BP701T · B.Pharm Semester VII · Theory Notes

Biostatistics and Research Methodology

Unit I — Basics of Biostatistics and Descriptive Statistics (9 hours)

Unit I (9 hours)

Pills & Python Notes

From-zero teaching standard: every technical term is defined before use.

Syllabus coverage: Unit I (9 hours).

PILLS & PYTHON

Pharmacy • Data • AI • Healthcare

Instagram @pillsandpython · pillsandpython.com · YouTube @PillsAndPython

CHAPTER 1: Basic Concepts of Biostatistics — Variables, Data, Collection and Preprocessing

1. Learning Objectives

After studying this chapter, you will be able to:

- Define biostatistics and explain why it is needed in pharmacy
- Explain what a variable is and classify variables (qualitative vs quantitative; discrete vs continuous)
- Define data and describe methods of data collection
- Explain the steps of data preprocessing and data cleaning
- Identify common errors in data and describe how to correct them

2. Definitions

Biostatistics: The science of collecting, organising, summarising, presenting and analysing data that relate to living things (humans, animals, plants, medicines, microbes) and drawing conclusions from it.

Variable: Anything that can change or take different values from one person, tablet, patient or sample to another. Example: blood pressure, tablet weight, age.

Data: The actual values collected for the variables. Example: 100 tablet weights written down in a notebook.

Data collection: The planned process of gathering information (numbers, answers, measurements) from people, samples or records.

Data preprocessing and cleaning: The steps of checking the collected data for mistakes, filling or removing missing values, and arranging it in a usable form before analysis.

3. Terminology Table

Term	Simple meaning
Biostatistics	Statistics applied to biology, medicine and pharmacy
Variable	Something that varies and can be measured or counted
Data	Collected values of variables
Observation	One single collected value
Qualitative variable	Variable described in words (e.g., blood group)
Quantitative variable	Variable described in numbers (e.g., weight in mg)
Discrete variable	Whole-number variable (e.g., number of tablets)
Continuous variable	Variable that can take any decimal value (e.g., height)
Primary data	Data you collect yourself
Secondary data	Data collected by someone else

Term	Simple meaning
Raw data	Data exactly as collected, not yet cleaned or arranged
Outlier	A value that is far away from all other values
Missing value	A value that was not recorded

4. Simple Explanations

What is biostatistics? (start from zero)

Imagine a pharmacist wants to know: "Is my new cough syrup really working?" She gives it to 100 patients and records who got better and who did not. Those numbers, by themselves, are just a list. **Biostatistics** is the tool that turns that list into an answer — "Yes, it works for about 80 out of 100 patients."

So biostatistics has five jobs: (1) **collect** the data properly, (2) **organise** it, (3) **summarise** it (averages, percentages), (4) **present** it (tables, charts), and (5) **analyse** it to reach a conclusion. The "bio" part only means the data comes from living things — patients, tablets, bacteria, animals.

Example: A quality-control chemist weighs 20 tablets of paracetamol. The weights are the data; the average weight tells her whether the batch is uniform. That is biostatistics at work.

What is a variable?

A **variable** is simply anything that can take different values. A fixed thing is not a variable — a **constant** is. The word "variable" comes from "vary" = to change.

- The weight of a tablet is a variable, because different tablets weigh slightly different amounts (498 mg, 501 mg, 499 mg...).
- The number "1 tablet in one blister" is a constant if every blister has exactly one tablet.

Example: In a study of blood pressure, *blood pressure* is the variable (120, 130, 125 mmHg...). The data are the actual readings written in the register.

What is data?

Data is the plural of "datum" (one observation). If a variable is the *question* ("What is the tablet weight?"), then the data are the *answers* (498 mg, 501 mg, 499 mg...). Raw data is just the pile of answers before any cleaning or counting.

Example: A nurse records the fasting blood sugar of 10 patients: 95, 102, 110, 98, 130, 105, 99, 115, 108, 101 mg/dL. That list of ten numbers is the data.

5. Classifications

5.1 Types of variables

TECHNICAL CONCEPT: Variables are classified as qualitative or quantitative.

SIMPLE EXPLANATION: Ask one question — "Can I describe this variable in numbers?" If yes, it is quantitative (quantity = number). If it is described in words or categories, it is qualitative (quality = kind).

- **Qualitative variables (categorical):** described in words/categories. Example: blood group (A, B, AB, O), gender (male, female), tablet defect (yes/no), pain level described as mild/moderate/severe.

- **Quantitative variables (numerical):** described in numbers. Example: weight (mg), age (years), number of tablets, blood pressure (mmHg).

EXAMPLE:

```
Patient No.   Blood group (qualitative)   Fasting sugar mg/dL (quantitative)
1             A                           95
2             O                           110
3             B                           98
```

KEY TAKEAWAY: Numbers → quantitative; words/categories → qualitative.

5.2 Discrete vs continuous (sub-types of quantitative variables)

TECHNICAL CONCEPT: Quantitative variables are either discrete or continuous.

SIMPLE EXPLANATION: Count the whole numbers only, or can you have decimals too?

- **Discrete:** Only whole numbers are possible — you count them. Example: number of tablets in a strip (10, 15 — never 10.5), number of patients in a ward.
- **Continuous:** Any value, including decimals, is possible — you measure them. Example: height (165.5 cm), tablet weight (498.3 mg), temperature (37.2 °C).

EXAMPLE: Number of side effects reported by a patient (0, 1, 2) → discrete. Patient's body weight (58.5 kg) → continuous.

KEY TAKEAWAY: Counted = discrete; measured (with decimals) = continuous.

5.3 Types of data (by source)

- **Primary data:** You collect it yourself, first-hand. Example: you interview 50 patients about their symptoms.
- **Secondary data:** Collected by someone else, and you use it. Example: hospital admission records, WHO reports, published research papers.

5.4 Methods of data collection

1. **Direct observation:** Watch and record. Example: counting how many patients visit the OPD each hour.
2. **Interview:** Ask questions face-to-face or on the phone. Example: asking patients about side effects of a medicine.
3. **Questionnaire:** A written set of questions filled by the respondent. Example: a survey form given to 200 pharmacy students.
4. **Experiment:** Controlled measurements in a laboratory. Example: recording dissolution % of tablets at different times.
5. **Records/registers:** Using already-recorded information. Example: collecting last year's prescription data from the hospital pharmacy.

6. Processes

6.1 Steps of data collection

Step 1 — Define the objective. Decide clearly what question you want to answer. Example: "What is the average weight of paracetamol tablets in batch B-12?"

Step 2 — Decide the variable(s). List exactly what you will measure. Example: weight of each tablet in mg.

Step 3 — Choose the method. Observation, interview, questionnaire, experiment or records — whichever fits the objective.

Step 4 — Prepare the tool. Make the questionnaire, calibration sheet, or recording format. Example: a table with columns "Tablet No." and "Weight (mg)".

Step 5 — Collect the data. Carry out the measurements or interviews carefully and honestly, without skipping.

Step 6 — Check and store. Review the filled sheets the same day; store safely (register, spreadsheet).

6.2 Steps of data preprocessing and cleaning

Step 1 — Enter and arrange. Transfer data into a table or spreadsheet, one row per observation, one column per variable.

Step 2 — Check for missing values. Look for blank cells. Decide: fill them (only if sensible), or leave out that observation.

Step 3 — Check for wrong entries. Impossible values must be corrected or removed. Example: a tablet weight entered as "49 mg" when others are ~500 mg is likely a typing error for "499 mg".

Step 4 — Check for outliers. Values far from the rest may be errors or genuine extreme values — investigate before deleting.

Step 5 — Check formats and units. All weights must be in the same unit (mg, not a mix of mg and g); dates in the same format.

Step 6 — Coding and recoding. Convert words into codes for analysis. Example: male = 1, female = 2; yes = 1, no = 0.

Step 7 — Final verification. Recheck totals and a few random entries against the original records.

7. Textual Diagrams / Flowcharts

Flowchart: From question to conclusion

```
Research question
      |
What to measure?  (identify variables)
      |
How to measure?   (choose collection method)
      |
Collect the data  (field/lab work)
      |
Clean the data    (missing values, errors, formats)
      |
Organise + summarise  (tables, averages)
      |
Analyse + conclude
```

Diagram: Variable classification tree

```
                VARIABLE
              (anything that varies)
              /            \
      Qualitative           Quantitative
    (words/categories)      (numbers)
       |                   /          \
  e.g. blood group     Discrete       Continuous
  e.g. gender          (whole numbers  (any value,
                        only)          decimals ok)
                        |                   |
                  e.g. no. of tablets   e.g. tablet weight
                  e.g. no. of patients  e.g. blood pressure
```

8. > INFOGRAPHIC CONCEPT

INFOGRAPHIC CONCEPT: A one-page visual titled "The Data Journey of a Tablet Batch". Left panel shows a balance and a QC chemist with the caption "COLLECT — weigh 20 tablets (primary data, experiment method)". Middle panel shows a checklist with magnifying glass: "CLEAN — find missing entries, fix typing errors, keep one unit (mg)". Right panel shows a neat table and a bar chart: "ORGANISE + SUMMARISE — frequency table and average weight". A winding arrow connects the three panels, and each variable type appears as a labelled tag: weight in mg (quantitative, continuous), batch number (qualitative), number of defective tablets (quantitative, discrete).

9. Tables

Table 1: Primary vs secondary data

Feature	Primary data	Secondary data
Collected by	The researcher herself/himself	Someone else
Time and cost	More time and cost	Less time and cost
Suitability	Exactly fits the research question	May not fit perfectly
Example	Interviewing patients about side effects	Hospital pharmacy issue registers

Table 2: Data collection methods at a glance

Method	What happens	Pharmacy example
Observation	Watch and record	Counting patients at the counter per hour
Interview	Ask questions directly	Asking patients if they completed the antibiotic course
Questionnaire	Written questions, self-filled	Survey on self-medication among students
Experiment	Controlled measurement	Dissolution test of tablets at 15, 30, 45 min
Records	Use existing records	Using last year's ADR reports

10. Worked Examples and Calculation Boxes

Worked example 1: Identifying variable types

A hospital study records for each patient: (a) age in years, (b) blood group, (c) number of medicines prescribed, (d) body temperature in °C.

- (a) Age in years → quantitative, continuous (can be 45.5 years)
- (b) Blood group → qualitative (A, B, AB, O)
- (c) Number of medicines → quantitative, discrete (1, 2, 3... whole numbers)
- (d) Body temperature → quantitative, continuous (37.2 °C)

CALCULATION BOX 1: Simple percentage summary

Problem: Out of 200 patients given a medicine, 160 showed improvement. What percentage improved?

Solution:

- Step 1: Write the fraction: improved ÷ total = 160 ÷ 200
- Step 2: Divide: 160 ÷ 200 = 0.8
- Step 3: Multiply by 100 to get percentage: 0.8 × 100 = 80

Final answer: 80% of patients improved.

CALCULATION BOX 2: Spotting and correcting a data error

Problem: Ten tablet weights (mg) were recorded as: 500, 502, 499, 501, 50, 500, 503, 498, 501, 500. Identify the suspicious value and suggest the correction.

Solution:

- Step 1: List all values and look at the pattern: nine values are between 498 and 503.
- Step 2: One value, 50, is far below the rest — it is an outlier and almost certainly a typing error (the "0" of 500 was missed).
- Step 3: Check the original record; if it reads "500", correct the entry to 500.

Final answer: The value 50 is the error; correct it to 500 after verification.

11. Analogies

- **Biostatistics is like a pharmacist's balance.** Raw data is like a heap of powder on the pan — useless until weighed (summarised) properly.
- **Variables are like medicine labels.** Just as every label has a name, strength and batch number, every variable has a name, type and unit.
- **Data cleaning is like filtering a syrup.** You remove dust and fibres (errors, missing values) so the final product is clear and safe to use.
- **Discrete vs continuous is like tablets vs syrup.** Tablets are counted whole (discrete); syrup is measured and can be 5.5 mL (continuous).

12. Pharmacy Relevance

- **Quality control:** Tablet weight variation, hardness, friability and dissolution data are collected and analysed batch-wise — pure biostatistics.
- **Clinical pharmacy:** Patient blood pressure readings before and after treatment are data that decide whether a drug works.
- **Pharmacovigilance:** ADR (adverse drug reaction) reports are collected through forms, cleaned, coded and analysed to detect drug safety signals.
- **Community pharmacy:** Prescription audit data (number of antibiotics prescribed per 100 prescriptions) helps monitor rational drug use.

13. India Context

- The **National Family Health Survey (NFHS)** and **Census of India** are large examples of secondary data collected by government agencies and used by pharmacy and public-health researchers.
- **Pharmacovigilance Programme of India (PvPI)** collects ADR data from hospitals across India through standardised reporting forms — an example of primary data collected by interview/record methods.
- ICMR and state health departments routinely publish disease statistics (for example, TB and diabetes numbers) that students can use as secondary data for projects.

14. Key Points

1 Biostatistics = collection, organisation, summary, presentation and analysis of data from living things.

2 A variable is anything that can take different values; data are the actual collected values.

3 Qualitative variables are described in words; quantitative variables in numbers.

4 Discrete variables take whole numbers only; continuous variables can take decimal values.

5 Primary data is collected first-hand; secondary data is taken from other sources.

6 Data collection methods: observation, interview, questionnaire, experiment, records.

7 Data cleaning steps: arrange, find missing values, correct wrong entries, check outliers, unify formats and units, code categories, verify.

15. Common Confusions

- **Variable vs data:** The variable is the *thing being measured* (tablet weight); the data are the *actual measurements* (498 mg, 501 mg...). Students often use the two words as if they are the same.
- **Discrete vs continuous:** "Number of tablets" is discrete, but "weight of tablets" is continuous — the same object can give both types depending on what is measured.
- **Primary vs secondary:** Primary does not mean "better"; it means "collected by you". Secondary data like government health surveys can be perfectly reliable.
- **Cleaning vs deleting:** Cleaning does not mean deleting everything unusual. An outlier must be investigated — it may be a genuine extreme value.

16. Common Mistakes

1 Writing data with mixed units (some weights in mg, some in g) and analysing them together.

2 Deleting outliers without checking whether they are typing errors or real values.

3 Leaving blank cells without noting them, so the analysis silently uses fewer observations.

4 Coding categories inconsistently (male = 1 in one file, male = 0 in another).

5 Collecting data without a clear objective, then not knowing what to analyse.

17. Mnemonics

- **"CO-OPS-PA"** for the five jobs of biostatistics: **C**ollect, **O**rganise, **O** (summarise → "O" for overview), **P**resent, **A**nalyse. (Say: "Co-ops pa!")
- **"Q-words, Q-numbers"**: **Q**ualitative = words, **Q**uantitative = numbers.
- **"Count it = Discrete; Measure it = Continuous"** — tablets are counted, syrup is measured.
- Cleaning steps: **"A M W O F C V"** — **A**rrange, **M**issing values, **W**rong entries, **O**utliers, **F**ormats, **C**ode, **V**erify ("A Memorable Work Of Fine Careful Verification").

18. Exam Summary

Biostatistics is the science of collecting, organising, summarising, presenting and analysing data relating to living things. A **variable** is anything that can take different values; **data** are the collected values of variables. Variables are **qualitative** (words/categories, e.g., blood group) or **quantitative** (numbers). Quantitative variables are **discrete** (whole numbers only, e.g., number of tablets) or **continuous** (any value including decimals, e.g., weight). Data are **primary** (collected first-hand) or **secondary** (from other sources). Methods of collection: observation, interview, questionnaire, experiment and records. **Data preprocessing/cleaning** means arranging data, finding missing values, correcting wrong entries, checking outliers, unifying formats and units, coding categories and verifying against originals.

19. One-Minute Revision

- Biostatistics: collect → organise → summarise → present → analyse data from living things.
- Variable = what varies; Data = the collected values.
- Qualitative = words; Quantitative = numbers.
- Discrete = whole numbers (counted); Continuous = decimals allowed (measured).
- Primary = self-collected; Secondary = others' data.
- Collection: observation, interview, questionnaire, experiment, records.
- Cleaning: arrange → missing → wrong entries → outliers → formats → code → verify.

20. Questions

Short-answer questions (2–3 marks)

1 Define biostatistics in your own words.

2 What is the difference between a variable and data? Give one pharmacy example.

3 Classify: (a) number of capsules in a bottle, (b) dissolution % of a tablet, (c) patient's blood group.

4 What is primary data? Give two examples from pharmacy practice.

5 List the steps of data cleaning in order.

Long-answer questions (5–10 marks)

1 Define biostatistics and explain its five functions with a pharmacy example for each.

2 Explain the classification of variables with suitable pharmaceutical examples.

3 Describe the various methods of data collection, with merits and limitations of each.

4 What is data preprocessing and cleaning? Explain each step in detail.

MCQs (10, with answer key)

Q1. Biostatistics is the science of: a) Making medicines b) Collecting, organising, summarising, presenting and analysing data from living things c) Only calculating averages d) Writing prescriptions **Answer: b**

Q2. Which of the following is a qualitative variable? a) Tablet weight b) Blood group c) Number of patients d) Temperature **Answer: b**

Q3. The number of tablets in a strip is an example of: a) Continuous variable b) Discrete variable c) Constant d) Qualitative variable **Answer: b**

Q4. Data collected by the researcher herself for the first time is called: a) Secondary data b) Primary data c) Raw constant d) Published data **Answer: b**

Q5. A questionnaire filled in by the patient herself is an example of which collection method? a) Observation b) Interview c) Questionnaire d) Experiment **Answer: c**

Q6. A value far away from all other values in a dataset is called: a) Constant b) Outlier c) Variable d) Datum **Answer: b**

Q7. Dissolution percentage of a tablet (e.g., 85.5%) is: a) Qualitative b) Discrete quantitative c) Continuous quantitative d) A constant **Answer: c**

Q8. Which step comes FIRST in data cleaning? a) Coding categories b) Arranging data in a table c) Checking outliers d) Deleting all unusual values **Answer: b**

Q9. Hospital admission registers used by a student for a project are: a) Primary data b) Secondary data c) Experimental data d) Qualitative constants **Answer: b**

Q10. Converting "male/female" into 1/2 for analysis is called: a) Observation b) Coding c) Experimentation d) Sampling **Answer: b**

Viva questions

1 What does the "bio" in biostatistics mean?

2 Give one example each of discrete and continuous variables from a pharmacy.

3 Why should all data be in the same unit before analysis?

4 What will you do if you find a blank cell in your data sheet?

5 Name three methods of data collection.

Application-based questions

1 A QC officer records the hardness of 10 tablets (kg/cm^2): 4.5, 4.8, 4.6, 45, 4.7, 4.5, 4.9, 4.6, 4.8, 4.7. Identify the data problem and explain how you would handle it as part of data cleaning.

2 You want to study self-medication among college students in your city. Which data collection method would you choose and why? Write any four questions you would include.

3 A researcher downloads last year's ADR reports from a government website for her study. Is this primary or secondary data? Justify.

21. References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)

CHAPTER 2: Population, Sample and Sampling Methods

1. Learning Objectives

After studying this chapter, you will be able to:

- Define population and sample and explain their relationship
- Explain why sampling is important in pharmacy and research
- List and explain the four probability sampling methods: random, systematic, stratified, cluster
- List and explain the three non-probability sampling methods: convenience, purposive, snowball
- Choose a suitable sampling method for a given pharmacy research situation

2. Definitions

Population: The entire group of individuals, items or units about which we want to draw a conclusion. Example: all paracetamol tablets manufactured in one year.

Sample: A smaller part of the population that is actually studied. Example: 100 tablets taken from that year's production.

Sampling: The process of selecting a sample from the population.

Sampling frame: A complete list of all members of the population from which the sample is drawn. Example: a register of all registered pharmacies in a district.

Sampling unit: One single member of the population. Example: one tablet, one patient, one pharmacy.

Probability sampling: A sampling method in which every member of the population has a known, non-zero chance of being selected.

Non-probability sampling: A sampling method in which members are selected by convenience or judgement, and not every member has a known chance of selection.

3. Terminology Table

Term	Simple meaning
Population	The whole group you want to know about
Sample	The small part you actually study
Census	Studying every member of the population
Sampling frame	The list of all population members
Sampling unit	One member (one tablet, one patient)
Random sampling	Selection purely by chance, like a lottery
Systematic sampling	Selecting every k-th member from a list
Stratified sampling	Dividing the population into groups (strata) first, then sampling from each
Cluster sampling	Dividing the population into clusters and studying whole clusters
Convenience sampling	Taking whoever is easily available
Purposive sampling	Selecting members deliberately for a purpose
Snowball sampling	One selected person refers the next person
Sampling error	The difference between sample result and true population result

Term	Simple meaning
Bias	A systematic (one-sided) error in selection

4. Simple Explanations

Population vs sample (start from zero)

TECHNICAL CONCEPT: The population is the complete set of units of interest; the sample is a subset of it.

SIMPLE EXPLANATION: Think of a huge vessel of dal being cooked for a hostel. The whole vessel is the **population** — all the dal. You do not need to eat the whole vessel to know whether the salt is right; you taste one spoonful. That spoonful is the **sample**. If the dal was stirred well, that one spoon tells you about the whole vessel.

EXAMPLE: A factory makes 10,00,000 (one million) tablets in a batch. The population = all 10,00,000 tablets. Testing all of them would destroy the whole batch (destructive testing). So the QC chemist tests a sample of 100 tablets and concludes about the whole batch.

KEY TAKEAWAY: Population = whole group; sample = the small part actually studied.



Why is sampling important?

TECHNICAL CONCEPT: Sampling allows valid conclusions about a population with limited time, cost and effort.

SIMPLE EXPLANATION: Studying the whole population (called a **census**) is often impossible. Imagine testing every tablet in a batch — the test itself breaks the tablet, so nothing would be left to sell! Or interviewing every diabetic patient in India — it would take years and crores of rupees.

EXAMPLE: To check the average weight of 5,00,000 tablets, weighing every tablet would take weeks. Weighing a sample of 200 tablets takes one afternoon and gives a reliable answer.

KEY TAKEAWAY: Sampling saves time, money and effort, and is the only option when testing destroys the product.

Probability vs non-probability sampling — the big division

TECHNICAL CONCEPT: In probability sampling every unit has a known chance of selection; in non-probability sampling selection depends on convenience or judgement.

SIMPLE EXPLANATION: Probability sampling is like a fair lottery — every ticket has an equal chance, so the result is fair and can represent everyone. Non-probability sampling is like asking your friends — quick and easy, but your friends may not represent the whole class.

KEY TAKEAWAY: Probability methods are fair and representative; non-probability methods are quick but may be biased.

5. Classifications

5.1 Probability sampling methods

TECHNICAL CONCEPT: Four main methods — simple random, systematic, stratified, cluster.

(a) Simple random sampling

SIMPLE EXPLANATION: Every member has an equal chance, like drawing names from a box. Use a lottery or a random-number table/generator.

EXAMPLE: From a list of 500 pharmacy students, pick 50 by lottery for a survey on study habits.

Steps:

1 Number every member of the population (1 to 500).

2 Generate random numbers (lottery slips or computer).

3 Select the members matching those numbers.

(b) Systematic sampling

SIMPLE EXPLANATION: Pick every k-th member from an ordered list. If you need 50 out of 500, k = 500 ÷ 50 = 10, so pick every 10th name.

EXAMPLE: In a hospital OPD register, interview every 10th patient to study waiting time.

Steps:

1 Calculate the interval k = population size ÷ sample size.

2 Choose a random starting point between 1 and k.

3 Select every k-th member after that.

```
Register: 1  2  3  4  5  6  7  8  9 [10] 11 12 ... [20] ...
                                      ↑                ↑
                                   selected         selected
(k = 10, start = 10)
```

Caution: If the list has a hidden pattern (e.g., every 10th bed is near the nurse station), systematic sampling can go wrong.

(c) Stratified sampling

SIMPLE EXPLANATION: First divide the population into natural groups (**strata**), then take a random sample from EACH group. This guarantees every group is represented.

EXAMPLE: A college has 300 B.Pharm, 200 D.Pharm and 100 M.Pharm students. To survey 60 students fairly, take 30 from B.Pharm, 20 from D.Pharm and 10 from M.Pharm (same proportion), each chosen randomly.

Steps:

1 Identify the strata (e.g., course, gender, age group).

2 Decide how many to take from each stratum (usually in proportion to its size).

3 Draw a random sample within each stratum.

(d) Cluster sampling

SIMPLE EXPLANATION: Divide the population into **clusters** (natural small groups like villages, wards, colleges), randomly select a few whole clusters, and study EVERYONE in those clusters.

EXAMPLE: To study medicine use in a district with 100 villages, randomly pick 10 villages and interview all households in those 10 villages.

KEY DIFFERENCE (strata vs clusters):

- Stratified: sample from EVERY group, only some members of each.
- Cluster: sample only SOME groups, but ALL members of the chosen groups.

5.2 Non-probability sampling methods

(a) Convenience sampling

SIMPLE EXPLANATION: Take whoever is easily available. Quick and cheap, but the sample may not represent the population.

EXAMPLE: Interviewing patients sitting in the waiting hall of your own hospital about drug prices.

(b) Purposive (judgemental) sampling

SIMPLE EXPLANATION: The researcher deliberately selects specific people who are most informative for the purpose of the study.

EXAMPLE: For a study on insulin handling, deliberately selecting only diabetic patients who use insulin — because only they can answer the questions.

(c) Snowball sampling

SIMPLE EXPLANATION: Start with one or two suitable people, and ask each of them to refer the next person — the sample grows like a rolling snowball.

EXAMPLE: To study self-medication with sleeping pills (a sensitive topic), one user refers another user, who refers the next.

6. Processes

6.1 General steps of a sampling plan

Step 1 — Define the population. Be exact. "All B.Pharm students of our college" — not just "students".

Step 2 — Prepare the sampling frame. Get or make the complete list (register, roll list, batch records).

Step 3 — Choose the sampling method. Random, systematic, stratified, cluster, convenience, purposive or snowball — based on the objective, time and resources.

Step 4 — Decide the sample size. How many units are needed? (Sample-size calculation is covered in later chapters.)

Step 5 — Select the sample. Follow the method's procedure honestly — no picking favourites.

Step 6 — Collect data from the selected units. Only from the sample, not from outsiders.

6.2 Step-by-step: simple random sampling with a lottery

1 Write the names/numbers of all population members on identical slips.

2 Mix the slips thoroughly in a box.

3 Without looking, draw the required number of slips.

4 Those drawn form the sample.

7. Textual Diagrams / Flowcharts

Flowchart: Choosing a sampling method

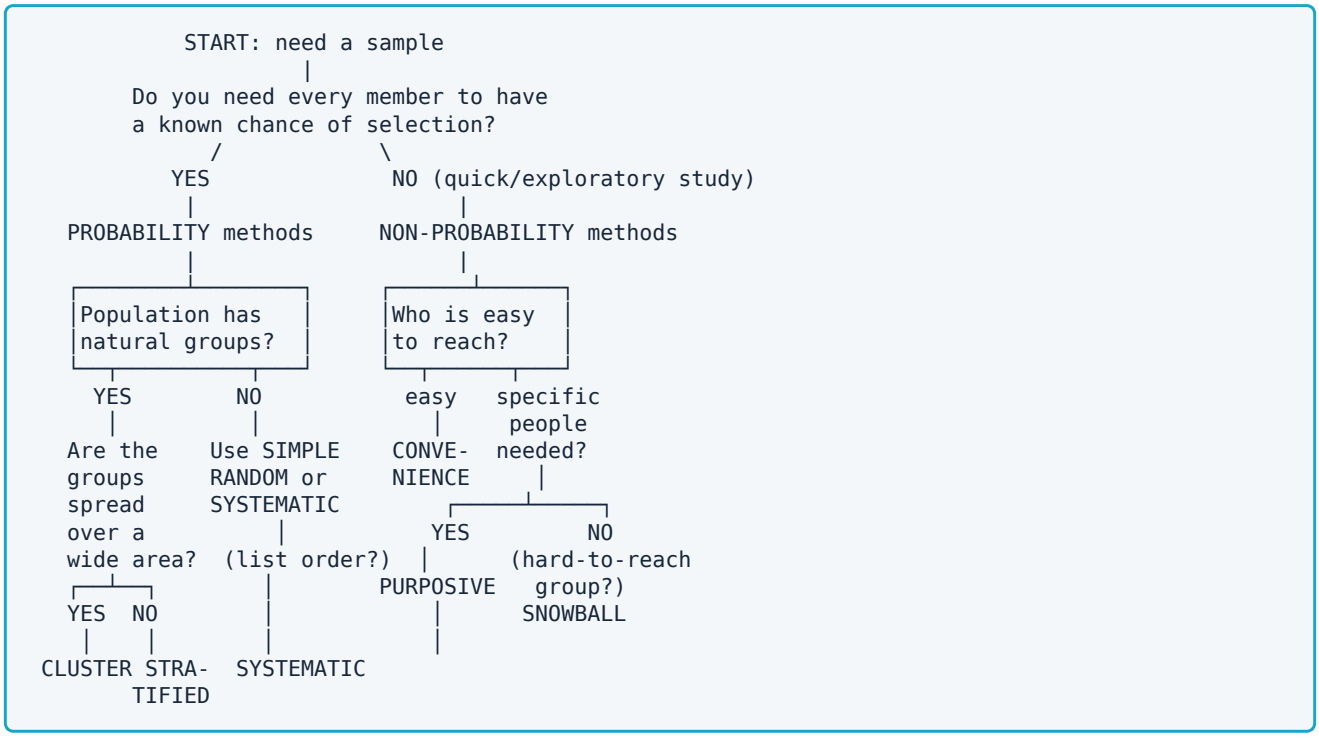
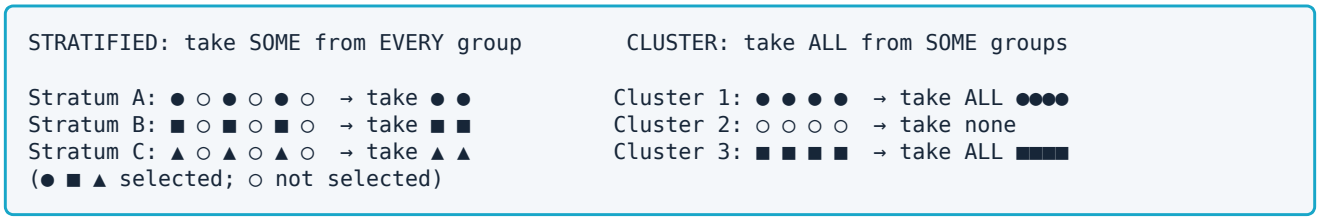


Diagram: Stratified vs cluster (side by side)



8. > INFOGRAPHIC CONCEPT

INFOGRAPHIC CONCEPT: A one-page visual titled "Seven Ways to Pick a Sample". Seven panels in two rows: top row (blue header "PROBABILITY — everyone gets a fair chance") shows (1) lottery box for random, (2) a numbered list with every 10th name highlighted for systematic, (3) three jars labelled B.Pharm/D.Pharm/M.Pharm each contributing balls for stratified, (4) a map with a few whole villages circled for cluster. Bottom row (orange header "NON-PROBABILITY — quick but careful") shows (5) a hand grabbing nearby people for convenience, (6) a magnifying glass picking insulin users for purposive, (7) a growing chain of people icons for snowball. A footer strip reads: "Probability = fair and representative. Non-probability = fast but possibly biased."

9. Tables

Table 1: Probability vs non-probability sampling

Feature	Probability sampling	Non-probability sampling
Chance of selection	Known for every member	Not known
Selection basis	Random/chance	Convenience or judgement
Representativeness	Usually representative	May be biased
Statistical inference	Valid (results can be generalised)	Limited
Time and cost	More	Less
Examples	Random, systematic, stratified, cluster	Convenience, purposive, snowball

Table 2: The four probability methods compared

Method	How it works	Best when	Watch out for
Simple random	Lottery / random numbers	Population is small and uniform	Needs a complete list
Systematic	Every k-th member	List is available, quick work needed	Hidden patterns in the list
Stratified	Sample from every stratum	Population has clear subgroups	Strata must be correctly defined
Cluster	Study whole selected clusters	Population spread over a wide area	Clusters must resemble each other

Table 3: The three non-probability methods compared

Method	How it works	Best when	Limitation
Convenience	Whoever is available	Pilot study, very limited time	Highly biased possible
Purposive	Deliberately chosen experts/cases	Specific information needed	Researcher's judgement may be wrong
Snowball	Referrals chain	Hard-to-reach or sensitive groups	Sample may be too similar

10. Worked Examples and Calculation Boxes

CALCULATION BOX 1: Systematic sampling interval

Problem: A hospital has 600 registered diabetic patients. You want a systematic sample of 60 patients. Find the sampling interval k and list the first three selected patients if the random start is 4.

Solution:

- Step 1: Formula: k = population size ÷ sample size = N ÷ n
- Step 2: Substitute: k = 600 ÷ 60 = 10
- Step 3: Random start = 4, so select patient 4, then add k each time: 4, 4+10 = 14, 14+10 = 24

Final answer: k = 10; first three patients are numbers 4, 14 and 24.

CALCULATION BOX 2: Proportional stratified sampling

Problem: A pharmacy college has 400 B.Pharm, 300 D.Pharm and 100 Pharm.D students (total 800). You need a stratified sample of 80 students in proportion to each course. How many from each?

Solution:

- Step 1: Fraction to sample: 80 ÷ 800 = 1/10 (10%)
- Step 2: B.Pharm: 400 × 1/10 = 40 students
- Step 3: D.Pharm: 300 × 1/10 = 30 students
- Step 4: Pharm.D: 100 × 1/10 = 10 students
- Step 5: Check: 40 + 30 + 10 = 80 ✓

Final answer: 40 B.Pharm, 30 D.Pharm, 10 Pharm.D students.

Worked example: Choosing the right method

- **Situation A:** Study drug use in all villages of a district → **cluster sampling** (villages are natural clusters; travelling to every village is costly).
- **Situation B:** Compare HbA1c levels between male and female diabetics → **stratified sampling** (gender strata ensure both groups are represented).
- **Situation C:** Quick feedback on a new hospital pharmacy counter → **convenience sampling** (ask patients present today).
- **Situation D:** Study of patients who experienced a rare side effect → **purposive sampling** (only those patients have the needed information).

11. Analogies

- **Sampling is tasting dal from a vessel.** You don't eat the whole vessel; one well-stirred spoonful tells you about all of it.
- **Random sampling is a fair lottery.** Every ticket has an equal chance — that is why the result is trusted.

- **Stratified sampling is making mixed pakoras.** You take some onion, some potato and some chilli pieces so every flavour is in the plate.
- **Snowball sampling is a rolling snowball.** It starts small and grows as each person brings the next.
- **Sampling frame is the attendance register.** Without the full list, you cannot fairly pick names from the class.

12. Pharmacy Relevance

- **Quality control:** Every batch of tablets is a population; the few tablets tested for weight, hardness and dissolution form the sample. Acceptance or rejection of the whole batch depends on this sample.
- **Bioequivalence studies:** A sample of healthy volunteers represents the population of future users of a generic medicine.
- **Drug utilisation studies:** Sampling prescriptions from a hospital pharmacy tells whether antibiotics are being overused.
- **Pharmacovigilance:** ADR reports are a sample of all adverse events; they signal drug safety problems in the population.

13. India Context

- The **National Family Health Survey (NFHS)** uses stratified and cluster sampling to cover India's huge, spread-out population — studying every household would be impossible.
- The **Census of India** is one of the rare examples of a true census (whole population counted), done once every ten years.
- In Indian pharmacy colleges, student projects commonly use convenience sampling (classmates, nearby pharmacies) because of limited time — students must honestly mention this limitation in their reports.

14. Key Points

1 Population = the whole group of interest; sample = the part actually studied.

2 Sampling saves time, cost and effort, and is essential when testing destroys the product.

3 A census studies everyone; a sample studies a part.

4 Probability sampling: every member has a known chance — random, systematic, stratified, cluster.

5 Non-probability sampling: selection by convenience/judgement — convenience, purposive, snowball.

6 Stratified = some from EVERY group; cluster = all from SOME groups.

7 Sampling frame is the complete list used for selection; bias is a one-sided selection error.

15. Common Confusions

- **Stratified vs cluster:** The most confused pair. Remember: stratified takes a *little from all* groups; cluster takes *all from a few* groups.
- **Sample vs sampling:** The sample is the *group selected*; sampling is the *process of selecting*.
- **Random vs haphazard:** Random sampling follows a proper chance procedure (lottery, random numbers). Just grabbing "anyone" is convenience sampling, not random sampling.
- **Population vs sample size:** "Population" means the whole group, not a large number. A sample of 10,000 can still be a sample if the population is 10 lakh.

16. Common Mistakes

1 Calling convenience sampling "random sampling" just because the selection felt casual.

2 Using systematic sampling on a list with a repeating pattern (e.g., every 10th entry is a holiday).

3 Forgetting to define the population clearly before sampling ("patients" — which hospital? which disease? which year?).

4 In stratified sampling, taking equal numbers from unequal strata instead of proportional numbers.

5 Generalising results from a convenience sample to the whole population as if it were a random sample.

17. Mnemonics

- Probability four: **"RaSyStCl"** — **Ra**ndom, **Sy**stematic, **St**ratified, **Cl**uster ("Rassi stuckle" — say it fast!).
- Non-probability three: **"CPS"** — **C**onvenience, **P**urposive, **S**nowball (like a "CPS" exam you clear quickly — and non-probability is the quick method).
- **"Little from All = straTified; All from liTTle = clusTer"** — the T's remind you which is which.
- Systematic interval: **"k = N ÷ n"** — "k for k-th, N for total Number, n for needed number".

18. Exam Summary

The **population** is the entire group about which conclusions are needed; the **sample** is the smaller part actually studied. **Sampling** is the selection process. Sampling is important because studying the whole population (census) costs too much time and money and is impossible when testing destroys the product (e.g., tablet testing). **Probability sampling** gives every member a known chance of selection: (1) **simple random** — lottery/random numbers, equal chance; (2) **systematic** — every k-th member where k = N ÷ n; (3) **stratified** — divide into strata, sample from every stratum; (4) **cluster** — select a few whole clusters and study all members in them. **Non-probability sampling** uses convenience or judgement: (1) **convenience** — whoever is available; (2) **purposive** — deliberately chosen informative cases; (3) **snowball** — referrals grow the sample. Probability methods are representative and allow generalisation; non-probability methods are quick but may be biased.

19. One-Minute Revision

- Population = whole; Sample = studied part; Sampling = selection process.

- Census = everyone; usually too costly → take a sample.
- Probability (fair, known chance): Random, Systematic (every k-th, k = N/n), Stratified (some from every group), Cluster (all from some groups).
- Non-probability (quick, may be biased): Convenience, Purposive, Snowball.
- Sampling frame = complete list; Sampling unit = one member.
- Bias = one-sided selection error.

20. Questions

Short-answer questions (2–3 marks)

1 Define population and sample with one pharmacy example.

2 What is a sampling frame? Give an example.

3 Calculate the sampling interval for a systematic sample of 40 from a population of 800.

4 Differentiate between stratified and cluster sampling in two lines.

5 When is snowball sampling used?

Long-answer questions (5–10 marks)

1 Define population and sample. Explain the importance of sampling in pharmaceutical research with examples.

2 Describe the four probability sampling methods with a pharmacy example and steps for each.

3 Describe the three non-probability sampling methods. When would you prefer them over probability methods?

4 Compare probability and non-probability sampling under suitable headings.

MCQs (10, with answer key)

Q1. The entire group about which conclusions are drawn is called the: a) Sample b) Population c) Census d) Stratum **Answer: b**

Q2. Studying every member of the population is called: a) Sampling b) Census c) Stratification d) Randomisation **Answer: b**

Q3. In which method does every member have a known chance of selection? a) Convenience b) Purposive c) Simple random d) Snowball **Answer: c**

Q4. For a population of 1,000 and a desired sample of 100, the systematic sampling interval k is: a) 5 b) 10 c) 100 d) 1,000 **Answer: b**

Q5. Dividing students into B.Pharm, D.Pharm and Pharm.D groups and sampling from each group is: a) Cluster sampling b) Stratified sampling c) Snowball sampling d) Convenience sampling **Answer: b**

Q6. Selecting a few whole villages and studying all households in them is: a) Stratified sampling b) Systematic sampling c) Cluster sampling d) Random sampling **Answer: c**

Q7. Asking patients present in the waiting hall to fill a feedback form is: a) Random sampling b) Convenience sampling c) Stratified sampling d) Systematic sampling **Answer: b**

Q8. Snowball sampling grows through: a) Lottery draws b) Referrals from selected members c) Every 10th name d) Government lists **Answer: b**

Q9. A complete list of all population members used for sampling is the: a) Sampling unit b) Sampling frame c) Stratum d) Cluster **Answer: b**

Q10. Which sampling method is most likely to give a biased sample? a) Simple random b) Stratified c) Convenience d) Systematic **Answer: c**

Viva questions

1 Why can't we test every tablet in a batch?

2 What is the difference between a sample and the sampling process?

3 Give the formula for the systematic sampling interval.

4 Name the two broad categories of sampling methods.

5 When would you use purposive sampling?

Application-based questions

1 You must estimate the average waiting time at a hospital pharmacy counter that serves about 500 patients a day. Design a sampling plan: define the population, choose a method, and justify your choice.

2 A researcher wants to compare medicine-adherence between rural and urban diabetics in a district. Which probability method fits best? Explain with steps.

3 For a study on misuse of sleeping pills among college students (a sensitive topic), which non-probability method would you use and why?

21. References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)

CHAPTER 3: Descriptive Statistics — Central Tendency, Dispersion and Frequency Diagrams

1. Learning Objectives

After studying this chapter, you will be able to:

- Distinguish descriptive statistics from inferential statistics

- Construct a frequency distribution from raw data
- Calculate mean, median and mode with all steps shown
- Calculate range, variance and standard deviation with all steps shown
- Explain degrees of freedom, quartiles, skewness and kurtosis in simple words
- Calculate quartiles for a small dataset
- Draw and interpret diagrammatic representations of frequency distributions (histogram, frequency polygon, ogive, bar and pie diagrams)

2. Definitions

Descriptive statistics: Methods that describe and summarise data (averages, tables, charts) without going beyond the data collected.

Inferential statistics: Methods that use sample data to draw conclusions (inferences) about the whole population.

Frequency distribution: A table showing how often each value (or class of values) occurs in the data.

Mean: The arithmetic average — sum of all values divided by the number of values.

Median: The middle value when data are arranged in ascending order.

Mode: The value that occurs most frequently.

Range: Maximum value minus minimum value.

Variance: The average of the squared differences of values from the mean.

Standard deviation (SD): The square root of variance; the typical distance of values from the mean.

Degrees of freedom: The number of values free to vary after a statistic (like the mean) is fixed; for a sample it is n − 1.

Quartiles: Three values (Q1, Q2, Q3) that divide ordered data into four equal parts.

Skewness: A measure of the asymmetry (lopsidedness) of a distribution.

Kurtosis: A measure of how peaked or flat a distribution is compared to the normal curve.

3. Terminology Table

Term	Simple meaning
Descriptive statistics	Summarising the data you have
Inferential statistics	Concluding about the population from the sample
Frequency	How many times a value occurs
Class interval	A group range in a frequency table (e.g., 110–119)
Central tendency	The "centre" of the data (mean, median, mode)
Dispersion	How spread out the data are
Deviation	How far one value is from the mean

Term	Simple meaning
Squared deviation	Deviation multiplied by itself (removes minus signs)
Sample SD (s)	SD calculated with n − 1 in the denominator
Population SD (σ)	SD calculated with N in the denominator
IQR	Interquartile range = Q3 − Q1
Histogram	Bar chart of a frequency distribution (bars touch)
Ogive	Cumulative frequency curve
Normal curve	Bell-shaped symmetric curve

4. Simple Explanations

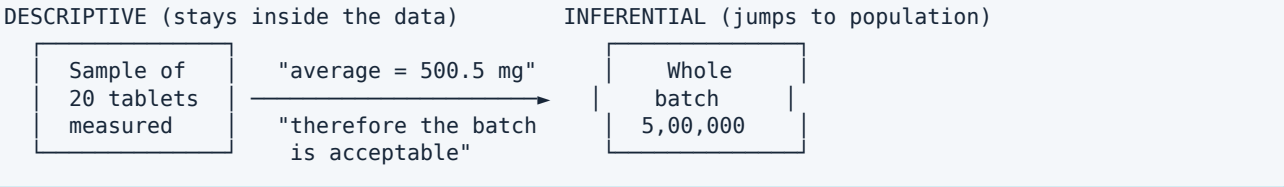
4.1 Descriptive vs inferential statistics (start from zero)

TECHNICAL CONCEPT: Descriptive statistics summarise the sample; inferential statistics generalise from the sample to the population.

SIMPLE EXPLANATION: Descriptive statistics is like describing the spoonful of dal you tasted — "this spoon is salty." Inferential statistics is the next step — "so the whole vessel is probably salty." Description stays with what you measured; inference jumps from the sample to the population.

EXAMPLE: You weigh 20 tablets and report "average weight = 500.5 mg, range 498–503 mg" — that is **descriptive**. You then conclude "therefore the whole batch of 5,00,000 tablets meets the weight specification" — that is **inferential**.

KEY TAKEAWAY: Descriptive = "what the data look like"; inferential = "what the data tell us about the bigger group".



4.2 Frequency distribution (start from zero)

TECHNICAL CONCEPT: A frequency distribution organises raw data into values/classes with their counts.

SIMPLE EXPLANATION: Raw data is a messy pile. A frequency distribution is like sorting that pile into labelled boxes and writing how many items went into each box. First list each distinct value (or group), then count (tally) how many times it appears.

EXAMPLE: Fifteen patient systolic BP readings (mmHg): 112, 118, 121, 125, 129, 132, 135, 138, 122, 127, 131, 119, 124, 128, 134. Grouped into classes:

BP class (mmHg)	Tally	Frequency
110–119	III	3
120–129	IIII II	7

BP class (mmHg)	Tally	Frequency
130–139	IIII	5
Total		15

KEY TAKEAWAY: Frequency distribution = values (or classes) + how many times each occurs. Always check the total matches the number of observations.

4.3 Measures of central tendency (start from zero)

TECHNICAL CONCEPT: Mean, median and mode each locate the centre of the data in a different way.

SIMPLE EXPLANATION: If someone asks "what is the *typical* tablet weight?", you need one number to represent all. Three candidates:

- **Mean** = the fair-share average (add everything, divide equally).
- **Median** = the middle person in a line arranged by height.
- **Mode** = the most popular value (the winner of the votes).

EXAMPLE: 5 tablets weighing 498, 500, 502, 499, 501 mg → mean (fair share) = 2500 ÷ 5 = 500 mg; arranged (498, 499, 500, 501, 502) the middle = median 500 mg. For BP readings 120, 130, 120, 125, 130, 130, 128, the value 130 occurs three times → mode = 130 mmHg.

KEY TAKEAWAY: Mean = average; median = middle; mode = most frequent. For clean symmetric data they are nearly equal.

4.4 Measures of dispersion (start from zero)

TECHNICAL CONCEPT: Dispersion measures how spread out the values are around the centre.

SIMPLE EXPLANATION: Two batches can both average 500 mg, yet one batch has all tablets between 498–502 mg (good, uniform) and the other has tablets from 480–520 mg (bad, non-uniform). The average alone hides this! **Dispersion** tells you the spread.

- **Range** = max − min (quickest, roughest).
- **Variance** = average of squared differences from the mean.
- **Standard deviation** = square root of variance (back in original units, e.g., mg).

EXAMPLE: Batch A weights: 499, 500, 501 (range = 2). Batch B weights: 490, 500, 510 (range = 20). Same mean (500), very different spread — dispersion reveals the difference.

KEY TAKEAWAY: Central tendency tells you the centre; dispersion tells you the spread. Report both, always.

4.5 Degrees of freedom (start from zero)

TECHNICAL CONCEPT: Degrees of freedom = number of independent values free to vary.

SIMPLE EXPLANATION: Suppose 3 numbers must average to 10 (sum = 30). You can freely choose the first two numbers (say 8 and 12), but the third is then FORCED to be 10 (because 8+12+10 = 30). So out of 3 numbers, only 2 are free → degrees of freedom = 3 − 1 = 2. In general, df = n − 1.

EXAMPLE: For a sample of 10 tablet weights, once the mean is fixed, 9 weights are free and the 10th is forced → df = 9. That is why sample variance divides by (n − 1).

KEY TAKEAWAY: df = n − 1 for a sample, because fixing the mean uses up one "freedom".

4.6 Quartiles (start from zero)

TECHNICAL CONCEPT: Quartiles split ordered data into four equal parts: Q1 (25%), Q2 = median (50%), Q3 (75%).

SIMPLE EXPLANATION: Line up the data from smallest to largest, like students in a height line. Q2 is the middle student (the median). Q1 is the middle of the *lower half*; Q3 is the middle of the *upper half*.

EXAMPLE: 8 dissolution values (%): 82, 85, 88, 90, 92, 94, 96, 99.

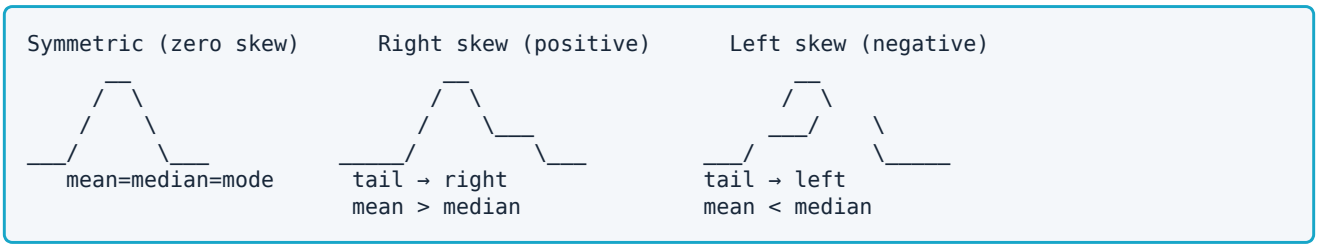
- Q2 (median) = (90 + 92) ÷ 2 = 91
- Q1 = middle of (82, 85, 88, 90) = (85 + 88) ÷ 2 = 86.5
- Q3 = middle of (92, 94, 96, 99) = (94 + 96) ÷ 2 = 95
- IQR = Q3 − Q1 = 95 − 86.5 = 8.5

KEY TAKEAWAY: Quartiles show where the lower quarter, middle and upper quarter of the data lie. IQR = Q3 − Q1 measures the spread of the middle 50%.

4.7 Skewness and kurtosis (start from zero)

TECHNICAL CONCEPT: Skewness measures asymmetry; kurtosis measures peakedness relative to the normal (bell) curve.

SIMPLE EXPLANATION: A perfectly symmetric bell curve has zero skewness. If a few very large values stretch the right side into a long tail, the curve leans left and the tail points right → **positively (right) skewed**. If a few very small values stretch the left side → **negatively (left) skewed**. **Kurtosis** describes the peak: a tall sharp peak (heavy tails) vs a flat low peak.



EXAMPLE: Hospital stay durations are usually right-skewed: most patients stay 2–4 days, but a few stay 30+ days, pulling the mean above the median.

KEY TAKEAWAY: Skewness = which side the tail is on; the mean is pulled towards the tail. Kurtosis = how peaked or flat the curve is.

5. Classifications

5.1 Types of statistics

	Descriptive statistics	Inferential statistics
Purpose	Describe/summarise collected data	Draw conclusions about the population
Goes beyond the data?	No	Yes
Tools	Mean, median, mode, SD, tables, charts	Hypothesis tests, confidence intervals, p-values (later units)
Example	"Average weight of 20 tablets = 500.5 mg"	"The batch meets the specification"

5.2 Measures of central tendency — when to use which

- **Mean:** Best for symmetric numerical data without outliers. (Most common in pharmacy QC.)
- **Median:** Best when outliers exist or data are skewed (e.g., hospital stay days, income).
- **Mode:** Best for categorical data (e.g., most common blood group, most prescribed antibiotic).

5.3 Measures of dispersion — the family

Range (simplest) → Variance (squared units) → Standard deviation (original units, most useful) → IQR (middle-50% spread, outlier-resistant).

6. Processes

6.1 Steps to build a frequency distribution

Step 1 — Arrange raw data in ascending order. **Step 2 —** Decide the number of classes (roughly 5–10 for small data) and the class width. **Step 3 —** Write the class intervals (e.g., 110–119, 120–129). **Step 4 —** Tally each observation into its class. **Step 5 —** Count tallies = frequency of each class; check the total.

6.2 Steps to calculate mean, median, mode

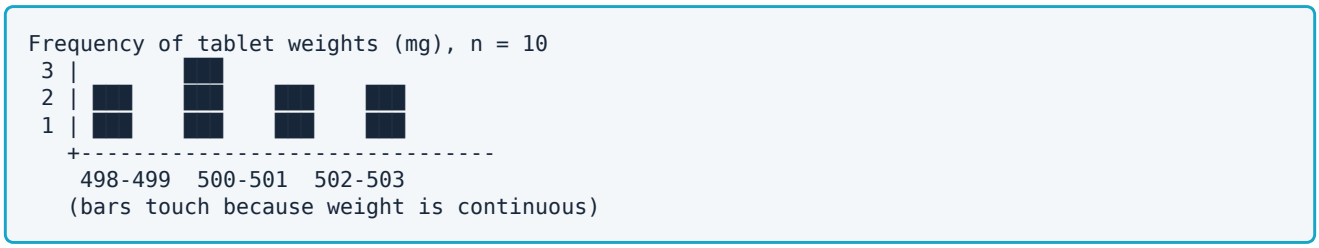
- **Mean:** add all values → divide by n.
- **Median:** arrange ascending → if n is odd, take the middle ((n+1)/2-th) value; if n is even, average the two middle values (n/2-th and (n/2+1)-th).
- **Mode:** count each value → the value with the highest count.

6.3 Steps to calculate variance and SD

Step 1 — Calculate the mean. **Step 2 —** Subtract the mean from each value (deviations). **Step 3 —** Square each deviation. **Step 4 —** Add all squared deviations. **Step 5 —** Divide by $(n - 1)$ for sample variance, or by N for population variance. **Step 6 —** Take the square root → standard deviation.

7. Textual Diagrams / Flowcharts

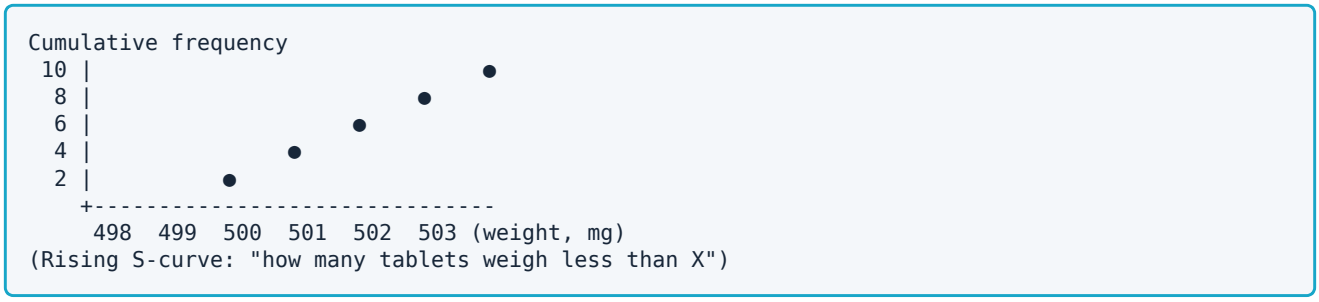
Histogram (bars touch — for continuous data)



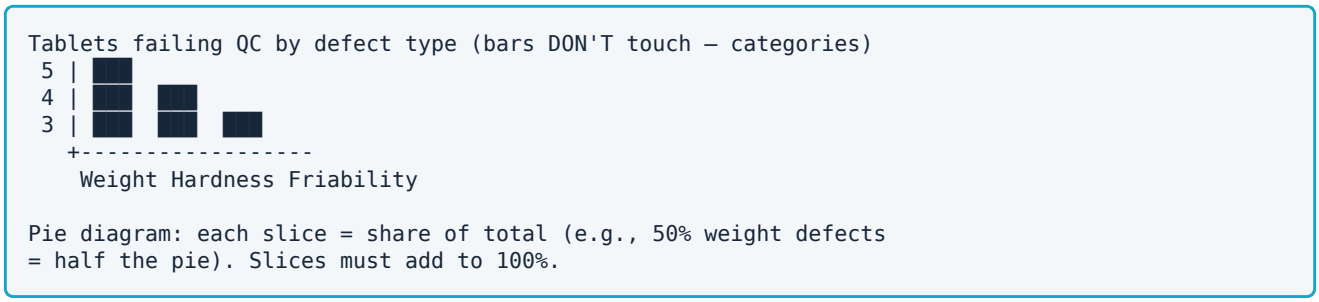
Frequency polygon (join midpoints)



Ogive — cumulative frequency curve ("less than" type)



Bar diagram (bars separated — for categories) and pie diagram



8. > INFOGRAPHIC CONCEPT

INFOGRAPHIC CONCEPT: A one-page visual titled "Describe Any Data in 6 Steps". Six connected panels: (1) a messy pile of numbers labelled "RAW DATA — 10 tablet weights"; (2) sorted boxes labelled "FREQUENCY TABLE — count what falls where"; (3) a target with an arrow at the bullseye labelled "CENTRE — mean 500.5, median 500.5, mode 499/500/501/502"; (4) arrows spreading outward labelled "SPREAD — range 5 mg, SD 1.58 mg"; (5) a bell curve with a right-side tail labelled "SHAPE — skewness and kurtosis"; (6) a histogram, pie chart and ogive labelled "SHOW — pick the right diagram". A banner across the bottom reads: "Centre + Spread + Shape + Picture = complete description."

9. Tables

Table 1: Mean vs median vs mode

Feature	Mean	Median	Mode
Definition	Sum ÷ n	Middle value	Most frequent value
Uses all values?	Yes	No (position only)	No (frequency only)
Affected by outliers?	Yes, strongly	No	No
Works for categories?	No	No (needs order)	Yes
Pharmacy example	Average tablet weight	Middle dissolution value	Most common blood group

Table 2: Dispersion measures compared

Measure	Formula idea	Units	Best feature
Range	max − min	Same as data	Quickest
Variance	average of squared deviations	Squared units (mg^2)	Basis for advanced statistics
Standard deviation	√variance	Same as data (mg)	Most used and interpretable
IQR	Q3 − Q1	Same as data	Not affected by outliers

Table 3: Which diagram for which data

Diagram	Use for	Bars touch?
Histogram	Continuous data frequency distribution	Yes
Frequency polygon	Comparing two distributions	(lines, not bars)
Ogive	Cumulative frequencies, finding median graphically	(curve)
Bar diagram	Categorical data	No
Pie diagram	Parts of a whole (percentages)	(slices)

10. Worked Examples and Calculation Boxes

Dataset used below (tablet weights, mg, n = 10): 498, 502, 500, 501, 499, 503, 500, 502, 499, 501

CALCULATION BOX 1: Mean

Problem: Calculate the mean tablet weight.

Solution:

- Step 1: Add all values: 498 + 502 + 500 + 501 + 499 + 503 + 500 + 502 + 499 + 501
- Step 2: Add in pairs: (498+502) = 1000; (500+501) = 1001; (499+503) = 1002; (500+502) = 1002; (499+501) = 1000
- Step 3: Total = 1000 + 1001 + 1002 + 1002 + 1000 = 5005
- Step 4: Divide by n = 10: 5005 ÷ 10 = 500.5

Final answer: Mean = 500.5 mg.

CALCULATION BOX 2: Median

Problem: Calculate the median of the same 10 weights.

Solution:

- Step 1: Arrange ascending: 498, 499, 499, 500, 500, 501, 501, 502, 502, 503
- Step 2: n = 10 (even), so median = average of the 5th and 6th values
- Step 3: 5th value = 500; 6th value = 501
- Step 4: (500 + 501) ÷ 2 = 1001 ÷ 2 = 500.5

Final answer: Median = 500.5 mg.

CALCULATION BOX 3: Mode

Problem: Find the mode of: 120, 130, 120, 125, 130, 130, 128 (systolic BP, mmHg).

Solution:

- Step 1: Count each value: 120 appears 2 times; 125 appears 1 time; 128 appears 1 time; 130 appears 3 times
- Step 2: The highest count is 3, for the value 130

Final answer: Mode = 130 mmHg.

CALCULATION BOX 4: Range

Problem: Calculate the range of the tablet-weight dataset.

Solution:

- Step 1: Maximum = 503 mg; minimum = 498 mg
- Step 2: Range = max − min = 503 − 498 = 5

Final answer: Range = 5 mg.

CALCULATION BOX 5: Sample variance and sample standard deviation (full steps)

Problem: Calculate the sample variance (s^2) and sample SD (s) of the tablet-weight dataset. Mean = 500.5 mg.

Solution:

- Step 1: Find each deviation (value − mean) and square it:

Value	Deviation (x − 500.5)	Squared deviation
498	−2.5	6.25
502	+1.5	2.25
500	−0.5	0.25
501	+0.5	0.25
499	−1.5	2.25
503	+2.5	6.25
500	−0.5	0.25
502	+1.5	2.25
499	−1.5	2.25
501	+0.5	0.25

- Step 2: Add the squared deviations: 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 + 0.25 + 2.25 + 2.25 + 0.25 = 22.50
- Step 3: Degrees of freedom = n − 1 = 10 − 1 = 9
- Step 4: Sample variance s^2 = 22.50 ÷ 9 = 2.50 mg^2
- Step 5: Sample SD s = √2.50. Since 1.58 × 1.58 = 2.4964 ≈ 2.50, s ≈ 1.58 mg

Final answer: Sample variance = 2.50 mg^2; sample SD = 1.58 mg.

CALCULATION BOX 6: Quartiles and IQR

Problem: Dissolution results (%) of 8 tablets: 82, 85, 88, 90, 92, 94, 96, 99. Find Q1, Q2, Q3 and IQR.

Solution:

- Step 1: Data are already ascending; n = 8
- Step 2: Q2 (median) = average of 4th and 5th values = (90 + 92) ÷ 2 = 182 ÷ 2 = 91
- Step 3: Lower half = 82, 85, 88, 90 → Q1 = average of 2nd and 3rd = (85 + 88) ÷ 2 = 173 ÷ 2 = 86.5
- Step 4: Upper half = 92, 94, 96, 99 → Q3 = average of 2nd and 3rd = (94 + 96) ÷ 2 = 190 ÷ 2 = 95
- Step 5: IQR = Q3 − Q1 = 95 − 86.5 = 8.5

Final answer: Q1 = 86.5, Q2 = 91, Q3 = 95, IQR = 8.5%.

11. Analogies

- **Mean is the fair-share division.** Like dividing 10 laddus equally among 5 children — each gets 2, even if the original pieces differed.
- **Median is the middle student in a height line.** Even if the tallest student is a giant, the middle student's height doesn't change — medians resist outliers.

- **SD is the "typical wobble".** If tablet weights wobble only ±1.5 mg around the mean, the process is steady; a ±20 mg wobble means trouble.

12. Pharmacy Relevance

- **Tablet weight uniformity:** Mean weight near the target + small SD = uniform batch. IP specifications use such summaries for batch release.
- **Dissolution testing:** Mean % drug released at 30 min and its SD decide whether a batch passes; quartiles summarise the middle 50% of tablets.
- **Clinical readings:** Mean reduction in BP after treatment (descriptive) supports the claim that the drug works (inferential).

13. India Context

- Indian Pharmacopoeia (IP) tests like weight variation and dissolution are reported using means and ranges — the descriptive statistics of this chapter in daily QC use.
- Government health reports (e.g., NFHS) and pharmacy college projects present data with averages, percentages and bar diagrams before attempting any advanced analysis.

14. Key Points

1 Descriptive statistics summarise data; inferential statistics generalise to the population.

2 Frequency distribution = values/classes with their counts (tally first, then count).

3 Mean = sum ÷ n; median = middle value; mode = most frequent value.

4 Range = max − min; variance = average of squared deviations; SD = √variance.

5 Sample variance divides by (n − 1); population variance divides by N.

6 Degrees of freedom for a sample = n − 1.

7 Quartiles divide ordered data into quarters; IQR = Q3 − Q1.

8 Skewness = asymmetry (mean pulled towards the tail); kurtosis = peakedness.

9 Histogram: bars touch (continuous data); bar diagram: bars separated (categories); ogive: cumulative curve; pie: parts of a whole.

15. Common Confusions

- **Variance vs SD:** Variance is in squared units (mg^2) and hard to interpret; SD (mg) is the interpretable one. Students often report variance when the question asks for SD.
- **Sample vs population formulas:** Sample variance divides by (n − 1); population variance divides by N. Using the wrong one is the commonest calculation error.
- **Median position with even n:** With 10 values the median is the average of the 5th AND 6th values — not just the 5th.

- **Histogram vs bar diagram:** Histogram bars touch (continuous classes); bar diagram bars are separated (distinct categories).
- **Skewness direction:** The skew is named after the TAIL, not the hump. Tail on the right = right (positive) skew.
- **Mode of continuous data:** With all-distinct measurements (like our weights), every value occurs once — mode is not meaningful there; use grouped classes instead.

16. Common Mistakes

1 Forgetting to arrange data in ascending order before finding the median or quartiles.

2 Dividing by n instead of (n − 1) when calculating sample variance/SD.

3 Reporting the mean for highly skewed data (e.g., hospital stay) where the median is better.

4 Drawing a histogram with gaps between bars, or a bar diagram with touching bars.

5 Rounding off at every step instead of only at the final answer, accumulating error.

6 Writing the SD with squared units (mg^2) — SD is always in the original units.

17. Mnemonics

- **"Mean needs Money-sense: add all, divide by all"** — sum ÷ n.
- **"Median = Middle"** — both start with M; arrange first, then take the middle.
- **"Mode = Most"** — both start with Mo; the most frequent wins.
- **"SD is the Square-root Doctor"** — it cures the squared units of variance back to normal units.
- **"n minus 1 is the sample's price"** — sample variance always pays (n − 1).
- **"Quarters need Quartiles"** — Q1, Q2, Q3 cut data into 4 equal quarters.

18. Exam Summary

Descriptive statistics describe and summarise collected data (means, tables, charts); **inferential statistics** use sample data to conclude about the population. A **frequency distribution** lists values or classes with their frequencies. Measures of central tendency: **mean** (sum ÷ n), **median** (middle value of ordered data), **mode** (most frequent value). Measures of dispersion: **range** (max − min), **variance** (average of squared deviations from the mean), **standard deviation** (√variance, in original units). Sample variance divides by (n − 1) because degrees of freedom = n − 1. **Quartiles** (Q1, Q2, Q3) divide ordered data into four equal parts; **IQR** = Q3 − Q1. **Skewness** measures asymmetry (mean is pulled towards the tail); **kurtosis** measures peakedness. Diagrammatic representation: **histogram** (touching bars, continuous data), **frequency polygon** (lines joining midpoints), **ogive** (cumulative frequency curve), **bar diagram** (separated bars, categories), **pie diagram** (parts of a whole).

19. One-Minute Revision

- Descriptive = summarise; Inferential = conclude about population.

- Mean = sum/n; Median = middle; Mode = most frequent.
- Range = max−min; Variance = avg of squared deviations; SD = √variance.
- Sample variance ÷ (n−1); df = n−1.
- Q1/Q2/Q3 split data into quarters; IQR = Q3−Q1.
- Skew = tail side (mean follows the tail); Kurtosis = peakedness.
- Histogram bars touch; bar-diagram bars don't; ogive = cumulative curve; pie = % shares.

20. Questions

Short-answer questions (2–3 marks)

1 Differentiate between descriptive and inferential statistics with one example each.

2 Define frequency distribution. What is a class interval?

3 State the formulae for sample variance and sample standard deviation.

4 What are degrees of freedom? Give the value for a sample of size 12.

5 Define skewness and kurtosis in one line each.

6 When is the median preferred over the mean?

Long-answer questions (5–10 marks)

1 Explain the measures of central tendency with formulae, merits and limitations of each.

2 Explain range, variance and standard deviation. Why is SD preferred over variance for reporting?

3 Explain quartiles, skewness and kurtosis with diagrams.

4 Describe the diagrammatic representation of frequency distributions: histogram, frequency polygon, ogive, bar and pie diagrams.

MCQs (10, with answer key)

Q1. Which of the following is a descriptive statistic? a) Testing whether a new drug works b) Reporting the average weight of 20 tablets c) Estimating the population mean d) Calculating a p-value **Answer: b**

Q2. The middle value of ordered data is the: a) Mean b) Mode c) Median d) Range **Answer: c**

Q3. For the values 3, 5, 7, 5, 9, the mode is: a) 5 b) 6 c) 7 d) 3 **Answer: a**

Q4. Range of 12, 18, 25, 9, 30 is: a) 21 b) 18 c) 30 d) 9 **Answer: a**

Q5. Standard deviation is: a) The square of variance b) The square root of variance c) Twice the variance d) Half the range **Answer: b**

Q6. Sample variance is calculated by dividing the sum of squared deviations by: a) n b) n + 1 c) n − 1 d) 2n **Answer: c**

Q7. Q2 is the same as the: a) Mean b) Mode c) Median d) Range **Answer: c**

Q8. In a right-skewed distribution: a) Mean < median b) Mean > median c) Mean = median d) Mode > mean **Answer: b**

Q9. In a histogram, the bars: a) Are always separated b) Touch each other c) Are drawn vertically only d) Show cumulative frequency **Answer: b**

Q10. IQR is equal to: a) Q3 + Q1 b) Q3 − Q1 c) Q2 − Q1 d) Max − min **Answer: b**

Viva questions

1 What is the difference between variance and standard deviation?

2 Why do we divide by (n − 1) for sample variance?

3 If mean > median, what type of skewness is present?

4 Which diagram would you use for categorical data — histogram or bar diagram?

5 What does an ogive show?

Application-based questions

1 Ten capsules have weights (mg): 250, 252, 248, 251, 250, 249, 253, 250, 251, 252. Calculate the mean, median, range and sample SD with all steps.

2 A QC report shows dissolution data with mean = 92% and SD = 8%. A second batch shows mean = 92% and SD = 2%. Which batch is more uniform, and why? What does this teach about reporting the mean alone?

3 Hospital stay (days) for 9 patients: 2, 3, 2, 4, 3, 5, 4, 3, 30. Calculate the mean and median. Which is the better "typical" value here, and what does the difference tell you about skewness?

21. References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)