BP201T · B.PHARM THEORY

Applied Biostatistics and Data Analytics for Pharmaceutical Sciences

Unit I — Descriptive Statistics

Pills & Python Notes

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

Syllabus coverage: Unit I (6 hours).

PILLS & PYTHON

Pharmacy • Data • AI • Healthcare

Instagram @pillsandpython · pillsandpython.com · YouTube @PillsAndPython

Chapter 1 — Descriptive Statistics (Unit I, 6 hours)

1.1 Learning objectives

After studying this chapter, you will be able to:

1. Define statistics and biostatistics in your own words.
2. Classify any pharmacy measurement as nominal, ordinal, interval, or ratio data.
3. Name the main sources of data in pharmacy practice and research.
4. Calculate mean, median, and mode by hand and interpret each one.
5. Calculate range, variance, and standard deviation and explain what they tell you.
6. Recognise skewness and describe what a skewed distribution means for biological measurements.
7. Use Python (NumPy and Pandas) to compute descriptive statistics and interpret the output.

1.2 What is statistics? What is biostatistics?

Statistics is the science of collecting, organising, summarising, and drawing conclusions from numbers. **Biostatistics** is statistics applied to living things and medicine — patients, drugs, diseases, trials. **Data analytics** is the broader modern practice of using data (often with computers) to find patterns and support decisions.

> **Analogy:** A pharmacist's dispensing register is raw data. Statistics is the skill of reading that register and answering questions like "which medicine sold most this month?" — summarising many numbers into a few meaningful ones.

> **Key takeaway:** Statistics turns piles of numbers into answers you can act on.

1.3 Types of data — nominal, ordinal, interval, ratio

Before choosing any calculation, you must know what *type* your data is. There are four levels of measurement:

Type	Meaning	Key question	Pharmacy example	Can you average it?
Nominal	Named categories, no order	"Which group?"	Blood group (A/B/AB/O); dosage form (tablet/capsule/syrup)	No — names have no average
Ordinal	Ordered categories, gaps unknown	"What rank?"	Pain score mild/moderate/severe; tablet coating quality grade 1–5	Cautious — order exists but gaps are unequal
Interval	Numbers, equal gaps, **no true zero**	"How much more?"	Temperature in °C; calendar year	Yes (mean OK), but ratios are meaningless
Ratio	Numbers, equal gaps, **true zero**	"How many times more?"	Tablet weight in mg; blood drug concentration; number of patients	Yes — all maths allowed

Simple explanation of the zero rule: 0 mg of drug means *no drug at all* — a true zero, so weight is ratio data, and "100 mg is twice 50 mg" makes sense. But 0 °C does *not* mean "no temperature" (it is just the freezing point of water), so "20 °C is twice as hot as 10 °C" is wrong — temperature in °C is only interval data.

> **Analogy:** Nominal is like sorting tablets by *colour*, ordinal is like ranking cricket teams *1st, 2nd, 3rd* (you don't know by how much 1st beats 2nd), interval is like a *thermometer*, and ratio is like a *weighing balance* (zero weight = nothing there).

Mnemonic: **N**o **O**rder, **I**nterval-no-zero, **R**atio-real-zero — "**N**ames, **O**rder, **I**ntervals, **R**atios."

⚠️ **Terminology alert:** "Discrete" vs "continuous" is a *different* split from the four types above. Discrete = countable whole values (number of tablets dispensed); continuous = measurable on a scale (tablet weight 500.4 mg). A ratio variable can be discrete (number of patients) or continuous (weight).

Key takeaway: The data type decides which statistics are legal. Never take the mean of nominal data.

1.4 Sources of data in pharmacy

Source	What data is collected	Example question it answers
Clinical trials	Drug effects in volunteer/patient groups	"Does the new tablet lower BP better than the old one?"
Pharmacovigilance	Adverse drug reaction (ADR) reports	"Is this side effect linked to our drug?"
Quality control (QC)	Measurements on manufactured batches	"Is every tablet in this batch within weight limits?"
PK (pharmacokinetic) studies	Drug concentration in blood over time	"How fast is the drug absorbed and removed?"

India context: In India, ADR reports flow into the Pharmacovigilance Programme of India (PvPI), and clinical trials are registered in the Clinical Trials Registry – India (CTRI). Every such database is a source of data that needs exactly the statistics in this course.

Key takeaway: Wherever a pharmacist measures, counts, or reports — statistics follows.

1.5 Measures of central tendency — mean, median, mode

Central tendency = "the middle value" — one number that represents a whole group.

The mean (average)

Definition: Add all values, divide by how many there are.

$$\text{Mean} = (\text{sum of all values}) / (\text{number of values})$$

Worked example (illustrative): Weights of 10 tablets (mg): 502, 498, 505, 500, 503, 497, 501, 499, 504, 500.

- Step 1 — add them: 502 + 498 = 1000; + 505 = 1505; + 500 = 2005; + 503 = 2508; + 497 = 3005; + 501 = 3506; + 499 = 4005; + 504 = 4509; + 500 = **5009**.
- Step 2 — divide by 10: 5009 ÷ 10 = **500.9 mg**.
- Interpretation: the "typical" tablet weighs about 500.9 mg — very close to the 500 mg target.

The median (middle value)

Definition: Arrange values in order; the median is the middle one (or the average of the two middle ones if the count is even).

Worked example: Sorted weights: 497, 498, 499, 500, 500, 501, 502, 503, 504, 505.

- 10 values (even) → middle two are the 5th and 6th: 500 and 501.
- Median = (500 + 501) ÷ 2 = **500.5 mg**.

The mode (most frequent value)

Definition: The value that appears most often. Here, **500 mg** appears twice — mode = 500 mg. (A dataset can have no mode, one mode, or several.)

Which one to use? Mean uses every value but is pulled by extremes (one 700 mg tablet would drag the mean up). Median ignores extremes — better for skewed data like hospital bills. Mode is the only choice for nominal data (e.g., "most common blood group").

Analogy: Mean is like the *see-saw balance point* — one heavy child at the end tips the whole board. Median is like the *middle child in a line* — one giant at the end doesn't move the middle child.

Key takeaway: Mean = balance point, median = middle person, mode = most popular. Pick the one your data type and shape allow.

1.6 Measures of dispersion — range, variance, standard deviation

Dispersion = "how spread out the values are." Two batches can have the same mean weight but very different quality.

Range

Definition: Largest value − smallest value. Here: 505 − 497 = **8 mg**. Simple but uses only two values — one bad tablet can fool it.

Variance

Definition: The average of the *squared* differences from the mean. Steps:

1. Find the mean (500.9).
2. Subtract the mean from each value (the "deviation").
3. Square each deviation (squaring removes minus signs and punishes big deviations).
4. Average the squared deviations.

Worked example (same tablet weights):

Tablet	Weight	Deviation (w − 500.9)	Squared
1	502	+1.1	1.21
2	498	−2.9	8.41
3	505	+4.1	16.81
4	500	−0.9	0.81
5	503	+2.1	4.41
6	497	−3.9	15.21
7	501	+0.1	0.01
8	499	−1.9	3.61
9	504	+3.1	9.61
10	500	−0.9	0.81
Total			**60.91**

- **Population variance** (whole batch described): 60.91 ÷ 10 = **6.091 mg²**.
- **Sample variance** (10 tablets standing in for a huge batch): 60.91 ÷ (10 − 1) = 60.91 ÷ 9 = **6.768 mg²**.

Standard deviation (SD)

Definition: The square root of variance — it brings the units back to mg, so you can interpret it directly.

- Population SD = √6.091 ≈ **2.47 mg**. Sample SD = √6.768 ≈ **2.60 mg**.

Interpretation: Tablet weights typically wander about 2.5 mg from the mean. A small SD = consistent, good-quality batch; a large SD = uneven batch, QC problem.

> ⚠️ **Terminology alert — population vs sample variance:** When your data *is* the whole group, divide by **n**; when it is a *sample* of a bigger group, divide by **n − 1**. Software differs: NumPy's `std()` uses n by default, Pandas' `.std()` uses n − 1. This single default has confused generations of students.

> **Analogy:** Mean tells you where the *crowd stands*; SD tells you how *tightly the crowd huddles*. A huddle (small SD) is predictable; a scattered crowd (large SD) is not.

> **Key takeaway:** Central tendency says "where", dispersion says "how spread". A good QC report always gives both.

1.7 Skewness — the shape of the distribution

Skewness describes whether a distribution leans to one side.

- **Symmetric:** mean ≈ median (balanced, like our tablet weights).
- **Right-skewed (positive skew):** a long tail to the right; a few very large values drag the **mean above the median**.
- **Left-skewed (negative skew):** a long tail to the left; a few very small values drag the **mean below the median**.

> **Worked example (illustrative):** Hospital stay lengths (days) for 7 patients: 2, 3, 3, 4, 4, 5, 30.

- Mean = (2+3+3+4+4+5+30) ÷ 7 = 51 ÷ 7 ≈ **7.3 days**. Median = **4 days**.
- One 30-day stay pulls the mean up → right-skewed. Reporting "average stay 7.3 days" would mislead; the median (4 days) is the honest "typical" stay.

Pharmacy relevance: Biological measurements are often skewed — drug clearance times, hospital costs, and waiting times usually have right skew (most patients cluster low, a few stay very long). Always look at the shape before choosing mean vs median.

Mnemonic: "The mean **follows the tail**" — tail to the right, mean sits right of the median.

> **Key takeaway:** Skewness tells you which "average" to trust. Skewed data → prefer the median.

1.8 Descriptive statistics in Python — NumPy and Pandas

Python is the calculator for this course. Two libraries matter here:

- **NumPy** — fast number crunching on lists/arrays.
- **Pandas** — tables (called DataFrames) with labelled rows and columns, like an Excel sheet.

Snippet 1 — NumPy basics (illustrative tablet weights):

```
import numpy as np

weights = np.array([502, 498, 505, 500, 503, 497, 501, 499, 504, 500])
```

Line-by-line: `import numpy as np` loads the NumPy library and shortens its name to np; `np.array([...])` stores our 10 weights as a NumPy array (a fast, maths-ready list).

```
print("Mean:", np.mean(weights))        # average of the values
print("Median:", np.median(weights))    # middle value
print("SD (population):", np.std(weights))          # divide by n
print("SD (sample):", np.std(weights, ddof=1))      # divide by n-1
print("Range:", np.max(weights) - np.min(weights))  # max minus min
```

Sample output:

```
Mean: 500.9
Median: 500.5
SD (population): 2.4681748...
SD (sample): 2.60158...
Range: 8
```

Interpretation: The mean (500.9 mg) matches our hand calculation exactly — software confirms the maths. Note the two SD values differ because of the n vs n − 1 default: `ddof=1` means "delta degrees of freedom = 1", i.e. divide by (n − 1).

Snippet 2 — Pandas on a small QC table (illustrative):

```
import pandas as pd

qc = pd.DataFrame({
    "batch": ["A", "A", "A", "B", "B", "B"],
    "weight_mg": [502, 498, 500, 510, 490, 505]
})
print(qc.groupby("batch")["weight_mg"].agg(["mean", "std", "min", "max"]))
```

Line-by-line: `pd.DataFrame({...})` builds a table with two columns; `groupby("batch")` splits rows by batch; `["weight_mg"]` picks the weight column; `.agg([...])` computes four summaries per batch at once.

Sample output (values computed by running the code):

```
       mean        std  min  max
batch
A      500.0   2.00  498  502
B      501.67  10.41 490  510
```

Interpretation: Both batches average ~500 mg, but batch B's SD (10.4) is five times batch A's (2.0) — batch B is the inconsistent one. Same mean, different quality: exactly why dispersion matters.

> **Key takeaway:** `df.describe()` and `groupby` give you in one line what took a page of hand calculation — but you must still know what each number means.

1.9 Pharmacy relevance (Unit I)

- **QC:** Every batch-release decision (weight variation, content uniformity, dissolution) is a mean ± SD judgement against pharmacopoeial limits.
- **Clinical trials:** Baseline tables ("Table 1") in every trial report summarise age, BP, and lab values with mean ± SD or median — you must read them fluently.
- **Pharmacovigilance:** Counting ADR types (nominal data) and summarising time-to-onset (skewed, so median) are daily tasks.

1.10 Common confusions

1. **Mean vs median:** mean is the balance point; median is the middle person. They are equal only in symmetric data.
2. **Variance vs SD:** variance is in *squared* units (mg^2) and hard to interpret; SD is in original units (mg). Always report SD to humans.
3. **n vs n − 1:** sample statistics divide by n − 1. Pandas does this by default; NumPy does not.
4. **"Average" is not always the mean** — in skewed data the honest "average" is often the median.

1.11 Common mistakes

- Taking the mean of nominal data (e.g., "average blood group").
- Reporting only the mean without any dispersion measure.
- Using mean ± SD on heavily skewed data without mentioning the skew.
- Mixing up population and sample SD in software.

1.12 One-minute revision (Unit I)

- Data types: **Nominal** (names) → **Ordinal** (ranks) → **Interval** (no true zero) → **Ratio** (true zero, all maths OK).
- Pharmacy data sources: clinical trials, pharmacovigilance, QC, PK studies.
- Centre: mean (balance point), median (middle), mode (most frequent).
- Spread: range (max−min), variance (mean of squared deviations), SD = √variance.
- Sample variance divides by **n − 1**; Pandas `.std()` does this, NumPy `std()` needs `ddof=1`.
- Skew: mean **follows the tail**; skewed → trust the median.

- Python: NumPy for numbers, Pandas for tables; `groupby` + `agg` summarises by batch.

1.13 Practice questions (Unit I)

MCQs

1. Blood group (A, B, AB, O) is which type of data?

a) Nominal b) Ordinal c) Interval d) Ratio

2. Which measure is most appropriate for the "typical" hospital stay when stays are right-skewed?

a) Mean b) Median c) Mode d) Range

3. `pandas.Series.std()` divides the squared deviations by:

a) n b) n − 1 c) n − 2 d) √n

4. In a right-skewed distribution:

a) mean < median b) mean > median c) mean = median d) mode > mean

Short questions

5. Define nominal, ordinal, interval, and ratio data with one pharmacy example each.
6. A QC chemist reports "mean tablet weight 500 mg" for two batches. Why is this report incomplete?
7. Distinguish between population variance and sample variance.

Long / application-based

8. (Illustrative data) Dissolution (%) of 8 tablets: 92, 95, 88, 94, 91, 96, 93, 90. Calculate mean, median, range, sample variance and sample SD showing every step, and write a two-line QC interpretation.
9. Explain, with a pharmacy example, how the mean "follows the tail" in skewed data.
10. Write a Pandas snippet that reads a DataFrame of `batch` and `hardness_kg` and prints mean ± SD per batch; explain each line and what the output tells a QC officer.

Unit I infographic concept

- Data types: **Nominal** (names) → **Ordinal** (ranks) → **Interval** (no true zero) → **Ratio** (true zero, all maths OK).
- Pharmacy data sources: clinical trials, pharmacovigilance, QC, PK studies.
- Centre: mean (balance point), median (middle), mode (most frequent).
- Spread: range (max−min), variance (mean of squared deviations), SD = √variance.
- Sample variance divides by **n − 1**; Pandas `.std()` does this, NumPy `std()` needs `ddof=1`.
- Skew: mean **follows the tail**; skewed → trust the median.
- Python: NumPy for numbers, Pandas for tables; `groupby` + agg summarises by batch.

Combined references

1. Lane, D.M., *Introduction to Statistics*. Houston: Rice University. Available at: https://onlinestatbook.com/2/index.html
2. Walters, S.J., Campbell, M.J. and Machin, D., *Medical Statistics: A Textbook for the Health Sciences*. Hoboken, NJ: Wiley-Blackwell.
3. Rowe, P., *Essential Statistics for the Pharmaceutical Sciences*. London: Pharmaceutical Press.
4. Bolton, S. and Bon, C., *Pharmaceutical Statistics: Practical and Clinical Applications*. New York: Informa Healthcare.
5. De Muth, J.E., *Basic Statistics and Pharmaceutical Statistical Applications*. Boca Raton, FL: CRC Press.