $f_0 = 9$, following frequency $f_2 = 12$, lower boundary $L = 19.5$, and width $h = 10$:

$$\text{Mode} = 19.5 + \left(\frac{16 - 9}{2(16) - 9 - 12}\right) 10 = 19.5 + \frac{7}{11}(10) \approx 25.86 \text{ minutes.}$$

All three results are estimates because only grouped frequencies are available. The modal class can also be reported directly as 20–29 minutes.

5.8 Choosing and reporting a measure of central tendency

Measure	Calculation / interpretation	Suitable use and caution
Arithmetic mean	Sum all values and divide by the number of observations; for grouped data use $\sum fx / \sum f$.	Useful for approximately symmetric numerical data; sensitive to extreme values.
Median	Middle ordered value; for grouped data estimate with the median-class formula.	Useful for skewed numerical data and when extreme values are present.
Mode	Most frequent value or category; grouped data use the modal class or grouped-mode estimate.	Useful for nominal categories and identifying the most common result; may be non-unique.
Harmonic mean	$H = n/\sum(1/x)$ for positive values.	Useful for particular rate averages; strongly affected by small positive values and cannot be applied to zero values.
Geometric mean	$G = \exp(\sum \ln x/n)$ for positive values.	Useful for multiplicative changes and positively skewed ratio data; requires positive values.

How to present a complete statistical solution

For every worked problem, show the following in order:

1. State the measure and write the correct formula.
2. Identify n or $\sum f$, and show any class midpoint or cumulative frequency needed.
3. Substitute the values before simplifying.
4. Give the numerical result with suitable rounding and the correct unit.
5. Interpret the result in the context of the question, noting when a grouped-data estimate is being used.

Chapter 6

Measures of Dispersion

6.1 Range

$$\text{Range} = x_{\max} - x_{\min}.$$

Range is simple but depends only on the two extreme values.

6.2 Variance and standard deviation

For a population:

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

For a sample:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}, \qquad s = \sqrt{s^2}.$$

Use the population formula when the dataset is treated as the entire population of interest; use the sample formula when estimating variability from a sample.

6.3 Coefficient of variation

For a positive mean, $\text{CV} = (s/\bar{x}) \times 100\%$. It compares relative variability between measurements on suitable ratio scales. Do not use it uncritically when the mean is near zero or the scale has no meaningful zero.

6.4 Worked examples

Worked Example

Example 1: Range. Values are 9, 12, 15, 16, 20. **Solution.** Range $= 20 - 9 = 11$.

Worked Example

Example 2: Population variance and SD. Data are 2, 4, 6. **Solution.** Mean $= 4$. Squared deviations are $4, 0, 4$, sum $= 8$. Population variance $= 8/3 = 2.667$; population SD $= \sqrt{2.667} \approx 1.633$.

Worked Example

Example 3: Sample variance and SD. Use the same values 2, 4, 6 as a sample. **Solution.** $\bar{x} = 4$ and sum of squared deviations $= 8$. Sample variance $= 8/(3-1) = 4$; sample SD $= \sqrt{4} = 2$.

Worked Example

Example 4: Coefficient of variation. A method has mean 50 and sample SD 5. **Solution.** $CV = (5/50) \times 100 = 10\%$.

Worked Example

Example 5: Compare precision. Method A: mean 100, SD 4. Method B: mean 20, SD 2. **Solution.** $CV_A = 4/100 \times 100 = 4\%$; $CV_B = 2/20 \times 100 = 10\%$. Method A has lower relative variation in this hypothetical comparison.

Class Work:

1. Calculate range for 11, 13, 13, 18, 22.
2. Calculate population variance and SD for 1, 3, 5.
3. Calculate sample variance and SD for 5, 7, 8, 10.
4. Method X has mean 40 and SD 2; Method Y has mean 80 and SD 8. Compare their CVs.

Chapter 7

Elements of Probability

7.1 Basic ideas

A probability is a number from 0 to 1. An event with probability 0 is impossible under the model; probability 1 means certain. The complement rule is $P(A^c) = 1 - P(A)$. For mutually exclusive events, $P(A \cup B) = P(A) + P(B)$. In general,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

For independent events, $P(A \cap B) = P(A)P(B)$. Conditional probability is $P(A \mid B) = P(A \cap B)/P(B)$ when $P(B) > 0$.

7.2 Worked examples

Worked Example

Example 1: Complement. A test result is positive with probability 0.08. Find the probability of not positive. **Solution.** $P(\text{not positive}) = 1 - 0.08 = 0.92$.

Worked Example

Example 2: Mutually exclusive outcomes. In a simplified model, probabilities of outcome A and outcome B are 0.20 and 0.35, and they cannot occur together. **Solution.** $P(A \cup B) = 0.20 + 0.35 = 0.55$.

Worked Example

Example 3: Independent events. Suppose two independent specimens each have probability 0.1 of contamination. Find the probability both are contaminated. **Solution.** $0.1 \times 0.1 = 0.01$.

Worked Example

Example 4: Conditional probability. Of 100 specimens, 40 are from group A; 10 of group A are positive. Find the proportion positive among group A. **Solution.** $P(\text{positive} \mid A) = 10/40 = 0.25$.

Worked Example

Example 5: At least one. Two independent tests each have probability 0.9 of detecting a condition. Find the probability at least one detects it. **Solution.** Probability neither detects it is $(0.1)(0.1) = 0.01$. Thus probability at least one detects it is $1 - 0.01 = 0.99$.

Class Work:

1. If $P(A) = 0.27$, calculate $P(A^c)$.
2. If mutually exclusive events have probabilities 0.12 and 0.23, find the probability of either event.
3. Two independent samples each have probability 0.2 of being positive. Find the probability both are positive.
4. Explain the difference between independent and mutually exclusive events.

Chapter 8

Probability Distributions

8.1 Discrete distributions

A probability distribution lists the possible values of a random variable and their probabilities. The probabilities must be non-negative and sum to 1.

8.1.1 Binomial distribution

Use a binomial model when there are a fixed number n of independent trials, each with the same probability p of success, and each trial has two outcomes. The probability of exactly x successes is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Mean $= np$ and variance $= np(1-p)$.

8.1.2 Poisson distribution

A Poisson model may describe counts in a fixed interval when events occur at a stable average rate and under appropriate independence assumptions:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad E(X) = \mathrm{Var}(X) = \lambda.$$

8.1.3 Normal distribution

A normal distribution is continuous, symmetric and bell-shaped. Its mean and standard deviation are μ and σ. Standardisation uses $Z = (X - \mu)/\sigma$. Probability areas are obtained from a standard normal table or software.

8.2 Worked examples

Worked Example

Example 1: Binomial exact probability. Suppose $n = 3$, $p = 0.2$. Find $P(X = 1)$. **Solution.** $\binom{3}{1}(0.2)^1(0.8)^2 = 3(0.2)(0.64) = 0.384$.

Worked Example

Example 2: Binomial mean and variance. If $n = 20$ and $p = 0.05$, calculate the mean and variance. **Solution.** Mean $= np = 20(0.05) = 1$. Variance $= np(1-p) = 20(0.05)(0.95) = 0.95$.

Worked Example

Example 3: Poisson probability. If a process averages $\lambda = 2$ events per interval, find the probability of zero events. **Solution.** $P(X = 0) = e^{-2}2^0/0! = e^{-2} \approx 0.1353$.

Worked Example

Example 4: Normal standard score. A measurement is 115, with reference model mean 100 and SD 10. Find z. **Solution.** $z = (115 - 100)/10 = 1.5$. The observation is 1.5 standard deviations above the model mean.

Worked Example

Example 5: Expected count. A screening programme tests 300 people and the assumed positive probability is 0.04. Find the expected number positive under a binomial model. **Solution.** $E(X) = np = 300(0.04) = 12$.

Class Work:

1. For $n = 4, p = 0.5$, calculate $P(X = 2)$.
2. For a Poisson model with $\lambda = 3$, write the formula for $P(X = 1)$ and evaluate it approximately.
3. If $X = 72$, $\mu = 60$, and $\sigma = 6$, calculate the standard score.
4. State two conditions for using a binomial model.

Chapter 9

Hypothesis Testing and Statistical Significance

9.1 The logic of a test

A hypothesis test uses sample data to assess a claim about a population parameter. The **null hypothesis** H_0 usually represents no difference or a specified reference value. The **alternative hypothesis** H_1 represents the difference or relationship of interest.

A p-value is the probability, assuming the null hypothesis and model assumptions are true, of obtaining a test statistic at least as extreme as the observed one. It is not the probability that the null hypothesis is true. Choose a significance level α before analysis; a common teaching value is 0.05. Statistical significance does not by itself imply clinical importance.

9.2 One-sample t test

When the population SD is unknown and a suitable sample-based model is appropriate:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad df = n - 1.$$

The test assumptions, direction (one- or two-sided), significance level and reference distribution must be stated.

9.3 Worked examples

Worked Example

Example 1: State hypotheses. A laboratory asks whether the population mean differs from 100. **Solution.** $H_0 : \mu = 100$; $H_1 : \mu \neq 100$. This is a two-sided test.

Worked Example

Example 2: Calculate a one-sample t. Reference mean 100; sample mean 104; sample SD 10; $n = 25$. **Solution.** Standard error $= 10/\sqrt{25} = 10/5 = 2$. Therefore $t = (104-100)/2 = 2.00$, with $df = 24$. A final decision requires the chosen α and the appropriate t distribution (or software) and assumptions.

Worked Example

Example 3: Interpret a p-value. A valid test reports $p = 0.03$ and the preselected $\alpha = 0.05$. **Solution.** Since $0.03 < 0.05$, reject H_0 under this testing rule. Report the result and effect estimate; do not say that there is a 3% probability that H_0 is true.

Worked Example

Example 4: Non-significant result. A test reports $p = 0.18$ with $\alpha = 0.05$. **Solution.** Since $0.18 > 0.05$, fail to reject H_0. This does not prove equality; the study may be imprecise or have limited sample size.

Worked Example

Example 5: Confidence and clinical relevance. A very large study finds a mean difference of 0.2 units with $p < 0.001$. **Solution.** The difference is statistically significant under the model, but its clinical or laboratory importance depends on context, measurement units, uncertainty and a meaningful threshold.

9.4 Basic decision table

Condition	Decision	Careful wording
$p \leq \alpha$	Reject H_0	Evidence against H_0 under assumptions
$p > \alpha$	Fail to reject H_0	Insufficient evidence against H_0

Class Work:

1. Write two-sided hypotheses for a reference mean of 50.
2. A sample has $\bar{x} = 52$, $s = 8$, $n = 16$, and $\mu_0 = 50$. Calculate the t statistic.
3. Explain why $p = 0.04$ does not mean there is a 4% chance the null hypothesis is true.
4. Distinguish statistical significance from clinical significance.

Chapter 10

Analysis of Variance (ANOVA)

10.1 Purpose

One-way ANOVA compares the means of three or more independent groups for a quantitative outcome. Its null hypothesis is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. The alternative is that at least one population mean differs. ANOVA does not by itself identify which pairs differ; follow-up comparisons may be needed with appropriate control of multiple testing.

Typical assumptions are independent observations, approximately normal residuals within groups, and similar population variances. Study design and data structure determine whether one-way ANOVA is appropriate.

10.2 ANOVA calculations

$$SS_T = \sum_i \sum_j (x_{ij} - \bar{x})^2, \quad SS_B = \sum_i n_i(\bar{x}_i - \bar{x})^2, \quad SS_W = SS_T - SS_B$$

$$df_B = k - 1, \quad df_W = N - k, \quad MS_B = \frac{SS_B}{df_B}, \quad MS_W = \frac{SS_W}{df_W}, \quad F = \frac{MS_B}{MS_W}.$$

10.3 Worked example: three laboratory methods

Hypothetical measurements:

Method A	Method B	Method C
8	10	13
9	11	14
10	12	15

Group means are $\bar{x}_A = 9$, $\bar{x}_B = 11$, $\bar{x}_C = 14$; grand mean $= (27 + 33 + 42)/9 = 102/9 = 11.333$.

$$SS_B = 3(9 - 11.333)^2 + 3(11 - 11.333)^2 + 3(14 - 11.333)^2$$

$$\approx 3(5.444) + 3(0.111) + 3(7.111) = 38.000.$$

Within-group squared deviations: Method A has $(-1)^2 + 0^2 + 1^2 = 2$; Method B has 2; Method C has 2. Thus $SS_W = 6$. Here $df_B = 2$, $df_W = 6$, $MS_B = 19$, $MS_W = 1$, and $F = 19$. The result indicates strong separation in this illustrative dataset; a formal conclusion should compare the statistic with an appropriate reference distribution or use software and verify assumptions.

10.4 Four worked examples

Worked Example

Example 1: Hypotheses. Compare mean results from three assay methods. **Solution.** $H_0 : \mu_A = \mu_B = \mu_C$. H_1 : at least one mean differs.

Worked Example

Example 2: Degrees of freedom. There are 4 groups and 28 observations. **Solution.** Between-group df $= k - 1 = 4 - 1 = 3$. Within-group df $= N - k = 28 - 4 = 24$. Total df $= N - 1 = 27$.

Worked Example

Example 3: Mean square. If $SS_B = 36$ and $k = 4$, calculate MS_B. **Solution.** $df_B = 4 - 1 = 3$, so $MS_B = 36/3 = 12$.

Worked Example

Example 4: F statistic. If $MS_B = 12$ and $MS_W = 3$, calculate F. **Solution.** $F = 12/3 = 4$. Statistical significance requires the appropriate F reference distribution and degrees of freedom.

Class Work:

1. State the null and alternative hypotheses for comparing four treatment means.
2. Find between- and within-group degrees of freedom for $k = 3$, $N = 24$.
3. If $SS_W = 40$ and $N - k = 20$, find MS_W.
4. Explain why a significant ANOVA result does not show exactly which groups differ.

Chapter 11

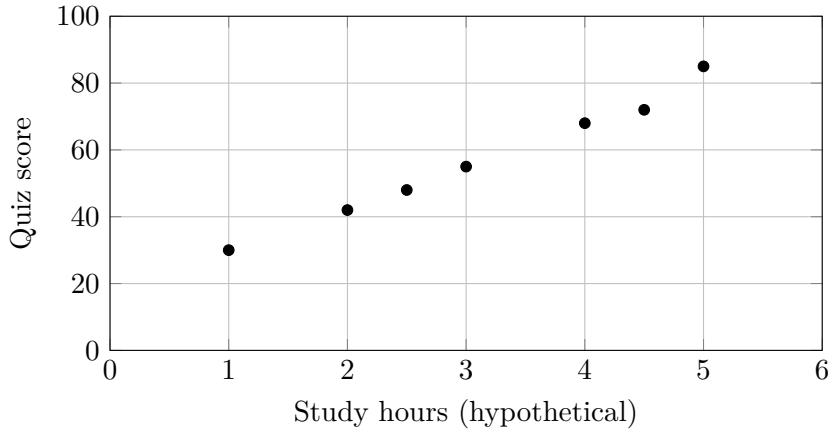
Correlation Analysis

11.1 Meaning

Correlation quantifies the direction and strength of a linear association between two quantitative variables. Pearson's correlation coefficient r lies between -1 and $+1$. Values near $+1$ indicate a strong positive linear association; values near -1 indicate a strong negative linear association; values near zero indicate little linear association. A near-zero Pearson coefficient does not rule out a nonlinear relationship.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

11.2 Scatter plot



The plotted points show a positive association. This does not prove that study hours alone caused the scores; other variables may contribute.

11.3 Four worked examples

Worked Example

Example 1: Interpret $r = 0.82$. Solution. This indicates a strong positive linear association in the sample. It does not establish causation.

Worked Example

Example 2: Interpret $r = -0.76$. Solution. This indicates a strong negative linear association: larger values of one variable tend to accompany smaller values of the other, in a roughly linear pattern.

Worked Example

Example 3: Interpret $r = 0.03$. Solution. There is almost no observed linear association. Inspect a scatter plot because a nonlinear pattern could still exist.

Worked Example

Example 4: Correlation versus causation. A study finds that hospital attendance and diagnostic tests rise together. **Solution.** The positive correlation may reflect more patients, a policy change or another common cause. Correlation alone cannot establish that one variable causes the other.

Class Work:

1. Interpret $r = 0.65$ in a biomedical context.
2. Explain the meaning of $r = -1$, $r = 0$ and $r = +1$ for linear association.
3. Why should a scatter plot be examined before interpreting Pearson correlation?
4. Give one example of a confounding variable in a biomedical study.

Chapter 12

Simple Linear Regression

12.1 Purpose and equation

Regression models the mean relationship between an outcome Y and predictor X. Simple linear regression is

$$\hat{Y} = a + bX,$$

where b is the estimated slope and a is the intercept. The slope is the estimated change in mean outcome for a one-unit increase in X. Prediction should remain within a sensible range of the observed data; extrapolation can be misleading.

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x}.$$

12.2 Four worked examples

Worked Example

Example 1: Interpret a fitted line. A fitted model is $\hat{Y} = 12 + 2.5X$. **Solution.** Intercept is 12 predicted units at $X = 0$ (if meaningful). Slope is 2.5: each one-unit increase in X is associated with an increase of 2.5 predicted units in Y.

Worked Example

Example 2: Predict from a line. Use $\hat{Y} = 12 + 2.5X$ to predict at $X = 4$. **Solution.** $\hat{Y} = 12 + 2.5(4) = 12 + 10 = 22$.

Worked Example

Example 3: Calculate slope. Given $\sum(x - \bar{x})(y - \bar{y}) = 30$ and $\sum(x - \bar{x})^2 = 10$. **Solution.** $b = 30/10 = 3$.

Worked Example

Example 4: Calculate intercept. If $\bar{x} = 4$, $\bar{y} = 18$, and $b = 3$. **Solution.** $a = \bar{y} - b\bar{x} = 18 - 3(4) = 6$. The fitted line is $\hat{Y} = 6 + 3X$.

Worked Example

Example 5: Limitation. A model based on ages 18–50 is used to predict an outcome at age 95. **Solution.** This is extrapolation beyond the observed age range. The relationship may not continue, so the prediction requires strong justification and may be unreliable.

Class Work:

1. Interpret slope $b = -1.2$ in context.
2. Use $\hat{Y} = 5 + 4X$ to predict Y when $X = 7$.
3. If $\bar{x} = 3$, $\bar{y} = 14$, and $b = 2$, calculate the intercept.
4. State two limitations of regression prediction.

Chapter 13

Experimental Design in Clinical and Laboratory Research

13.1 Key principles

Randomisation Allocating eligible experimental units to groups using a chance process, helping balance known and unknown factors on average.

Replication Including multiple independent experimental units per treatment to estimate variability and improve precision.

Control group A comparison group that helps separate treatment effects from background changes, placebo effects or routine practice.

Blinding Concealing treatment or group assignment from participants, assessors or analysts where feasible to reduce bias.

Standardisation Applying procedures consistently across groups.

13.2 Common designs

A parallel-group design assigns different participants or specimens to different treatments. A paired design compares two measurements on the same unit or on matched units. A completely randomised design assigns units to treatments entirely by chance. Blocking groups similar units before randomisation when an important source of variation is known.

13.3 Four worked examples

Worked Example

Example 1: Randomisation. Thirty independent specimens are assigned by computer to Method A or Method B. **Solution.** Random allocation reduces the risk that one method receives systematically different specimens. The specimens, not repeated readings from the same specimen, are the experimental units if treatment is applied at specimen level.

Worked Example

Example 2: Replication. One specimen is measured ten times using a new assay. **Solution.** These repeated measurements describe repeatability for that specimen but do not

provide ten independent biological specimens. Independent replication is needed to generalise across specimens.

Worked Example

Example 3: Control group. A study evaluates a new prevention programme. **Solution.** A suitable comparison group receiving usual practice may help estimate the programme's additional effect, provided allocation, follow-up and other conditions are handled appropriately.

Worked Example

Example 4: Blocking. A laboratory knows that machine batch affects results and compares two reagents. **Solution.** Units can be grouped by machine batch, then each reagent allocated within each batch. This helps account for batch variability in the comparison.

Worked Example

Example 5: Identify design elements. A study compares three storage temperatures using 12 independently prepared samples per temperature. **Solution.** Experimental units: prepared samples. Factor: storage temperature. Treatments: the three temperature levels. Outcome: the measured analyte concentration after storage. Replication: 12 samples per level. Randomisation: random assignment of samples to storage conditions.

Class Work:

1. Define randomisation, replication and a control group.
2. Explain why ten readings from one specimen are not equivalent to ten independent specimens.
3. Propose a design comparing two laboratory methods and identify the outcome variable.
4. Explain how blinding may reduce measurement bias.

Chapter 14

Integrated Biomedical Applications and Revision

14.1 Choosing an appropriate method

Research question	Data/design	Possible method
What is the average haemoglobin in a sample?	One quantitative variable	Mean and SD; median and IQR if skewed
How common is a condition?	Cases and a defined population	Prevalence or incidence, with clear denominator
Do three independent methods have different means?	Quantitative outcome, three groups	One-way ANOVA if assumptions are reasonable
Are two quantitative variables linearly associated?	Paired quantitative values	Scatter plot and Pearson correlation
Can one variable predict another?	Outcome and predictor	Regression, with residual checks
Does a sample mean differ from a reference value?	Quantitative sample	One-sample t test if appropriate

14.2 Integrated worked example

A hypothetical laboratory records haemoglobin values (g/dL):

$$11,\ 12,\ 13,\ 12,\ 14,\ 15,\ 12,\ 13,\ 11,\ 14.$$

Step 1: Sort the data. $11, 11, 12, 12, 12, 13, 13, 14, 14, 15$.

Step 2: Mean.

$$\bar{x} = \frac{11 + 12 + 13 + 12 + 14 + 15 + 12 + 13 + 11 + 14}{10} = \frac{127}{10} = 12.7.$$

Step 3: Median. The fifth and sixth ordered values are 12 and 13, so median $= (12 + 13)/2 = 12.5$.

Step 4: Mode. The most frequent value is 12, so mode = 12.

Step 5: Range. $15 - 11 = 4$ g/dL.

Step 6: Interpretation. In this hypothetical sample, the mean is 12.7 g/dL, median 12.5 g/dL, mode 12 g/dL and range 4 g/dL. These summaries describe only this dataset; clinical interpretation requires appropriate reference intervals, patient context and validated measurement procedures.

14.3 Revision questions

1. Define biostatistics and give five biomedical applications.
2. Distinguish a population parameter from a sample statistic.
3. Classify nominal, ordinal, discrete and continuous variables with examples.
4. Construct a frequency table and bar chart for $2, 3, 3, 4, 4, 4, 5, 6, 6, 7$.
5. Calculate the crude birth rate for 720 live births in a mid-year population of 36,000.
6. Find the mean, median and mode of $4, 5, 5, 7, 9, 12$.
7. Calculate the range and population standard deviation of $2, 4, 4, 6$.
8. A test is positive with probability 0.12. Find the probability of a negative result.
9. For $n = 5$ and $p = 0.2$, calculate the binomial probability of exactly one success.
10. State null and alternative hypotheses for a test of whether μ differs from 25.
11. Explain one-way ANOVA and state its main assumptions.
12. Interpret a Pearson correlation of $r = -0.70$.
13. Fit the intercept if $\bar{x} = 5$, $\bar{y} = 17$ and slope $b = 2$.
14. Describe randomisation, replication and control groups.
15. Explain why association should not automatically be interpreted as causation.

Selected Answers to Class Work

Chapter 1: Answers should identify statistics as methods for collecting, summarising, analysing and interpreting data; applications include laboratory quality evaluation, treatment comparisons, surveillance, screening and planning research.

Chapter 2: Age is usually quantitative (often continuous in principle); blood group is nominal; positive culture count is discrete; severity low/medium/high is ordinal. Every answer should define the target population and sample clearly.

Chapter 3: For $4, 5, 4, 6, 7, 5, 4, 6, 5, 8$, frequencies are 4:3, 5:3, 6:2, 7:1, 8:1. For 18 of 90 positive, percentage $= 20\%$ and angle $= 72^\circ$.

Chapter 4: Birth rate $= (900/60,000) \times 1000 = 15$ per 1,000. Death rate $= (150/30,000) \times 1000 = 5$ per 1,000. Case-fatality $= (16/400) \times 100 = 4\%$.

Chapter 5: For 6, 8, 8, 9, 10, 13: mean $= 54/6 = 9$, median $= (8 + 9)/2 = 8.5$, mode $= 8$.

Chapter 6: Range of 11, 13, 13, 18, 22 is 11. Population variance for 1, 3, 5 is $8/3$ and SD ≈ 1.633. CV for X is 5%; for Y is 10%.

Chapter 7: If $P(A) = 0.27$, complement $= 0.73$. Mutually exclusive probabilities 0.12 and 0.23 sum to 0.35. Independent probability both positive $= 0.2^2 = 0.04$.

Chapter 8: For $n = 4, p = 0.5$, $P(X = 2) = \binom{4}{2}(0.5)^4 = 6/16 = 0.375$. For Poisson $\lambda = 3$, $P(X = 1) = 3e^{-3} \approx 0.1494$. Standard score $(72 - 60)/6 = 2$.

Chapter 9: For $\bar{x} = 52, s = 8, n = 16, \mu_0 = 50$, $t = (52 - 50)/(8/4) = 1$ with $df = 15$.

Chapter 10: For $k = 3, N = 24$, between-group df $= 2$, within-group df $= 21$. If $SS_W = 40$ and df within $= 20$, $MS_W = 2$.

Chapter 11: A coefficient $r = 0.65$ indicates a moderately strong positive linear association, subject to the context and scatter plot.

Chapter 12: With $\bar{x} = 3, \bar{y} = 14, b = 2$, intercept $a = 14 - 2(3) = 8$.

Appendix A

Advanced Worked Examples by Subtopic

This supplementary workbook expands the core lecture with university-level numerical problems. Each question is followed immediately by a worked solution. The problems use hypothetical data and reinforce both computation and interpretation. Where a test requires a reference distribution, assumptions or a critical value, the solution states what can be concluded from the information supplied and what additional information is needed.

A.1 Applications and interpretation of biostatistics

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 1.1

Question. A screening programme tests 800 people; 96 test positive. Calculate the observed positive proportion and explain what it does and does not estimate.

Worked solution. Working: $\hat{p} = 96/800 = 0.12 = 12\%$. This is the sample positive proportion. It estimates prevalence only if the sampling design, case definition and target population make that interpretation appropriate.

Problem 1.2

Question. Two methods give mean results 48.2 and 50.1 units. State the descriptive quantities needed before deciding whether the methods differ meaningfully.

Worked solution. Working: Report each sample size, mean, SD (or suitable robust spread), the mean difference $50.1 - 48.2 = 1.9$ units, and a confidence interval or appropriate paired/independent comparison based on the design. Means alone cannot quantify uncertainty.

Problem 1.3

Question. A surveillance unit records weekly case counts for 12 weeks. Name a suitable display and two cautions.

Worked solution. Working: Use a time-series line graph with week on the horizontal axis and case count on the vertical axis. Annotate changes in testing, reporting or population size;

counts alone may not equal incidence rates.

Problem 1.4

Question. A study estimates the mean glucose of all eligible clinic attendees from a random sample of 150. Identify descriptive and inferential tasks.

Worked solution. Working: The sample mean and SD are descriptive summaries. A confidence interval for the population mean and a test against a reference value are inferential tasks, provided assumptions and sampling design are suitable.

Problem 1.5

Question. A report says a treatment is effective because its group mean improved. Identify the missing evidence.

Worked solution. Working: Improvement alone does not rule out time effects, regression to the mean or confounding. Examine a comparison group, allocation design, effect estimate, uncertainty interval, missing data and clinically meaningful outcome criteria.

Problem 1.6

Question. A study reports a 95

Worked solution. Working: Midpoint $= (4.8 + 5.6)/2 = 10.4/2 = 5.2$ units. The interval width is 0.8 units; it quantifies uncertainty under the method's assumptions.

Problem 1.7

Question. A sample proportion is 0.18 from $n = 250$. Calculate the number of positive observations.

Worked solution. Working: Number positive $= 0.18(250) = 45$. The percentage is 18

Problem 1.8

Question. A report gives a relative risk of 1.8. Interpret it carefully.

Worked solution. Working: The exposed group has an estimated risk 1.8 times the unexposed group's risk over the defined period. Interpretation needs a confidence interval, valid comparison groups and consideration of confounding.

Problem 1.9

Question. A laboratory records 7, 9, 12, 12 and 15 minutes as turnaround times. Find the total and mean.

Worked solution. Working: Total $= 7 + 9 + 12 + 12 + 15 = 55$ minutes; mean $= 55/5 = 11$ minutes.

Problem 1.10

Question. A study measures 20 variables in 15 participants. Explain why a large variable count does not compensate for a small sample.

Worked solution. Working: Precision and model stability depend on independent information and study design. Many variables with few participants can cause overfitting, unstable estimates and low power.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.2 Populations, samples, parameters and sampling

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 2.1

Question. A university has 2,400 eligible students; 180 are sampled. The sample mean age is 20.4 years. Identify the population, sample, statistic and corresponding parameter.

Worked solution. Working: Population: all 2,400 eligible students. Sample: 180 selected students. Statistic: sample mean age 20.4 years. Parameter: the unknown mean age of all 2,400 eligible students.

Problem 2.2

Question. A frame lists 1,200 specimens. A random start is chosen from 1 to 12 and every 12th specimen is selected. Identify the method and sample size.

Worked solution. Working: This is systematic sampling. Approximate sample size $n = 1200/12 = 100$, assuming the list length and interval align. A periodic ordering pattern can introduce bias.

Problem 2.3

Question. A population contains 600 urban and 400 rural patients. A proportional stratified sample of 100 is needed. Find the sample allocation.

Worked solution. Working: Urban allocation $= 100(600/1000) = 60$. Rural allocation $= 100(400/1000) = 40$. Total $60 + 40 = 100$.

Problem 2.4

Question. A researcher selects only patients who attend on Monday morning. Explain the likely limitation.

Worked solution. Working: This is a convenience/time-specific sample. Monday attendees may differ from other patients in access, severity or work schedule, so selection bias may limit generalisability.

Problem 2.5

Question. Classify blood group, pain severity (mild/moderate/severe), colony count, and serum sodium concentration.

Worked solution. Working: Blood group: nominal categorical. Pain severity: ordinal categorical. Colony count: discrete quantitative. Serum sodium concentration: continuous quantitative.

Problem 2.6

Question. A population has 70

Worked solution. Working: Sample proportion = 84/120 = 0.70 = 70%, matching the stated population proportion in this sample. Agreement does not guarantee all other sample characteristics are representative.

Problem 2.7

Question. A sample contains 45 rural patients and 105 urban patients. Find the rural percentage.

Worked solution. Working: Total = 150. Rural percentage = 45/150(100) = 30%.

Problem 2.8

Question. A population of 5,000 is sampled at 4

Worked solution. Working: $n = 0.04(5000) = 200$ units.

Problem 2.9

Question. Explain the difference between non-response bias and random sampling error.

Worked solution. Working: Random sampling error is chance variation between samples. Non-response bias occurs if responders systematically differ from non-responders in ways related to the study variable.

Problem 2.10

Question. A study samples 25 patients from each of four hospitals. What sampling structure is this, and what should be considered in analysis?

Worked solution. Working: It is stratified sampling by hospital if sampling occurs within each hospital. Account for the design and hospital structure; equal sample sizes may require weighting for population estimates if hospital sizes differ.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.3 Frequency distributions and grouped data

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 3.1

Question. Construct a frequency distribution for 2, 3, 2, 4, 5, 3, 2, 4, 4, 5, 6, 3. Include relative and cumulative frequencies.

Worked solution. Working: Frequencies: 2:3, 3:3, 4:3, 5:2, 6:1; $n = 12$. Relative frequencies are $3/12 = .25$, .25, .25, $2/12 = .1667$, $1/12 = .0833$. Cumulative frequencies: 3, 6, 9, 11, 12.

Problem 3.2

Question. For classes 10–19, 20–29, 30–39, 40–49 with frequencies 4, 7, 11, 8, find class midpoints and cumulative frequencies.

Worked solution. Working: Midpoints: 14.5, 24.5, 34.5, 44.5. Cumulative frequencies: 4, 11, 22, 30. Total $n = 30$.

Problem 3.3

Question. A dataset has 48 observations. A class has frequency 9. Calculate its relative and percentage frequency.

Worked solution. Working: Relative frequency $= 9/48 = 0.1875$. Percentage frequency $= 0.1875(100) = 18.75\%$.

Problem 3.4

Question. Using classes 0–9, 10–19, 20–29 with frequencies 5, 12, 3, identify the modal class and estimate its midpoint.

Worked solution. Working: The modal class is 10–19 because it has the greatest frequency, 12. Its midpoint is $(10 + 19)/2 = 14.5$. The midpoint is not the exact mode.

Problem 3.5

Question. A grouped table has frequencies 6, 10, 14, 8, 2. Find total sample size and the cumulative percentage after the third class.

Worked solution. Working: Total $n = 6 + 10 + 14 + 8 + 2 = 40$. Cumulative frequency through class 3 is $6 + 10 + 14 = 30$. Cumulative percentage $= 30/40(100) = 75\%$.

Problem 3.6

Question. For frequencies 5, 8, 12, 9, 6, find total, modal frequency and modal class position.

Worked solution. Working: Total $= 5 + 8 + 12 + 9 + 6 = 40$. Highest frequency is 12, so the third class is modal.

Problem 3.7

Question. A value occurs 17 times in a sample of 85. Find percentage frequency.

Worked solution. Working: $17/85(100) = 20\%$.

Problem 3.8

Question. For classes 0–4, 5–9, 10–14, 15–19, find real boundaries and width for class 10–14.

Worked solution. Working: For integer measurements, real boundaries are 9.5 and 14.5. Width $= 14.5 - 9.5 = 5$.

Problem 3.9

Question. Cumulative frequencies are 4, 13, 29, 42. Find the class frequencies.

Worked solution. Working: First frequency $= 4$; second $= 13 - 4 = 9$; third $= 29 - 13 = 16$; fourth $= 42 - 29 = 13$.

Problem 3.10

Question. Grouped classes have midpoints 5, 15, 25 and frequencies 4, 7, 9. Estimate the grouped mean.

Worked solution. Working: $\sum fx = 5(4) + 15(7) + 25(9) = 20 + 105 + 225 = 350$. Total frequency $= 20$. Grouped mean $= 350/20 = 17.5$.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.4 Charts, pie charts, histograms, ogives and graphs

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 4.1

Question. A category represents 27 of 90 observations. Calculate its pie-chart angle and percentage.

Worked solution. Working: Percentage $= 27/90(100) = 30\%$. Angle $= 27/90(360^\circ) = 108^\circ$.

Problem 4.2

Question. A histogram class spans 20–30 and has frequency density 1.8. Find the bar area representing frequency.

Worked solution. Working: Frequency density $= f/$class width. Thus $f = 1.8(30 - 20) = 18$. Histogram area, not height alone, represents frequency when class widths differ.

Problem 4.3

Question. Cumulative frequencies at upper boundaries 10, 20, 30, 40 are 3, 11, 25, 32. Estimate the number of observations between 20 and 40.

Worked solution. Working: $CF(40) - CF(20) = 32 - 11 = 21$ observations.

Problem 4.4

Question. A line graph shows weekly cases 12, 18, 15, 9. Calculate the percentage decrease from week 2 to week 4.

Worked solution. Working: Decrease $= 18 - 9 = 9$. Percentage decrease $= 9/18(100) = 50\%$.

Problem 4.5

Question. Explain why a bar chart is preferred over a histogram for blood group counts, and state the key difference.

Worked solution. Working: Blood group is categorical, so use a bar chart. Bars are separated to show distinct categories. A histogram represents quantitative measurements grouped into adjoining intervals, so bars normally touch.

Problem 4.6

Question. A pie chart has 72 degrees for a category. Find its percentage and frequency if $n = 250$.

Worked solution. Working: Proportion $= 72/360 = 0.20$; percentage $= 20\%$; frequency $= 0.20(250) = 50$.

Problem 4.7

Question. An ogive has cumulative frequency 18 at boundary 30 and 31 at boundary 40. Estimate the number in the interval (30,40].

Worked solution. Working: Frequency in interval $= 31 - 18 = 13$.

Problem 4.8

Question. A class has width 5 and frequency 20. Calculate histogram frequency density.

Worked solution. Working: Density $= f/width = 20/5 = 4$.

Problem 4.9

Question. A line graph changes from 40 cases to 52 cases. Calculate percentage increase.

Worked solution. Working: Increase $= 12$. Percentage increase $= 12/40(100) = 30\%$.

Problem 4.10

Question. Explain why a truncated vertical axis can mislead readers of a bar chart.

Worked solution. Working: It exaggerates visual differences because bar lengths no longer reflect values from a zero baseline. Clearly mark any axis break and avoid distortion.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.5 Vital statistics and health indicators

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 5.1

Question. A district population is 42,000 and records 630 live births in a year. Calculate crude birth rate per 1,000.

Worked solution. Working: $CBR = (630/42000)(1000) = 15$ live births per 1,000 population per year.

Problem 5.2

Question. A population of 125,000 records 875 deaths in one year. Calculate crude death rate per 1,000.

Worked solution. Working: $CDR = (875/125000)(1000) = 7$ deaths per 1,000 population per year.

Problem 5.3

Question. Among 360 identified cases, 18 deaths are attributed to the disease. Calculate case-fatality percentage.

Worked solution. Working: $(18/360)(100) = 5\%$. The case definition and completeness of follow-up must be clear.

Problem 5.4

Question. At the start of follow-up, 2,500 disease-free people are at risk; 75 develop the disease. Calculate incidence proportion.

Worked solution. Working: $(75/2500)(100) = 3\%$ over the stated follow-up period.

Problem 5.5

Question. A cross-sectional survey finds 144 existing cases among 3,600 people. Calculate point prevalence and explain its interpretation.

Worked solution. Working: $(144/3600)(100) = 4\%$. Four percent of those assessed met the case definition at the survey time; this is not an incidence rate.

Problem 5.6

Question. A town has 18,000 people and 270 births in a year. Calculate crude birth rate.

Worked solution. Working: $(270/18000)(1000) = 15$ births per 1,000 population per year.

Problem 5.7

Question. A hospital identifies 520 cases and 13 disease-attributed deaths. Calculate case-fatality percentage.

Worked solution. Working: $(13/520)(100) = 2.5\%$.

Problem 5.8

Question. In a population of 4,000, 160 people have a condition at a survey date. Calculate prevalence.

Worked solution. Working: $(160/4000)(100) = 4\%$.

Problem 5.9

Question. Among 900 people initially disease-free, 27 develop disease during one year. Calculate incidence proportion.

Worked solution. Working: $(27/900)(100) = 3\%$ over one year.

Problem 5.10

Question. Explain why a crude death rate can mislead when comparing populations with different age structures.

Worked solution. Working: Older populations often have higher crude mortality because age strongly affects death risk. Age-specific rates or age-standardised rates may be needed for fairer comparison.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.6 Mean and weighted/frequency mean

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 6.1

Question. Calculate the mean of 14, 18, 21, 23, 24, 30.

Worked solution. Working: Sum $= 14 + 18 + 21 + 23 + 24 + 30 = 130$. $n = 6$. Mean $= 130/6 = 21.667$.

Problem 6.2

Question. Values 5, 10, 15, 20 have frequencies 3, 4, 2, 1. Find the frequency mean.

Worked solution. Working: $\sum fx = 5(3) + 10(4) + 15(2) + 20(1) = 105$. $\sum f = 10$. Mean $= 105/10 = 10.5$.

Problem 6.3

Question. A course grade is based on tests weighted 20

Worked solution. Working: Weighted mean $= .20(65) + .30(72) + .50(84) = 13 + 21.6 + 42 = 76.6$.

Problem 6.4

Question. A sample of 40 has mean 12.5 and a second sample of 60 has mean 14.0. Calculate the combined mean.

Worked solution. Working: Combined mean $= [40(12.5) + 60(14.0)]/(40 + 60) = (500 + 840)/100 = 13.4$.

Problem 6.5

Question. A single value 100 is added to the dataset 4, 6, 8, 10. Find the old and new means and explain the change.

Worked solution. Working: Old mean $= (4+6+8+10)/4 = 7$. New mean $= (28+100)/5 = 25.6$. The extreme value pulls the mean upward.

Problem 6.6

Question. Find the mean of 22, 24, 26, 29, 34.

Worked solution. Working: Sum $= 135$. Mean $= 135/5 = 27$.

Problem 6.7

Question. A frequency table has $x = 2, 4, 6, 8$ and $f = 1, 3, 4, 2$. Find the mean.

Worked solution. Working: $\sum fx = 2 + 12 + 24 + 16 = 54$; $\sum f = 10$. Mean $= 5.4$.

Problem 6.8

Question. A final score combines coursework 68 weighted 40

Worked solution. Working: $0.40(68) + 0.60(75) = 27.2 + 45 = 72.2$.

Problem 6.9

Question. Twenty samples have mean 14 and thirty samples have mean 16. Find combined mean.

Worked solution. Working: $(20(14) + 30(16))/50 = (280 + 480)/50 = 15.2$.

Problem 6.10

Question. The mean of 8 observations is 12.5. Find their total sum.

Worked solution. Working: $\sum x = n\bar{x} = 8(12.5) = 100$.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.7 Median and quartiles

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 7.1

Question. Find the median of 13, 5, 9, 7, 15, 11, 8.

Worked solution. Working: Ordered data: 5, 7, 8, 9, 11, 13, 15. With $n = 7$, the fourth value is the median: 9.

Problem 7.2

Question. Find the median of 4, 12, 7, 10, 15, 8.

Worked solution. Working: Ordered data: 4, 7, 8, 10, 12, 15. Median $= (8 + 10)/2 = 9$.

Problem 7.3

Question. For ordered values 2, 4, 5, 7, 8, 10, 12, 14, find the median and the lower/upper half medians using the median-of-halves convention.

Worked solution. Working: Median $= (7 + 8)/2 = 7.5$. Lower half 2,4,5,7 gives $Q_1 = (4 + 5)/2 = 4.5$. Upper half 8,10,12,14 gives $Q_3 = (10 + 12)/2 = 11$. Quartile conventions can differ by software.

Problem 7.4

Question. A dataset has mean 45 and median 38. What might this suggest, and what must be checked?

Worked solution. Working: Mean greater than median may indicate right-skewness or high values pulling the mean upward. Check a histogram/box plot and the raw data; the inequality alone does not prove skewness.

Problem 7.5

Question. Why may the median be more informative than the mean for hospital length of stay? Give a numerical illustration.

Worked solution. Working: Length of stay is often right-skewed. For 1, 2, 2, 3, 20 days, mean $= 28/5 = 5.6$ days while median $= 2$ days. The long stay strongly inflates the mean.

Problem 7.6

Question. Find the median of 18, 12, 20, 16, 14, 10, 22, 15, 13.

Worked solution. Working: Ordered: 10,12,13,14,15,16,18,20,22. Fifth value is 15.

Problem 7.7

Question. Find median of 3, 5, 9, 11, 14, 17, 20, 24.

Worked solution. Working: Median $= (11 + 14)/2 = 12.5$.

Problem 7.8

Question. A sorted dataset has 21 observations. Which position contains the median?

Worked solution. Working: Position $(n + 1)/2 = (21 + 1)/2 = 11$. The 11th ordered observation is the median.

Problem 7.9

Question. If $Q_1 = 12$ and $Q_3 = 28$, calculate the interquartile range.

Worked solution. Working: $IQR = Q_3 - Q_1 = 28 - 12 = 16$.

Problem 7.10

Question. A dataset has median 10 and IQR 4. Interpret these summaries.

Worked solution. Working: Half the observations are at or below 10 and half at or above, subject to convention. The middle 50

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.8 Mode and shape of distributions

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 8.1

Question. Find the mode(s) of 2, 4, 4, 5, 6, 6, 7.

Worked solution. Working: Values 4 and 6 each occur twice, more than the others. The data are bimodal.

Problem 8.2

Question. A sample contains blood groups A, B, AB, O, O, A, O. Identify the mode and explain why mean is unsuitable.

Worked solution. Working: O occurs three times and is the mode. Blood group is nominal, so arithmetic averaging has no meaningful interpretation.

Problem 8.3

Question. Data are 1, 2, 3, 4, 5. Is there a mode?

Worked solution. Working: Each value occurs once, so no unique mode exists.

Problem 8.4

Question. A frequency table gives values 1, 2, 3, 4 with frequencies 2, 7, 7, 1. Interpret the modal values.

Worked solution. Working: Values 2 and 3 tie at frequency 7, so the distribution is bimodal in the observed data.

Problem 8.5

Question. Compare mean, median and mode as summaries for a highly skewed laboratory turnaround-time distribution.

Worked solution. Working: The mean uses all values but is sensitive to long delays; median describes the midpoint robustly; mode identifies the most common value or class. Report median and IQR alongside graphical inspection when skewness is substantial.

Problem 8.6

Question. Find the mode of 4, 4, 5, 6, 6, 6, 7, 8.

Worked solution. Working: 6 occurs three times, the greatest frequency, so mode $= 6$.

Problem 8.7

Question. For categories A, B, B, C, A, B, C, identify mode.

Worked solution. Working: B occurs three times; A and C occur twice each. Mode is B.

Problem 8.8

Question. A distribution is left-skewed. Which is often larger: mean or median?

Worked solution. Working: The long left tail often pulls the mean downward, so mean is often less than median. This is a tendency, not an absolute rule.

Problem 8.9

Question. A distribution has two peaks. What is it called, and what should be investigated?

Worked solution. Working: It is bimodal. Investigate whether the sample combines two subpopulations, measurement processes or other meaningful groups.

Problem 8.10

Question. Why can the mode be unstable in continuous measurements?

Worked solution. Working: Exact values may rarely repeat, and the modal class can change with bin width or cut-points. Use suitable density displays and robust summaries.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.9 Range, variance and standard deviation

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 9.1

Question. Calculate the range of 18, 22, 17, 25, 31, 20.

Worked solution. Working: Maximum = 31, minimum = 17. Range = $31 - 17 = 14$.

Problem 9.2

Question. Calculate population variance and SD for 2, 4, 6, 8.

Worked solution. Working: Mean = 5. Squared deviations: 9,1,1,9; sum = 20. Population variance = $20/4 = 5$; SD = $\sqrt{5} = 2.236$.

Problem 9.3

Question. Calculate sample variance and SD for 3, 5, 8.

Worked solution. Working: Mean = $16/3 = 5.333$. Squared deviations sum = $(-7/3)^2 + (-1/3)^2 + (8/3)^2 = 114/9 = 12.667$. Sample variance = $12.667/(3-1) = 6.333$; sample SD = $\sqrt{6.333} = 2.517$.

Problem 9.4

Question. For values 6, 6, 6, 6, find range, variance and SD.

Worked solution. Working: Every observation equals 6. Range = 0; every deviation from the mean is zero; variance = 0 and SD = 0.

Problem 9.5

Question. A sample has mean 80, sample SD 12 and $n = 36$. Calculate the standard error of the mean.

Worked solution. Working: $SE = s/\sqrt{n} = 12/\sqrt{36} = 12/6 = 2$. Standard error describes sampling uncertainty in the mean, not the spread of individual observations.

Problem 9.6

Question. Find range for 32, 28, 41, 35, 29.

Worked solution. Working: Range = $41 - 28 = 13$.

Problem 9.7

Question. Find population variance and SD for 1, 2, 3, 4, 5.

Worked solution. Working: Mean = 3. Squared deviations sum = $4 + 1 + 0 + 1 + 4 = 10$. Population variance = $10/5 = 2$; SD = $\sqrt{2} \approx 1.414$.

Problem 9.8

Question. For sample data 4, 7, 9, find sample variance.

Worked solution. Working: Mean = 20/3. Squared deviations sum = $(-8/3)^2 + (1/3)^2 + (7/3)^2 = 114/9 = 12.667$. Sample variance = $12.667/2 = 6.333$.

Problem 9.9

Question. A population SD is 6. What is the variance?

Worked solution. Working: Variance = $SD^2 = 6^2 = 36$ squared units.

Problem 9.10

Question. Two datasets have equal means but SDs 2 and 7. Compare their spread.

Worked solution. Working: The second dataset has greater absolute dispersion around its mean. Whether this is important depends on units and context.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.10 Coefficient of variation and precision

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 10.1

Question. Method A has mean 40 and SD 2. Calculate CV.

Worked solution. Working: $CV = (2/40)(100) = 5\%$.

Problem 10.2

Question. Method B has mean 120 and SD 9. Calculate CV and compare with Method A above.

Worked solution. Working: $CV_B = (9/120)(100) = 7.5\%$. Method A has lower relative variability (5

Problem 10.3

Question. Two instruments have SDs 1.5 and 2.0, with means 30 and 50 respectively. Compare relative variation.

Worked solution. Working: $CV_1 = 1.5/30(100) = 5\%$. $CV_2 = 2/50(100) = 4\%$. Instrument 2 has lower relative variability despite its larger absolute SD.

Problem 10.4

Question. A measurement has mean near zero. Why can CV be misleading?

Worked solution. Working: Dividing by a very small mean makes CV extremely large or unstable; if the mean is zero, it is undefined. Use absolute spread measures and context instead.

Problem 10.5

Question. Explain why a lower CV alone does not prove an assay is more accurate.

Worked solution. Working: CV measures relative dispersion/precision, not closeness to a true or reference value. A method can be consistently biased and still have a small CV.

Problem 10.6

Question. Mean 75 and SD 3.75: calculate CV.

Worked solution. Working: $CV = 3.75/75(100) = 5\%$.

Problem 10.7

Question. Assay A: mean 25, SD 1.5; assay B: mean 80, SD 3.2. Calculate both CVs.

Worked solution. Working: A: $1.5/25(100) = 6\%$. B: $3.2/80(100) = 4\%$. B has lower relative dispersion.

Problem 10.8

Question. If CV is 8

Worked solution. Working: $SD = (CV/100)(mean) = 0.08(50) = 4$.

Problem 10.9

Question. If SD is 6 and CV is 12

Worked solution. Working: $0.12 = 6/\bar{x}$, so $\bar{x} = 6/0.12 = 50$.

Problem 10.10

Question. Why must units be comparable when using CV to compare two methods?

Worked solution. Working: CV is scale-free for ratio-scale data with meaningful positive means, but differences in analyte, operating range or measurement meaning can still make a direct comparison inappropriate.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.11 Probability rules and conditional probability

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 11.1

Question. If $P(A) = 0.35$, calculate the complement.
Worked solution. Working: $P(A^c) = 1 - 0.35 = 0.65$.

Problem 11.2

Question. Events A and B have probabilities 0.4 and 0.3 and are mutually exclusive. Find $P(A \cup B)$.
Worked solution. Working: $P(A \cup B) = P(A) + P(B) = 0.4 + 0.3 = 0.7$.

Problem 11.3

Question. If $P(A) = 0.5$, $P(B) = 0.4$, and $P(A \cap B) = 0.2$, calculate $P(A \cup B)$.
Worked solution. Working: $P(A \cup B) = 0.5 + 0.4 - 0.2 = 0.7$.

Problem 11.4

Question. Among 150 specimens, 60 are from ward X and 18 of these are positive. Estimate $P(\textit{positive} \mid X)$.
Worked solution. Working: Conditional proportion = $18/60 = 0.30$. Thus 30

Problem 11.5

Question. Two independent tests each have false-positive probability 0.05. Find the probability both are false positive and the probability at least one is false positive.
Worked solution. Working: Both = $0.05^2 = 0.0025$. At least one = $1 - P(\text{neither}) = 1 - (0.95)^2 = 1 - 0.9025 = 0.0975$. Independence is required.

Problem 11.6

Question. If $P(A) = 0.62$, find $P(A^c)$.
Worked solution. Working: $1 - 0.62 = 0.38$.

Problem 11.7

Question. If $P(A) = 0.7$, $P(B) = 0.5$, and events are independent, find intersection.
Worked solution. Working: $P(A \cap B) = 0.7(0.5) = 0.35$.

Problem 11.8

Question. If $P(A \cap B) = 0.12$ and $P(B) = 0.30$, calculate $P(A \mid B)$.

Worked solution. Working: $P(A \mid B) = 0.12/0.30 = 0.40$.

Problem 11.9

Question. Two mutually exclusive events have probabilities 0.18 and 0.22. Find probability neither occurs.

Worked solution. Working: Union $= 0.18 + 0.22 = 0.40$. Neither $= 1 - 0.40 = 0.60$.

Problem 11.10

Question. Explain why mutually exclusive events with positive probabilities are not independent.

Worked solution. Working: If mutually exclusive, $P(A \cap B) = 0$; independence would require $P(A \cap B) = P(A)P(B) > 0$ for positive probabilities. These conditions conflict.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.12 Binomial distribution

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 12.1

Question. For $n = 5, p = 0.2$, find $P(X = 2)$.
Worked solution. Working: $P(X = 2) = \binom{5}{2}(0.2)^2(0.8)^3 = 10(0.04)(0.512) = 0.2048$.

Problem 12.2

Question. For $n = 8, p = 0.1$, calculate the probability of no successes.
Worked solution. Working: $P(X = 0) = \binom{8}{0}(0.1)^0(0.9)^8 = (0.9)^8 = 0.4305$ approximately.

Problem 12.3

Question. A screening test has success probability 0.7 under a defined condition. For 10 independent subjects, find the expected number of successes and variance.
Worked solution. Working: Mean $= np = 10(0.7) = 7$. Variance $= np(1-p) = 10(0.7)(0.3) = 2.1$.

Problem 12.4

Question. For $n = 4, p = 0.25$, calculate $P(X \geq 1)$.
Worked solution. Working: $P(X \geq 1) = 1 - P(X = 0) = 1 - (0.75)^4 = 1 - 0.31640625 = 0.68359375$.

Problem 12.5

Question. State whether a binomial model is appropriate for 12 independent specimens with identical positive probability 0.08, and calculate $P(X = 1)$.
Worked solution. Working: It is appropriate if each specimen has two outcomes, trials are independent, n is fixed and p is constant. $P(X = 1) = \binom{12}{1}(0.08)(0.92)^{11} \approx 0.383$ (approximately).

Problem 12.6

Question. For $n = 6, p = 0.3$, calculate $P(X = 0)$.
Worked solution. Working: $P(0) = (1 - 0.3)^6 = 0.7^6 = 0.117649$.

Problem 12.7

Question. For $n = 5, p = 0.4$, calculate $P(X = 1)$.
Worked solution. Working: $\binom{5}{1}(0.4)(0.6)^4 = 5(0.4)(0.1296) = 0.2592$.

Problem 12.8

Question. For $n = 12, p = 0.25$, calculate mean and SD.

Worked solution. Working: Mean $= np = 3$. Variance $= 12(0.25)(0.75) = 2.25$; SD $= \sqrt{2.25} = 1.5$.

Problem 12.9

Question. For $n = 3, p = 0.1$, calculate $P(X \geq 1)$.

Worked solution. Working: $1 - P(0) = 1 - 0.9^3 = 1 - 0.729 = 0.271$.

Problem 12.10

Question. A binomial model has $n = 10, p = 0.2$. Find $P(X = 2)$.

Worked solution. Working: $\binom{10}{2}(0.2)^2(0.8)^8 = 45(0.04)(0.16777216) \approx 0.30199$.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.13 Poisson distribution

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 13.1

Question. For a Poisson variable with $\lambda = 2.5$, calculate $P(X = 0)$.
Worked solution. Working: $P(0) = e^{-2.5}(2.5)^0/0! = e^{-2.5} \approx 0.0821$.

Problem 13.2

Question. For $\lambda = 3$, calculate $P(X = 1)$.
Worked solution. Working: $P(1) = e^{-3}3^1/1! = 3e^{-3} \approx 0.1494$.

Problem 13.3

Question. A process averages 4 events per hour. Under a Poisson model, find the expected count in 3 hours.
Worked solution. Working: Expected count scales with time: $\lambda = 4(3) = 12$ events in 3 hours.

Problem 13.4

Question. For $\lambda = 2$, calculate $P(X = 2)$.
Worked solution. Working: $P(2) = e^{-2}2^2/2! = 2e^{-2} \approx 0.2707$.

Problem 13.5

Question. Explain two reasons a Poisson model might be unsuitable for laboratory event counts.
Worked solution. Working: It may be unsuitable if the event rate changes substantially over the interval, events cluster/depend on one another, or the variance is far from the mean due to overdispersion or underdispersion.

Problem 13.6

Question. For $\lambda = 1.5$, find $P(X = 0)$.
Worked solution. Working: $P(0) = e^{-1.5} \approx 0.2231$.

Problem 13.7

Question. For $\lambda = 4$, calculate $P(X = 2)$.
Worked solution. Working: $P(2) = e^{-4}4^2/2! = 8e^{-4} \approx 0.1465$.

Problem 13.8

Question. A clinic averages 6 arrivals per hour. Find expected arrivals over 30 minutes.
Worked solution. Working: Half an hour is 0.5 hours; expected count $= 6(0.5) = 3$.

Problem 13.9

Question. For $\lambda = 2.2$, calculate $P(X = 1)$.
Worked solution. Working: $P(1) = e^{-2.2}(2.2) \approx 0.2438$.

Problem 13.10

Question. Under a Poisson model, mean count is 5. State variance and SD.
Worked solution. Working: Variance $= \lambda = 5$. SD $= \sqrt{5} \approx 2.236$.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.14 Normal distribution and standard scores

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 14.1

Question. A measurement has $\mu = 70, \sigma = 8$. Find the z-score for $x = 82$.
Worked solution. Working: $z = (82 - 70)/8 = 1.5$.

Problem 14.2

Question. For $\mu = 100, \sigma = 15$, find the value corresponding to $z = -1.2$.
Worked solution. Working: $x = \mu + z\sigma = 100 + (-1.2)(15) = 82$.

Problem 14.3

Question. A normal model has mean 50 and SD 5. Standardise values 45 and 60.
Worked solution. Working: For 45, $z = (45 - 50)/5 = -1$. For 60, $z = (60 - 50)/5 = 2$.

Problem 14.4

Question. Explain the empirical 68–95–99.7 rule for a normal distribution.
Worked solution. Working: Approximately 68

Problem 14.5

Question. A measurement has z-score 2.1. Interpret it without making a diagnostic claim.
Worked solution. Working: It lies 2.1 standard deviations above the model mean. Whether it is unusual or clinically meaningful depends on model fit, reference population and measurement context.

Problem 14.6

Question. Find z for $x = 42$, $\mu = 50$, $\sigma = 4$.
Worked solution. Working: $z = (42 - 50)/4 = -8/4 = -2$.

Problem 14.7

Question. Find x for $z = 1.25$, $\mu = 80$, $\sigma = 6$.
Worked solution. Working: $x = 80 + 1.25(6) = 87.5$.

Problem 14.8

Question. A value is 2.5 SD below the mean 100 with SD 12. Find value.
Worked solution. Working: $x = 100 - 2.5(12) = 70$.

Problem 14.9

Question. Under a normal model, approximately what proportion lies between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$?

Worked solution. Working: Approximately 95

Problem 14.10

Question. Explain the difference between a percentile and a z-score.

Worked solution. Working: A z-score expresses distance from the mean in SD units. A percentile gives the proportion of observations below a value under the distribution.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.15 Hypothesis testing and one-sample t test

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 15.1

Question. A sample has $\bar{x} = 53, s = 8, n = 16$ and reference mean $\mu_0 = 50$. Calculate the one-sample t statistic and df.

Worked solution. Working: $SE = 8/\sqrt{16} = 2$. $t = (53 - 50)/2 = 1.5$. Degrees of freedom $= 16 - 1 = 15$.

Problem 15.2

Question. A sample has $\bar{x} = 104, s = 10, n = 25$ and $\mu_0 = 100$. State two-sided hypotheses and calculate t.

Worked solution. Working: $H_0 : \mu = 100$; $H_1 : \mu \neq 100$. $SE = 10/5 = 2$. $t = (104-100)/2 = 2$, $df = 24$.

Problem 15.3

Question. A test gives $p = 0.018$ with preselected $\alpha = 0.05$. State the decision and a correct interpretation.

Worked solution. Working: Since $0.018 < 0.05$, reject H_0 under the stated test. The data provide evidence against the null under model assumptions; the p-value is not the probability that H_0 is true.

Problem 15.4

Question. A test gives $p = 0.21$ with $\alpha = 0.05$. State the decision and what cannot be concluded.

Worked solution. Working: Fail to reject H_0. Do not conclude that the null is proved or that the groups are equivalent; precision and power should be considered.

Problem 15.5

Question. A paired study measures each of 12 specimens before and after treatment. Which values should a paired t test analyse?

Worked solution. Working: Calculate each specimen's difference (after minus before) and test whether the population mean difference is zero. The observations are paired; treating them as independent discards the pairing.

Problem 15.6

Question. For $\bar{x} = 28, s = 6, n = 9, \mu_0 = 25$, calculate t and df.

Worked solution. Working: $SE = 6/3 = 2$. $t = (28 - 25)/2 = 1.5$, $df = 8$.

Problem 15.7

Question. State hypotheses for testing whether the mean exceeds 12.

Worked solution. Working: $H_0 : \mu \leq 12$ (often tested at boundary $\mu = 12$); $H_1 : \mu > 12$. This is a right-tailed test.

Problem 15.8

Question. A confidence interval for a mean difference is -1.2 to 4.8. Does it exclude zero?

Worked solution. Working: No. Since zero lies within the interval, the data are compatible with no difference at the corresponding two-sided significance level, subject to the interval method.

Problem 15.9

Question. Explain Type I and Type II errors.

Worked solution. Working: Type I error rejects a true null hypothesis. Type II error fails to reject a false null hypothesis. Their probabilities depend on design, sample size, effect and significance threshold.

Problem 15.10

Question. A sample has $n = 49$ and SD 14. Calculate standard error.

Worked solution. Working: $SE = 14/\sqrt{49} = 14/7 = 2$.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.16 ANOVA and comparison of means

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 16.1

Question. There are $k = 5$ treatment groups and $N = 45$ observations. Calculate between-, within- and total degrees of freedom.

Worked solution. Working: Between $df = k - 1 = 4$. Within $df = N - k = 40$. Total $df = N - 1 = 44$; $4 + 40 = 44$.

Problem 16.2

Question. If $SS_B = 72$ across 4 groups and $SS_W = 120$ with 30 within-group df, calculate mean squares and F.

Worked solution. Working: $MS_B = 72/4 = 18$. $MS_W = 120/30 = 4$. $F = 18/4 = 4.5$. Compare with the appropriate F distribution for a formal p-value.

Problem 16.3

Question. Three group means are 10, 12 and 16 with equal $n = 5$. Calculate the grand mean.

Worked solution. Working: Grand mean $= [5(10) + 5(12) + 5(16)]/15 = (50 + 60 + 80)/15 = 190/15 = 12.667$.

Problem 16.4

Question. State the null hypothesis for comparing four independent laboratory methods and interpret a significant omnibus test.

Worked solution. Working: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. A significant omnibus test suggests at least one mean differs, but it does not identify which groups; suitable post-hoc comparisons are needed.

Problem 16.5

Question. List the core assumptions of one-way ANOVA and one possible remedy when variances differ substantially.

Worked solution. Working: Independence, approximate normality of residuals within groups, and homogeneity of variances are standard assumptions. If variances differ, consider Welch's ANOVA or an appropriate model, based on the design.

Problem 16.6

Question. For $k = 3, N = 36$, calculate degrees of freedom.

Worked solution. Working: Between $df = 2$; within $df = 36 - 3 = 33$; total $df = 35$.

Problem 16.7

Question. Given $SS_B = 48$, $df_B = 3$, $SS_W = 80$, $df_W = 20$, calculate F.

Worked solution. Working: $MS_B = 48/3 = 16$; $MS_W = 80/20 = 4$; $F = 16/4 = 4$.

Problem 16.8

Question. Four equal groups each have mean 20, 22, 24 and 26 with $n = 10$ each. Find grand mean.

Worked solution. Working: Equal sizes mean grand mean = $(20+22+24+26)/4 = 92/4 = 23$.

Problem 16.9

Question. Why does repeated measurement of the same sample not automatically satisfy ANOVA independence?

Worked solution. Working: Repeated values from the same unit are correlated. Use a repeated-measures or mixed model, or summarise at the independent experimental-unit level when justified.

Problem 16.10

Question. A significant ANOVA has $p = 0.01$. If $\alpha = 0.05$, state conclusion and next step.

Worked solution. Working: Reject equal-means null; evidence suggests at least one mean differs. Use planned contrasts or suitable post-hoc comparisons to locate differences while controlling multiplicity.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.17 Correlation analysis

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 17.1

Question. Interpret $r = 0.78$ between two quantitative laboratory variables.

Worked solution. Working: This indicates a fairly strong positive linear association in the sample. It does not prove causation or establish agreement between methods.

Problem 17.2

Question. Interpret $r = -0.62$ and explain the sign.

Worked solution. Working: There is a moderate-to-strong negative linear association: larger values of one variable tend to accompany smaller values of the other. The negative sign describes direction.

Problem 17.3

Question. A dataset has $r = 0.02$. What can and cannot be concluded?

Worked solution. Working: There is little observed linear association. This does not rule out nonlinear association, outliers or restricted range; inspect the scatter plot.

Problem 17.4

Question. Why is correlation not the same as agreement when comparing two assays?

Worked solution. Working: Two assays can correlate strongly because both increase across samples yet differ systematically in scale or bias. Agreement requires methods such as difference plots and analysis of bias/limits of agreement.

Problem 17.5

Question. A study reports correlation between exercise and lower blood pressure. Name two confounders and explain the limitation.

Worked solution. Working: Age and medication use could affect both exercise patterns and blood pressure. An observed association may partly reflect confounding; causal claims require design and analysis that address alternative explanations.

Problem 17.6

Question. Interpret $r = 0.91$ and state one caution.

Worked solution. Working: Strong positive linear association. Outliers, restricted range and confounding may influence it; it is not proof of causation.

Problem 17.7

Question. Interpret $r = -0.15$.

Worked solution. Working: Weak negative linear association in the sample. Its uncertainty and practical importance require sample size, confidence interval and context.

Problem 17.8

Question. If $r = 0$, what does this imply about linear association?

Worked solution. Working: Pearson linear correlation is zero in the sample. It does not imply independence or absence of nonlinear association.

Problem 17.9

Question. A scatterplot shows a curved U-shaped relationship but r is near zero. Explain.

Worked solution. Working: Positive and negative linear trends across the curve can cancel in Pearson's coefficient. A scatterplot is essential to detect nonlinear structure.

Problem 17.10

Question. Why does correlation have no measurement units?

Worked solution. Working: It standardises covariance by the product of the variables' SDs, so units cancel. Its value remains bounded between -1 and +1.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.18 Simple linear regression

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 18.1

Question. For $\hat{y} = 8 + 1.6x$, predict y at $x = 12$ and interpret the slope.

Worked solution. Working: $\hat{y} = 8 + 1.6(12) = 27.2$. Each one-unit increase in x is associated with a 1.6-unit increase in predicted mean y, within the model's applicable range.

Problem 18.2

Question. Given $\bar{x} = 6, \bar{y} = 20$ and slope $b = 2.5$, calculate intercept and write the fitted equation.

Worked solution. Working: $a = \bar{y} - b\bar{x} = 20 - 2.5(6) = 5$. Thus $\hat{y} = 5 + 2.5x$.

Problem 18.3

Question. Given $\sum(x - \bar{x})(y - \bar{y}) = 84$ and $\sum(x - \bar{x})^2 = 28$, calculate slope.

Worked solution. Working: $b = 84/28 = 3$.

Problem 18.4

Question. A model has intercept 15 and slope -0.4. Predict at $x = 20$ and interpret.

Worked solution. Working: $\hat{y} = 15 - 0.4(20) = 7$. The predicted outcome falls by 0.4 units per one-unit increase in predictor, on average, under the fitted linear model.

Problem 18.5

Question. Name three regression diagnostics or limitations to consider before reporting predictions.

Worked solution. Working: Examine residual pattern/linearity, constant residual variance, influential observations, residual distribution for inference, uncertainty intervals and extrapolation beyond the observed predictor range.

Problem 18.6

Question. For $\hat{y} = 4 + 3.2x$, predict at $x = 7$.

Worked solution. Working: $\hat{y} = 4 + 3.2(7) = 26.4$.

Problem 18.7

Question. Given $\bar{x} = 10, \bar{y} = 42, b = 1.8$, calculate intercept.

Worked solution. Working: $a = 42 - 1.8(10) = 24$.

Problem 18.8

Question. Given covariance numerator 56 and $S_{xx} = 14$, calculate slope.
Worked solution. Working: $b = 56/14 = 4$.

Problem 18.9

Question. A residual is observed $y = 31$ minus predicted $\hat{y} = 28$. Calculate residual.
Worked solution. Working: $e = y - \hat{y} = 31 - 28 = 3$, so the observation is 3 units above the fitted value.

Problem 18.10

Question. Why should a regression line not be used automatically outside the observed x-range?
Worked solution. Working: Extrapolation assumes the same functional relationship continues beyond the data. This may fail and prediction uncertainty can be large.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

A.19 Experimental design and applications

Attempt each question independently before studying the solution. Show the formula, substitution, arithmetic and interpretation.

Problem 19.1

Question. Design a comparison of two reagents using 40 independent specimens. Identify experimental units, treatment, outcome and randomisation.

Worked solution. Working: Experimental units are the 40 specimens if reagent is assigned at specimen level. Treatment is reagent A or B. Outcome is the measured analyte concentration or method difference. Randomly allocate specimens to reagents, with balanced allocation if feasible.

Problem 19.2

Question. A researcher measures one specimen 20 times. Explain the replication issue.

Worked solution. Working: The 20 readings are technical replicates from one biological unit. They estimate repeatability for that specimen but do not substitute for independent specimens when generalising across patients or samples.

Problem 19.3

Question. Explain how blocking by instrument batch could improve a reagent comparison.

Worked solution. Working: Batch is a known source of variation. Group specimens by batch and randomise reagent within each block; this can reduce residual variation and prevent treatment from being confounded with batch.

Problem 19.4

Question. A clinical trial has unequal dropout across arms. Name two risks and two reporting actions.

Worked solution. Working: Differential dropout can cause attrition bias and reduce precision. Report flow and reasons by arm, compare missingness patterns, predefine handling of missing data and conduct suitable sensitivity analyses.

Problem 19.5

Question. Distinguish randomisation, allocation concealment and blinding.

Worked solution. Working: Randomisation generates an unpredictable assignment sequence. Allocation concealment prevents recruiters from knowing the next assignment before enrolment. Blinding masks assignment after allocation from participants, clinicians, assessors or analysts where feasible.

Problem 19.6

Question. A study randomises 80 specimens equally to two methods. How many per group, and what is one benefit?

Worked solution. Working: $80/2 = 40$ per group. Randomisation reduces systematic allocation differences in expectation.

Problem 19.7

Question. A trial uses no control group and outcomes improve after treatment. Give two alternative explanations.

Worked solution. Working: Improvement could reflect natural recovery, regression to the mean, concurrent care, measurement drift or time trends. A comparison group helps assess these alternatives.

Problem 19.8

Question. Explain why the experimental unit must be identified before choosing sample size.

Worked solution. Working: Sample size must count independent units receiving treatment. Counting repeated readings as independent inflates effective sample size and can produce falsely narrow uncertainty.

Problem 19.9

Question. A laboratory compares two treatments but always runs treatment A in the morning and B in the afternoon. Identify the design flaw.

Worked solution. Working: Treatment is confounded with time of day; instrument drift or environmental changes could explain differences. Randomise or interleave run order.

Problem 19.10

Question. Name two ethical considerations in clinical research design.

Worked solution. Working: Obtain appropriate ethics approval and informed consent where required; minimise risk, protect confidentiality, justify sample size and ensure fair participant selection.

Class Work: For each problem above, write one sentence describing the assumptions or contextual information needed before using the result in a real biomedical study.

Suggested 15-Week Teaching Sequence

Week	Topic	Learning activity
1	Introduction, aims and applications	Biomedical case discussion
2	Populations, samples, parameters, statistics and variables	Classify study variables
3	Frequency distributions, tables, charts and graphs	Build frequency tables and ogives
4	Vital statistics and health indicators	Calculate rates
5	Descriptive statistics and central tendency	Assignment 1 due
6	Dispersion and coefficient of variation	Compare laboratory variability
7	Probability and introductory distributions	Mid-semester test
8	Probability distributions	Binomial, Poisson and normal exercises
9	Hypothesis testing and significance	Form hypotheses and interpret p-values
10	ANOVA and comparison of means	Assignment 2 due
11	Correlation	Scatter plots and r interpretation
12	Regression	Calculate and interpret a fitted line
13	Experimental design	Plan a laboratory experiment
14	Integrated biomedical applications	Select suitable methods
15	Revision and examination preparation	Integrated problem solving

Assessment Structure

Assessment component	Proposed weight
Assignment 1 (end of Week 5)	10%
Assignment 2 (end of Week 10)	10%
Mid-semester test (end of Week 7)	20%
Final examination	60%
Total	100%

Recommended Textbooks

The course outline recommends the following works. Consult the relevant chapters for fuller theoretical discussion and additional practice:

1. Daniel, W. W., and Cross, C. L. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley.
2. Rosner, B. *Fundamentals of Biostatistics*. Cengage Learning.
3. Pagano, M., and Gauvreau, K. *Principles of Biostatistics*. CRC Press.
4. Zar, J. H. *Biostatistical Analysis*. Pearson.
5. Motulsky, H. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. Oxford University Press.

Appendix B

Supplementary Rigorous Worked Examples from the Course Handout

This chapter integrates numerical questions from the uploaded *CALEB University Biostatistics* handout into the broader lecture notes. The original contexts and datasets are retained where practical. Some calculations are shown more fully, and obvious typographical or arithmetic inconsistencies in the handout are explicitly noted rather than repeated as correct results.

B.1 Extended tabular calculations for variance and standard deviation

For ungrouped observations, calculate the mean first, then each deviation $x - \bar{x}$, its square, and the sum of squares. Divide by n for a population variance or by $n - 1$ for a sample variance. For grouped data, use the class midpoint x and frequency f, so that the weighted sum of squared deviations is $\sum f(x - \bar{x})^2$.

Example SD1: Ungrouped observations with a full working table

Question. The turnaround times (minutes) for eight hypothetical specimens are $12, 14, 15, 18, 21, 24, 26, 30$. Find the mean, population variance, population standard deviation, sample variance and sample standard deviation.

Step 1: Calculate the mean.

$$\bar{x} = \frac{12 + 14 + 15 + 18 + 21 + 24 + 26 + 30}{8} = \frac{160}{8} = 20.$$

x	$x - \bar{x}$	$(x - \bar{x})^2$	Running total of squares
12	-8	64	64
14	-6	36	100
15	-5	25	125
18	-2	4	129
21	1	1	130
24	4	16	146
26	6	36	182
30	10	100	282
$\sum x = 160$	$\sum(x - \bar{x}) = 0$	$\sum(x - \bar{x})^2 = 282$	

Step 2: Population measures.

$$\sigma^2 = \frac{282}{8} = 35.25, \qquad \sigma = \sqrt{35.25} = 5.937.$$

Step 3: Sample measures.

$$s^2 = \frac{282}{8-1} = \frac{282}{7} = 40.286, \qquad s = \sqrt{40.286} = 6.347.$$

Interpretation. The sample standard deviation is larger because the sample variance uses $n-1$ rather than n in the denominator.

Example SD2: Grouped-data standard deviation from the uploaded handout

Question. The handout gives the following grouped weights. Complete the table and calculate the grouped mean, variance, standard deviation and coefficient of variation. Use the population-style grouped variance with denominator $\sum f$ as in the handout.

Class interval	Midpoint x	f	fx	$(x-64.1)^2$	$f(x-64.1)^2$
52–56	54	3	162	102.01	306.03
56–60	58	6	348	37.21	223.26
60–64	62	10	620	4.41	44.10
64–68	66	4	264	3.61	14.44
68–72	70	8	560	34.81	278.48
72–76	74	2	148	98.01	196.02
76–80	78	1	78	193.21	193.21
Total		34	2180		1255.54

Step 1: Mean.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2180}{34} = 64.1176 \approx 64.1.$$

Step 2: Variance and standard deviation using the handout's rounded mean.

$$\sigma^2 \approx \frac{\sum f(x-64.1)^2}{\sum f} = \frac{1255.54}{34} = 36.9276, \qquad \sigma \approx \sqrt{36.9276} = 6.0768.$$

Step 3: Coefficient of variation.

$$CV = \frac{6.0768}{64.1} \times 100\% = 9.48\%.$$

Interpretation. The grouped standard deviation is approximately 6.08 weight units, and the coefficient of variation is approximately 9.48%.

Accuracy note. The handout rounds the mean to 64.1 before calculating deviations. If the unrounded mean 2180/34 is used throughout, the sum of weighted squared deviations and the final variance will differ slightly. Grouped-data results are estimates because each class is represented by its midpoint.

Example SD3: Grouped data using the computational table

Question. A laboratory records the following grouped concentration measurements. Use a tabular method to estimate the mean and population standard deviation.

Class	Midpoint x	f	fx	$f(x-\bar{x})^2$
10–14	12	2	24	171.125
15–19	17	5	85	90.3125
20–24	22	8	176	4.5000
25–29	27	4	108	132.2500
30–34	32	1	32	115.5625
Total		20	425	513.7500

Step 1: Mean. $\bar{x} = \sum fx / \sum f = 425/20 = 21.25$. **Step 2: Weighted squared deviations.**

$$\begin{aligned}\sum f(x-\bar{x})^2 &= 2(12-21.25)^2 + 5(17-21.25)^2 \\ &\quad + 8(22-21.25)^2 + 4(27-21.25)^2 + (32-21.25)^2 \\ &= 171.125 + 90.3125 + 4.5 + 132.25 + 115.5625 = 513.75.\end{aligned}$$

$$\sigma^2 = \frac{513.75}{20} = 25.6875, \qquad \sigma = \sqrt{25.6875} = 5.068.$$

The table uses the actual mean in every squared-deviation term; this is essential for a valid variance calculation.

Example SD4: Comparing relative variability

Question. Method A has mean 64.1 and standard deviation 6.0768. Method B has mean 80 and standard deviation 10. Compare their relative variability.

$$CV_A = \frac{6.0768}{64.1} \times 100 = 9.48\%, \qquad CV_B = \frac{10}{80} \times 100 = 12.50\%.$$

Method A has lower relative variability in this comparison. This does not by itself establish accuracy or clinical suitability; bias and the measurement context must also be assessed.

Class Work:

1. Use a full table to calculate the sample standard deviation of $8, 10, 11, 15, 16$.
2. A grouped dataset has midpoints $5, 10, 15, 20$ and frequencies $3, 6, 8, 3$. Construct columns fx, $(x-\bar{x})^2$ and $f(x-\bar{x})^2$, then calculate the grouped population SD.
3. Two assays have means and SDs $(45, 3.6)$ and $(72, 4.32)$. Calculate both CVs and state which has lower relative variability.

B.2 Additional worked examples from the handout: frequency distributions and central tendency

Example FD1: Constructing a frequency distribution from experimental measurements

Question. The handout records 55 experimental measurements, with minimum 4.5 and maximum 52.0. Determine the range, estimate the number of classes using Sturges' rule, and give the frequency table reported in the handout.

$$R = 52.0 - 4.5 = 47.5, \qquad K = 1 + 3.322\log_{10}(55) \approx 1 + 3.322(1.7404) \approx 6.78 \approx 7.$$

$$w \approx \frac{R}{K} = \frac{47.5}{7} \approx 6.79 \approx 7.$$

Measurement class	Frequency
4.5–11.5	22
11.5–18.5	21
18.5–25.5	5
25.5–32.5	4
32.5–39.5	1
39.5–46.5	1
46.5–53.5	1
Total	55

The frequencies sum to 55. In a final table, define class-boundary conventions carefully so that each observation is counted once and only once.

Example FD2: Household feeding expenditure from the handout

Question. The handout records the following feeding expenditure (in thousands of naira) for 40 households in Ibadan in March 2023:

$$32, 22, 19, 18, 43, 42, 40, 43, 18, 21,$$
$$31, 26, 22, 25, 47, 40, 26, 32, 34, 22,$$
$$35, 38, 34, 41, 36, 25, 22, 45, 48, 26,$$
$$28, 38, 35, 19, 47, 28, 26, 35, 38, 35.$$

Construct seven classes and estimate the arithmetic, harmonic and geometric means from the grouped table.

Step 1: Range and class width.

$$R = 48 - 18 = 30, \qquad K = 7, \qquad w \approx \frac{30}{7} = 4.286.$$

Use a convenient class width of 5 and seven non-overlapping integer classes.

Class (N'000)	Midpoint x	Frequency f	fx	f/x	$f \ln x$
18–22	20	9	180	0.4500	26.961
23–27	25	6	150	0.2400	19.313
28–32	30	5	150	0.1667	17.006
33–37	35	7	245	0.2000	24.887
38–42	40	7	280	0.1750	25.822
43–47	45	5	225	0.1111	19.033
48–52	50	1	50	0.0200	3.912
Total		40	1280	1.3628	136.936

Step 2: Arithmetic mean.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1280}{40} = 32.00 \text{ thousand naira.}$$

Step 3: Harmonic mean.

$$H = \frac{\sum f}{\sum (f/x)} \approx \frac{40}{1.3628} \approx 29.35 \text{ thousand naira.}$$

Step 4: Geometric mean.

$$G = \exp\left(\frac{\sum f \ln x}{\sum f}\right) \approx \exp\left(\frac{136.936}{40}\right) \approx 30.67 \text{ thousand naira.}$$

Interpretation and limitation. These are grouped-data estimates based on class midpoints. They differ from calculations using all 40 raw observations because grouping replaces each observation with its class midpoint. For this dataset, the raw-data arithmetic mean is $1282/40 = 32.05$ thousand naira.

Example FD3: Standard deviation of the grouped household expenditure

Using the seven classes and midpoints in Example FD2, the grouped mean is $\bar{x} = 32$. Complete the dispersion table:

Midpoint x	f	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
20	9	−12	144	1296
25	6	−7	49	294
30	5	−2	4	20
35	7	3	9	63
40	7	8	64	448
45	5	13	169	845
50	1	18	324	324
Total	40			3290

$$\sigma^2 = \frac{3290}{40} = 82.25, \qquad \sigma = \sqrt{82.25} = 9.069 \text{ thousand naira.}$$

$$CV = \frac{9.069}{32} \times 100\% = 28.34\%.$$

The grouped SD is approximately N9,069 and the CV is approximately 28.34%; both are estimates based on the grouped midpoints.

Example CT1: Mean age from the handout

The handout lists the ages $13, 18, 18, 19, 17, 19.5, 19, 17, 18.5, 18, 17$ but describes them as ten students and sums only ten observations. The displayed list contains eleven values. This is an internal inconsistency in the source and should not be silently ignored. For the eleven values actually printed:

$$\sum x = 194, \qquad n = 11, \qquad \bar{x} = \frac{194}{11} = 17.636.$$

If the intended dataset was ten observations, the omitted or extra value must be clarified before reporting the mean. This is an important data-checking lesson.

Example CT2: Grouped mean from the handout

The handout provides class midpoints $54, 58, 62, 66, 70, 74, 78$ and frequencies $3, 6, 10, 4, 8, 2, 1$.

Midpoint x	Frequency f	fx	Cumulative frequency
54	3	162	3
58	6	348	9
62	10	620	19
66	4	264	23
70	8	560	31
74	2	148	33
78	1	78	34
Total	34	2180	

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2180}{34} = 64.1176 \approx 64.12.$$

The mean is an estimate based on class midpoints.

B.3 Additional probability-distribution examples from the handout

Example P1: Eight tosses of a fair coin

Let X be the number of heads in eight independent tosses. Then $X \sim \text{Bin}(8, 0.5)$ and

$$P(X = x) = \binom{8}{x}(0.5)^x(0.5)^{8-x} = \binom{8}{x}(0.5)^8.$$

1. **Exactly 3 heads:**
$$P(X = 3) = \binom{8}{3}(0.5)^8 = \frac{56}{256} = 0.21875.$$

2. **At least 1 head:**
$$P(X \geq 1) = 1 - P(X = 0) = 1 - \left(\frac{1}{2}\right)^8 = \frac{255}{256} = 0.99609.$$

3. **At most 2 heads:**
$$P(X \leq 2) = \frac{\binom{8}{0} + \binom{8}{1} + \binom{8}{2}}{256} = \frac{1 + 8 + 28}{256} = \frac{37}{256} = 0.14453.$$

4. **Between 2 and 4 heads inclusive:**
$$P(2 \leq X \leq 4) = \frac{\binom{8}{2} + \binom{8}{3} + \binom{8}{4}}{256} = \frac{28 + 56 + 70}{256} = \frac{154}{256} = 0.60156.$$

Example P2: Poisson model for incoming calls

The handout describes an average of 2.5 calls per minute. Let X be the number of calls in one minute and assume a Poisson model with $\lambda = 2.5$:

$$P(X = x) = \frac{e^{-2.5}(2.5)^x}{x!}.$$

Event	Working	Probability (approx.)
No calls	$e^{-2.5}$	0.0821
Exactly one call	$e^{-2.5}(2.5)$	0.2052
Exactly two calls	$e^{-2.5}(2.5)^2/2$	0.2565
At least one call	$1 - P(X = 0) = 1 - e^{-2.5}$	0.9179

These results depend on the Poisson assumptions and the stated one-minute interval.

Example P3: Normal distribution from the handout

Suppose $X \sim N(5, 4)$, where the second parameter is the variance. Thus $\mu = 5$ and $\sigma = \sqrt{4} = 2$. Standardise with $Z = (X - \mu)/\sigma$.

1. **Find $P(X \leq 7)$:**

$$z = \frac{7 - 5}{2} = 1, \qquad P(X \leq 7) = P(Z \leq 1) = 0.8413.$$

2. **Find $P(6 \leq X \leq 7)$:**

$$P(6 \leq X \leq 7) = \Phi(1) - \Phi(0.5) \approx 0.8413 - 0.6915 = 0.1498.$$

3. **Find c such that $P(X > c) = 0.025$:**

$$P(X \leq c) = 0.975, \quad z_{0.975} \approx 1.96, \quad \frac{c - 5}{2} = 1.96, \quad c = 5 + 3.92 = 8.92.$$

Example P4: Geometric waiting-time calculation

The handout states that the probability of a successful experiment is 0.60 and asks for the expected number of repetitions required to obtain five successes under a negative-binomial waiting-time model. With $r = 5$, $p = 0.60$ and $q = 0.40$:

$$E(X) = \frac{r}{p} = \frac{5}{0.60} = 8.333, \qquad \mathrm{Var}(X) = \frac{rq}{p^2} = \frac{5(0.40)}{0.60^2} = 5.556.$$

This model counts trials up to and including the fifth success; it does not require five successive successes.

B.4 University-standard integrated class work

Show formulas, substitutions, tabular calculations, units and a short interpretation for every question.

1. **Frequency distribution and mean.** Use the 40 household feeding-expenditure observations in the course handout. Construct seven non-overlapping classes, calculate class midpoints and frequencies, then estimate the arithmetic mean using $\sum fx / \sum f$. Explain why the grouped estimate may differ from the raw-data mean.

2. **Median and mode.** For the grouped distribution in Question 1, construct cumulative frequencies. Identify the median class and modal class, then use the appropriate grouped-data formulas to estimate the median and mode. State the class-boundary convention used.

3. **Standard deviation.** For Question 1, add columns fx, $(x - \bar{x})^2$ and $f(x - \bar{x})^2$. Calculate the grouped variance, SD and CV. Repeat using the sample-variance denominator $\sum f - 1$ and compare the two SDs.

4. **Binomial distribution.** A screening test has probability $p = 0.08$ of a positive result in each of 12 independent individuals. Calculate the probability of exactly two positives, at most one positive, and at least one positive. State the assumptions.

5. **Poisson distribution.** A laboratory receives an average of 3.2 urgent samples per hour. Calculate the probabilities of zero, exactly two, and at least four urgent samples in one hour.

6. **Normal distribution.** A measurement is normally distributed with mean 100 and standard deviation 12. Calculate $P(X < 88)$, $P(94 < X < 118)$ and the 90th percentile. Show each z-score.

7. **Hypothesis test.** A sample of 16 measurements has mean 52.4 and SD 4.8. Test $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$ at $\alpha = 0.05$. Calculate t, degrees of freedom and state how the critical value or p-value determines the decision.

8. **ANOVA.** Three independent methods produce results A: 8, 9, 10; B: 10, 11, 12; C: 13, 14, 15. Calculate group means, grand mean, SS_B, SS_W, degrees of freedom, mean squares and F.

9. **Correlation.** For paired observations $(1, 2), (2, 3), (3, 5), (4, 4), (5, 7)$, construct a table of deviations and cross-products and calculate Pearson's r. Interpret direction and strength without making a causal claim.

10. **Regression.** Using the paired observations in Question 9, calculate the least-squares slope and intercept, write the fitted line, and predict Y at $X = 6$. Explain why this prediction may be less reliable than interpolation.

11. **Experimental design.** Design a study comparing two laboratory methods with 48 independent specimens. Specify allocation, replication, control/comparison procedure, blinding where feasible, primary outcome and the analysis that matches the design.

Source note. The frequency-distribution, household expenditure, grouped-weight standard-deviation, coin-toss, Poisson-call and normal-distribution examples in this supplementary chapter are based on the uploaded *CALEB University Biostatistics* course handout. Expanded workings and clearly labelled corrections are provided for teaching accuracy.

Good Practice Checklist

Before reporting a biomedical statistical result, check that:

- the research question and target population are clearly stated;
- variables and units are defined;
- the sampling and study design are described;
- the denominator and period for each rate are explicit;
- the chosen summary or test matches the data and design;
- assumptions are assessed and limitations reported;
- estimates are presented with units and, where appropriate, confidence intervals;
- statistical significance is not confused with clinical importance;
- correlation is not described as proof of causation;
- privacy and ethical handling of biomedical data are respected.