B.PHARM SEMESTER II | BP201T

APPLIED BIOSTATISTICS AND DATA ANALYTICS

FOR PHARMACEUTICAL SCIENCES (THEORY)

UNIT I

Descriptive Statistics

Short Notes as per PCI B.Pharm Syllabus

What is covered in this unit

UNIT I TOPICS

Part A – Data: types of data (nominal, ordinal, interval, ratio), sources of data (clinical trials, pharmacovigilance, quality control, PK studies) and presentation of data.

Part B – Central tendency: mean, median, mode, their calculation and interpretation.

Part C – Dispersion: range, variance, standard deviation, CV and their interpretation with pharmaceutical data.

Part D – Shape: skewness, kurtosis and distribution shape in biological measurements.

Part E – Python: descriptive statistics with NumPy and Pandas, case studies, with focus on interpretation of results.

Part F – Practice: solved numerical problems and a list of formulae.

noteskarts.com | Noteskarts Prime App | YouTube: Noteskarts

Contents of Unit I

Part	Section	Topic
A	1	Introduction to biostatistics
A	2	Types of data in pharmaceutical sciences
A	3	Sources of data in pharmacy
A	4	Organisation and presentation of data
B	5	Arithmetic mean (with geometric mean)
B	6	Median and positional measures
B	7	Mode
B	8	Choosing and interpreting the right average
C	9	Range, variance, standard deviation and CV
D	10	Skewness, kurtosis and shape of distributions
D	11	Distribution shape in biological measurements
E	12	Getting started with Python
E	13	Descriptive statistics with NumPy
E	14	Descriptive statistics with Pandas
E	15	Skewness, outliers and graphs in Python
E	16	Case studies: Python analysis of pharmaceutical data
F	17	Solved numerical problems
F	18	Formulae used in this unit

Noteskarts — Smart Notes, Bright Future

PART A: BASICS OF DATA IN PHARMACEUTICAL SCIENCES

1. Introduction to Biostatistics

Statistics is the science of collecting, organising, presenting, analysing and interpreting numerical data so that sound decisions can be taken. **Biostatistics** is its application to biological, medical and health sciences. In pharmacy it is used in drug discovery, clinical trials, quality control, bioavailability and bioequivalence studies, pharmacovigilance and hospital practice.

Branches of statistics

Branch	Purpose	Pharmacy example
Descriptive statistics	Summarise and describe the data in hand using tables, graphs and numbers (mean, SD, etc.)	Average weight and SD of 20 tablets; percentage of patients with side effects
Inferential statistics	Draw conclusions about a large group (population) from a small group (sample); includes estimation and hypothesis testing	Is the drug better than placebo in all hypertensive patients? (Unit III)

This unit covers **descriptive statistics**, the first step of any data analysis. We always **describe** the data before we try to **infer** from it.

Basic terms

Term	Meaning	Example
Population	The complete group of interest	All tablets in a batch of 1 lakh; all diabetic patients in India
Sample	A smaller part of the population actually studied	20 tablets picked at random from the batch
Parameter	A number that describes a **population** (usually unknown); written with Greek letters: mean μ, SD σ	True mean weight μ of all tablets in the batch
Statistic	A number calculated from a **sample**; written with Roman letters: mean $\bar{x}$, SD s	Mean weight $\bar{x}$ of the 20 tablets = 500.8 mg
Variable	A characteristic that can take different values	Age, blood glucose, drug content
Observation / datum	One recorded value of a variable	One tablet weighing 502 mg
Raw data	Data as first collected, not yet organised	A list of 20 tablet weights
Frequency	Number of times a value (or class) occurs	500 mg appears 3 times

> **REMEMBER**
>
> **Parameter → Population (Greek letters). Statistic → Sample (English letters).** Both start with the same letter pattern: P for Parameter and Population, S for Statistic and Sample.

2. Types of Data in Pharmaceutical Sciences

Data may be classified in two ways: by **nature** (qualitative or quantitative) and by **level of measurement** (nominal, ordinal, interval, ratio). The type of data decides which graph and which statistic can be used.

A. Qualitative and quantitative data

Type	Description	Sub-type and examples
Qualitative (categorical)	Describes quality or category; not measured by numbers	Sex, blood group, tablet colour, dosage form, type of adverse reaction
Quantitative (numerical)	Measured or counted; arithmetic is meaningful	**Discrete:** whole numbers by counting (number of tablets, number of seizures, number of patients). **Continuous:** any value in a range, decimals possible (weight, BP, drug concentration, pH, time)

B. Scales of measurement

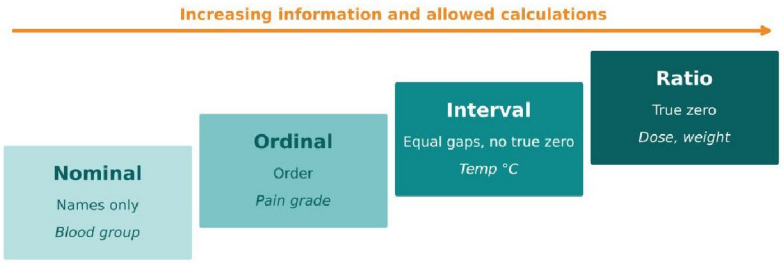

Figure 1: Four scales of measurement, from least to most informative

1. Nominal scale

- Data are **names or labels** with **no order**. Numbers, if used, are only codes (male = 1, female = 2).
- **Examples:** blood group (A, B, AB, O); route of administration (oral, IV, IM); dosage form (tablet, capsule, syrup); drug versus placebo group; adverse event (yes/no).
- **Statistics allowed:** counts, percentages, **mode**. Mean and median make no sense.
- A nominal variable with only two categories (yes/no, alive/dead) is called **binary or dichotomous**.

2. Ordinal scale

- Data have a **natural order or rank**, but the **gaps between ranks are not equal or not known**.
- **Examples:** severity of adverse reaction (mild < moderate < severe); pain score (none, mild, moderate, severe); stages of cancer (I to IV); NYHA heart failure class; Likert scale (strongly disagree to strongly agree); socio-economic class.
- **Statistics allowed:** counts, percentages, mode, **median**, percentiles. Mean is usually avoided because the gap between 'mild' and 'moderate' may not equal the gap between 'moderate' and 'severe'.

3. Interval scale

- Data have order and **equal gaps**, but there is **no true (absolute) zero**. Zero is only a convenient point, so ratios are not meaningful.
- **Examples:** temperature in °C or °F (40 °C is not 'twice as hot' as 20 °C); calendar year; some psychological or quality-of-life scores.
- **Statistics allowed:** mean, median, mode, SD, correlation and most tests.

4. Ratio scale

- Data have order, equal gaps and a **true zero** meaning 'nothing'. Ratios make sense (10 mg is twice 5 mg).
- **Examples:** weight, height, dose, drug concentration, Cmax, AUC, half-life, heart rate, blood glucose, tablet hardness, age, temperature in **kelvin**.
- **Statistics allowed:** everything, including the **geometric mean and coefficient of variation**.

Comparison of the four scales

Feature	Nominal	Ordinal	Interval	Ratio
Categories / names	Yes	Yes	Yes	Yes
Order or rank	No	**Yes**	Yes	Yes
Equal gaps	No	No	**Yes**	Yes
True zero	No	No	No	**Yes**
Mode	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Mean, SD	No	No (usually)	Yes	Yes
Pharmacy example	Blood group, dosage form	Pain grade, ADR severity	Temperature (°C)	Dose, weight, Cmax

QUICK MEMORY TRICK: NOIR

Nominal → **O**rdinal → **I**nterval → **R**atio. Each scale does everything the earlier one does, plus one more thing (order, then equal gaps, then true zero).

3. Sources of Data in Pharmacy

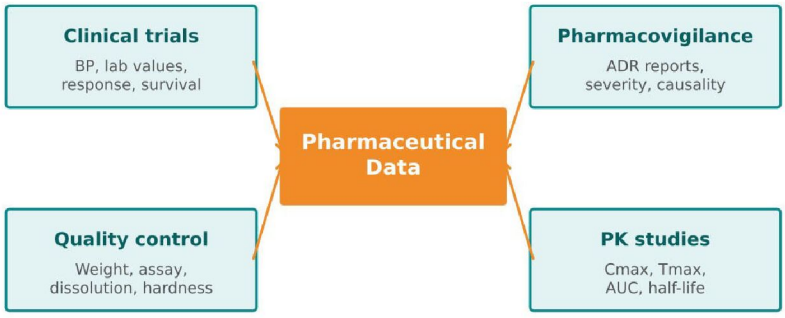

Figure 2: Main sources of data in the pharmaceutical sciences

1. Clinical trials

A **clinical trial** is a planned study on human volunteers to test the safety and efficacy of a drug, device or procedure. It is the richest source of pharmaceutical data.

Phase	Subjects	Main purpose	Typical data
Phase I	20–100 healthy volunteers (patients for cytotoxic drugs)	Safety, tolerability, dose range, pharmacokinetics	Adverse events, plasma levels, vital signs
Phase II	100–300 patients	Efficacy and best dose; further safety	Response rate, BP, HbA1c, biomarkers
Phase III	300–3000+ patients, often multi-centre randomised controlled trials	Confirm efficacy against standard treatment or placebo; compare side effects	Survival, relapse, symptom scores, adverse events
Phase IV	Thousands of patients after marketing	Long-term safety and rare adverse reactions (post-marketing surveillance)	Spontaneous reports, registry data

- **Data types found:** continuous (BP, glucose), ordinal (pain scale), nominal (responder yes/no) and time-to-event data (time to relapse).
- **Good Clinical Practice (GCP)** guidelines, ethics committee approval and informed consent control how the data are collected. Data are recorded in a **Case Record Form (CRF)**.
- **Randomisation, blinding and a control group** reduce bias so that differences in data can be attributed to the drug.

2. Pharmacovigilance

Pharmacovigilance (PV) is the science of detecting, assessing, understanding and preventing adverse effects of medicines. Its data come after the drug is in use.

- **Sources:** spontaneous (voluntary) ADR reports by doctors, pharmacists, nurses and patients; hospital records; cohort and case-control studies; registries; literature; and manufacturers' reports (periodic safety update reports, PSURs).
- **In India:** the **Pharmacovigilance Programme of India (PvPI)** run by the Indian Pharmacopoeia Commission (IPC), Ghaziabad, collects reports on a **Suspected Adverse Drug Reaction Reporting Form** through ADR Monitoring Centres. At the international level the WHO Uppsala Monitoring Centre keeps the **VigiBase** database.
- **Typical data:** patient details, suspected drug, dose, date of onset, type of reaction (nominal), **severity** (mild, moderate, severe: ordinal), **causality** (certain, probable, possible, unlikely: ordinal, WHO-UMC scale), outcome (recovered / not recovered / fatal: nominal).
- **Typical descriptive statistics:** number and percentage of reports by drug, age group or organ system; frequency of each severity grade.
- **Limitation:** spontaneous reports are **under-reported** and have no denominator (total number of users is unknown), so true incidence cannot be calculated.

3. Quality control (QC) and quality assurance

Pharmaceutical industry measures many properties of every batch to ensure that it meets pharmacopoeial standards. These measurements are **continuous ratio-scale data** collected repeatedly.

QC test	Data collected	What descriptive statistics tell
Weight variation	Weight of 20 tablets (mg)	Mean weight, how many tablets deviate more than the allowed limit
Content uniformity / assay	Drug content as % of label claim	Mean close to 100%, small SD or CV means uniform mixing
Dissolution	% drug released at each time point	Mean profile and variability between tablets or vessels

QC test	Data collected	What descriptive statistics tell
Hardness, friability, thickness, disintegration time	Newtons or kg/cm², % loss, mm, minutes	Consistency of compression force and granule quality
Fill volume / pH / viscosity	mL, pH units, centipoise	Batch-to-batch consistency
Microbial counts	Colony counts (discrete)	Median or geometric mean, since counts are often skewed

- **In-process control** and **control charts** (mean chart, range chart) use mean and range or SD over time to detect drift in the process.
- A small **SD or coefficient of variation** shows a well-controlled process; a large one shows poor control.

4. Pharmacokinetic (PK) studies

PK studies measure how the body handles a drug (absorption, distribution, metabolism, excretion). Blood samples are taken at fixed times after a dose and the drug concentration is measured.

PK parameter	Meaning	Data type
Cmax	Maximum (peak) plasma concentration	Continuous, ratio, often right-skewed
Tmax	Time at which Cmax occurs	Continuous (often reported as median)
AUC	Area under the plasma concentration-time curve; shows extent of absorption	Continuous, right-skewed
t½	Elimination half-life	Continuous
CL, Vd	Clearance and volume of distribution	Continuous

- Each volunteer gives one value of each parameter, so a **bioequivalence study** has a small data set (typically 12–36 subjects) summarised by **geometric mean** and CV.
- Because Cmax and AUC follow a skewed (log-normal) pattern, data are usually **log-transformed** before analysis (Part D).

4. Organisation and Presentation of Data

Raw data are difficult to read. We first **arrange** them and then **present** them as tables and graphs.

Array and frequency distribution

- **Array:** raw data arranged in ascending or descending order.
- **Frequency distribution:** a table showing how often each value (or class) occurs. For a **continuous** variable with many values, we group the data into **classes (class intervals)**.
- **Steps to make a grouped frequency table:** (1) find the range = largest − smallest; (2) choose 5 to 15 classes of **equal width** (class width h ≈ range ÷ number of classes); (3) write class limits without overlap; (4) put each value in its class by tally marks; (5) count the frequencies.

Term	Meaning
Class interval	A range of values, e.g. 490–500 mg
Class limits	Smallest (lower) and largest (upper) value of a class
Class width (h)	Upper limit − lower limit (for continuous classes such as 490–500, 500–510)
Mid-point (class mark)	(Lower limit + Upper limit) ÷ 2. Used to represent the class in calculations
Relative frequency	Frequency ÷ total number; × 100 gives **percentage frequency**
Cumulative frequency (cf)	Running total of frequencies. 'Less than' type adds from the lowest class

WORKED EXAMPLE 1: GROUPED FREQUENCY TABLE

Weights of 40 tablets from a batch were grouped as below. (Class 480–490 includes 480 up to but not including 490.)

Class (mg): 480–490 | 490–500 | 500–510 | 510–520 | 520–530

Frequency (f): 3 | 8 | 15 | 9 | 5 → total n = 40

Mid-point (x): 485 | 495 | 505 | 515 | 525

Cumulative frequency: 3 | 11 | 26 | 35 | 40

Relative frequency (%): 7.5 | 20 | 37.5 | 22.5 | 12.5

Interpretation: most tablets (15 of 40, 37.5%) lie in 500–510 mg, and 26 of 40 tablets (65%) weigh less than 510 mg. This table is used again in Part B and Part C.

Graphical presentation

Graph	Used for	Key points
Bar diagram	Nominal, ordinal or discrete data (e.g. number of ADRs by organ system)	Separate bars with gaps; height shows frequency
Pie chart	Parts of a whole (e.g. % of reports by severity)	Slices add to 100%; avoid with many categories
Histogram	Continuous data in classes (e.g. tablet weights)	**Bars touch each other**; area is proportional to frequency; shows shape of distribution
Frequency polygon	Compare two or more distributions	Join mid-points of the tops of histogram bars
Cumulative frequency curve (ogive)	Find median and percentiles graphically	Plot cumulative frequency against upper class limits
Line graph	Change over time (plasma concentration-time curve)	Time on the x-axis
Scatter plot	Relationship between two continuous variables (dose and response)	Each point is one subject (Unit IV)
Box-and-whisker plot	Median, spread and outliers; compare groups	Explained in Part D
Stem-and-leaf plot	Small data sets; shows shape and keeps actual values	Stem = leading digits, leaf = last digit

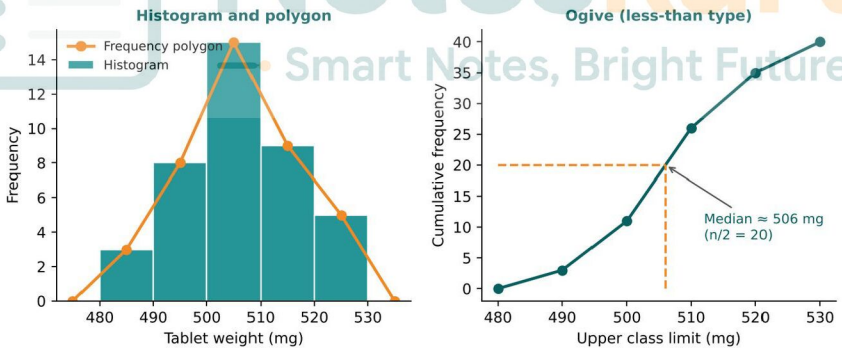

Figure 3: Histogram with frequency polygon, and a less-than ogive, drawn for the grouped tablet-weight data of Example 1

PART B: MEASURES OF CENTRAL TENDENCY

A **measure of central tendency** is a single value that represents the centre or typical value of a data set. The three main measures are the **mean, median and mode**. A good average should be easy to understand, rigidly defined, based on all observations, and not unduly affected by extreme values.

5. Arithmetic Mean

The **arithmetic mean** (average) is the sum of all observations divided by the number of observations. It is the most widely used measure for continuous ratio and interval data.

Ungrouped data

FORMULA

Sample mean: $\bar{x} = \Sigma x / n$ **Population mean:** $\mu = \Sigma x / N$

Σ (sigma) = sum of; x = each observation; n = number of observations in the sample.

WORKED EXAMPLE 2: MEAN OF TABLET WEIGHTS

Weights (mg) of 20 tablets: 498, 502, 505, 497, 500, 503, 499, 501, 496, 504, 500, 502, 498, 507, 495, 501, 499, 503, 500, 506.

$\Sigma x = 10016$ and $n = 20$.

$\bar{x} = 10016 / 20 = 500.8$ mg

Interpretation: the average tablet weight is 500.8 mg, very close to the target of 500 mg.

Grouped (frequency) data

- **Direct method:** $\bar{x} = \Sigma(f \cdot x) / n$, where x is the **mid-point** of each class, f its frequency and $n = \Sigma f$.
- **Short-cut (assumed mean) method:** choose an assumed mean A, class width h, and let $d = (x - A) / h$. Then **$\bar{x} = A + h \times \Sigma(f \cdot d) / n$**. This avoids large numbers.

WORKED EXAMPLE 3: MEAN FROM THE GROUPED TABLE (EXAMPLE 1)

Mid-points x: 485, 495, 505, 515, 525; frequencies f: 3, 8, 15, 9, 5; $n = 40$.

Direct method: $\Sigma fx = 1455 + 3960 + 7575 + 4635 + 2625 = 20250$, so **$\bar{x} = 20250 / 40 = 506.25$ mg**.

Short-cut method: $A = 505$, $h = 10$, $d = -2, -1, 0, 1, 2$; $\Sigma fd = -6 - 8 + 0 + 9 + 10 = 5$.

$\bar{x} = 505 + 10 \times 5 / 40 = 505 + 1.25 =$ **506.25 mg** (same answer).

Properties of the arithmetic mean

- The **sum of deviations** of all observations from the mean is **zero**: $\Sigma(x - \bar{x}) = 0$.
- The sum of squared deviations from the mean is **smaller than** from any other value (the 'least squares' property).
- If every value is increased by a constant, the mean increases by the same constant; if every value is multiplied by a constant, the mean is multiplied by it.
- It uses **every observation**, so it is **sensitive to extreme values (outliers)**.

Merits	Demerits
Easy to calculate and understand; uses all data; basis of SD, t-test, ANOVA and regression; best average for symmetrical data	Strongly pulled by outliers and skewed data; cannot be used for nominal data; cannot be found from open-ended classes; may not be an actual observed value (e.g. 2.4 children)

Geometric mean (GM)

- **GM = n-th root of ($x_1 \times x_2 \times \ldots \times x_n$)**, or **GM = antilog of the mean of the logarithms**: $GM = \exp[\ \Sigma \ln(x) / n\ ]$.
- Used for data that grow by **ratios or multiplication**, or are **right-skewed (log-normal)**: bacterial counts, antibody titres, **Cmax and AUC in bioequivalence**, growth rates, dilution series.
- Always smaller than or equal to the arithmetic mean (for positive values).

WORKED EXAMPLE 5: GEOMETRIC MEAN

Bacterial counts in three samples: 100, 1000 and 10000 colonies.

Arithmetic mean = 11100 / 3 = 3700. **Geometric mean = (100 × 1000 × 10000)^(1/3) = 1000.**

The GM (1000) represents the middle of the 'ten-fold' series, whereas the arithmetic mean (3700) is dragged up by the largest value.

6. Median and Other Positional Measures

The **median** is the **middle value** when the observations are arranged in order. Half the data lie below it and half above it. It is **not affected by extreme values**, so it is the best average for skewed data and for **ordinal data**.

Ungrouped data

RULE

Arrange the data in ascending order. If **n is odd**, median = the **((n + 1) / 2)-th** value. If **n is even**, median = the **average of the (n/2)-th and (n/2 + 1)-th** values.

WORKED EXAMPLE 6: ODD AND EVEN N

Half-lives (h) in five volunteers: 4.2, 5.1, 3.8, 6.0, 4.9. Arranged: 3.8, 4.2, **4.9**, 5.1, 6.0. n = 5 (odd), so the median is the 3rd value = **4.9 h**.

Add a sixth volunteer with 5.5 h. Arranged: 3.8, 4.2, 4.9, 5.1, 5.5, 6.0. n = 6 (even), so median = (4.9 + 5.1) / 2 = **5.0 h**.

Grouped data

FORMULA

Median = L + [(n/2 − cf) / f] × h

L = lower limit of the **median class** (the class containing the (n/2)-th value); cf = cumulative frequency of the class **before** the median class; f = frequency of the median class; h = class width; n = total frequency.

WORKED EXAMPLE 7: MEDIAN FROM THE GROUPED TABLE (EXAMPLE 1)

n/2 = 20. Cumulative frequencies: 3, 11, 26, 35, 40, so the 20th value falls in the class **500–510** (median class).

L = 500, cf = 11, f = 15, h = 10.

Median = 500 + [(20 − 11) / 15] × 10 = 500 + 6 = 506 mg

Quartiles, deciles and percentiles

- **Quartiles** divide ordered data into **4 equal parts**: Q1 (25th percentile), Q2 (= median, 50th percentile), Q3 (75th percentile).
- **Deciles** divide into 10 parts; **percentiles** into 100 parts. The **P-th percentile** is the value below which P% of observations fall.
- **Pharmacy use:** paediatric growth charts (a child at the 50th percentile of weight), reference ranges of lab tests (2.5th to 97.5th percentile), and **Q1 and Q3 in box plots**.
- **Interquartile range (IQR) = Q3 − Q1** (Unit I, Part C).

NOTE: DIFFERENT METHODS GIVE SLIGHTLY DIFFERENT QUARTILES

For the 20 tablet weights, hand calculation using the (n + 1)/4 position gives Q1 = 498.25, whereas Python (NumPy/Pandas, linear interpolation) gives **Q1 = 498.75**. Both are acceptable; always state the method or software used. The median (500.5) is the same in every method.

Merits of the median	Demerits
Not affected by outliers or skewness; can be used for ordinal data and open-ended classes; easy to find graphically from an ogive	Does not use every value; less suitable for further statistical calculation; needs arranging the data

7. Mode

The **mode** is the value that occurs **most often** (highest frequency). It is the only measure of central tendency that can be used for **nominal data**.

- **Ungrouped data:** find the value with the highest frequency. In the 20 tablet weights, **500 mg** occurs 3 times, more than any other value, so mode = 500 mg.
- A data set may have **no mode** (all values different), one mode (**unimodal**), two modes (**bimodal**) or many (**multimodal**). A bimodal histogram often means **two different groups are mixed** (e.g. fast and slow metabolisers of a drug, or two tablet batches).
- **Nominal example:** in a list of 50 ADR reports, 'rash' reported by 18 patients is the most common reaction, so 'rash' is the modal category.

Grouped data

FORMULA

Mode = L + [$(f_1 - f_0)$ / $(2f_1 - f_0 - f_2)$] × h

L = lower limit of the **modal class** (class with highest frequency); f_1 = its frequency; f_0 = frequency of the class before it; f_2 = frequency of the class after it; h = class width.

WORKED EXAMPLE 8: MODE FROM THE GROUPED TABLE (EXAMPLE 1)

Highest frequency is 15, so the modal class is **500–510**. L = 500, f_1 = 15, f_0 = 8, f_2 = 9, h = 10.

Mode = 500 + [(15 − 8) / (30 − 8 − 9)] × 10 = 500 + (7/13) × 10 = 505.4 mg

Relation between mean, median and mode

- For a **moderately skewed** distribution: **Mode = 3 × Median − 2 × Mean** (Karl Pearson's empirical relation).
- Check with Example 1: 3 × 506 − 2 × 506.25 = 505.5 mg, which is close to the calculated mode 505.4 mg.

Merits of the mode	Demerits
Easy to understand; not affected by extreme values; usable for nominal data; shows the most typical or popular value	May be absent or not unique; not based on all observations; unstable in small samples; hard to use in further calculations

8. Choosing and Interpreting the Right Average

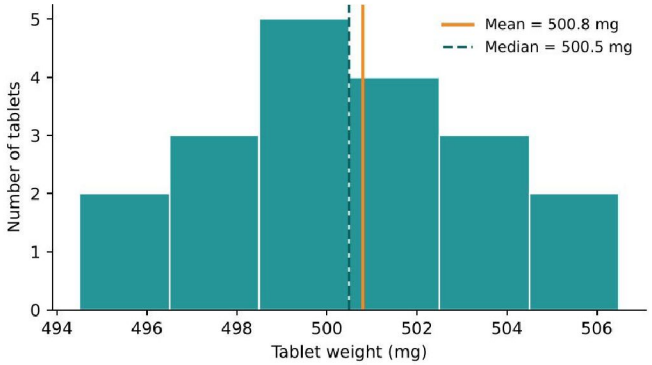

Figure 4: Histogram of the 20 tablet weights; mean (500.8 mg) and median (500.5 mg) are almost equal, showing a nearly symmetrical distribution

Comparison of mean, median and mode

Feature	Mean	Median	Mode
Definition	Sum ÷ number	Middle value	Most frequent value
Uses all values	Yes	No (position only)	No
Affected by outliers	**Very much**	Little	No
Data scale	Interval, ratio	Ordinal, interval, ratio	**All, including nominal**
Best for	Symmetrical data	**Skewed data**, ordinal data	Categories; most common value
Pharmacy example	Mean weight of tablets; mean fall in BP	Median survival time; median Tmax; median pain score	Most common ADR; most common blood group

Which average should I use?

Situation	Best measure
Symmetrical continuous data (tablet weight, height, Hb)	**Mean** (with SD)
Skewed data (income, hospital stay, survival time, Cmax)	**Median** (with IQR); or **geometric mean** for log-normal data
Ordinal data (pain score, ADR severity)	**Median** (and mode)
Nominal data (blood group, dosage form)	**Mode** (and percentages)
Ratio and growth data (bacterial counts, titres)	**Geometric mean**
Combining several groups of unequal size	**Weighted mean**

POSITION OF THE THREE AVERAGES TELLS YOU THE SHAPE

Mean ≈ Median ≈ Mode → symmetrical. **Mean > Median > Mode** → right (positive) skew. **Mean < Median < Mode** → left (negative) skew. Details in Part D.

PART C: MEASURES OF DISPERSION

An average alone does not tell the whole story. Two batches of tablets can both have a mean weight of 500 mg, but in one batch every tablet weighs 499–501 mg while in the other the weights range from 470 to 530 mg. **Dispersion (variability or spread)** tells how much the observations differ from each other and from the average. **A small spread means consistent, reliable data; a large spread means unreliable or varied data.**

Measure	Type	Symbol / formula
Range	Absolute	Largest − Smallest
Interquartile range and quartile deviation	Absolute	IQR = Q3 − Q1; QD = (Q3 − Q1)/2
Mean deviation	Absolute	$\Sigma\|x - \bar{x}\| / n$
Variance	Absolute	σ^2 (population), s^2 (sample)
Standard deviation	Absolute	σ (population), s (sample)
Standard error of the mean	Absolute	$SEM = s / \sqrt{n}$
Coefficient of variation	**Relative** (no units)	$CV\% = (s / \bar{x}) \times 100$

9. Range, Variance, Standard Deviation and CV

1. Range

- **Range = Largest value − Smallest value.** For the 20 tablet weights: 507 − 495 = **12 mg**.
- **Merits:** simplest to calculate; used in quality control (range charts) and for quick judgement.
- **Demerits:** depends on only two values; badly affected by outliers; ignores the data in between; increases with sample size.

2. Interquartile range and quartile deviation

- **IQR = Q3 − Q1** covers the **middle 50%** of the data. It is the measure of spread that goes with the **median** and is not affected by outliers.
- **Quartile deviation (semi-interquartile range) = (Q3 − Q1) / 2. Coefficient of quartile deviation = (Q3 − Q1)/(Q3 + Q1).**
- For the 20 tablet weights (Python method) Q1 = 498.75 and Q3 = 503.0, so **IQR = 4.25 mg**.

4. Variance

Variance is the average of the squared deviations from the mean.

	Population	Sample
Symbol	σ^2 (sigma squared)	s^2
Formula	$\sigma^2 = \Sigma(x - \mu)^2 / N$	$s^2 = \Sigma(x - \bar{x})^2 / (n - 1)$
Divisor	N (all members)	**n − 1** (degrees of freedom)
Use	When every member of the population is measured	When data are a sample (almost always)

> **WHY DIVIDE BY (N − 1) FOR A SAMPLE?**
>
> A sample is, on average, closer to its own mean than to the true population mean, so dividing by n would **underestimate** the true variance. Dividing by **n − 1** (**Bessel's correction**) removes this bias. **n − 1 is called the degrees of freedom.** For large n, the difference becomes tiny.

- **Short-cut (computational) formula:** $s^2 = [\ \Sigma x^2 - (\Sigma x)^2/n\ ] / (n - 1)$. It avoids calculating each deviation.
- The unit of variance is the **square** of the unit of the data (mg^2, $mmHg^2$), which is difficult to interpret. Hence SD is preferred.

5. Standard deviation (SD)

Standard deviation = square root of the variance. It is expressed in the **same unit** as the data and is the most widely used measure of spread. $\sigma = \sqrt{[\ \Sigma(x - \mu)^2/N\ ]}$; $s = \sqrt{[\ \Sigma(x - \bar{x})^2/(n - 1)\ ]}$.

WORKED EXAMPLE 10: VARIANCE AND SD (STEP BY STEP)

Half-lives (h) of five volunteers: 4.2, 5.1, 3.8, 6.0, 4.9. Mean $\bar{x}$ = 24.0 / 5 = 4.8.

Deviations ($x - \bar{x}$): −0.6, +0.3, −1.0, +1.2, +0.1 (they add to 0).

Squared deviations: 0.36, 0.09, 1.00, 1.44, 0.01. Sum = **2.90**.

Sample variance s^2 = 2.90 / (5 − 1) = 0.725 h^2; SD s = √0.725 = 0.85 h.

(Population formula would give 2.90/5 = 0.58 and σ = 0.76 h; use it only if these five were the whole population.)

Range = 6.0 − 3.8 = 2.2 h; Mean deviation = 3.2 / 5 = 0.64 h.

WORKED EXAMPLE 11: SD OF THE 20 TABLET WEIGHTS

$\bar{x}$ = 500.8 mg; sum of squared deviations = 201.2.

s^2 = 201.2 / 19 = 10.59 mg^2; s = 3.25 mg. (This matches the Python output in Part E.)

Interpretation: a typical tablet differs from the average weight by about 3.3 mg, which is a tiny 0.65% of 500 mg: the compression process is very consistent.

SD of grouped data

$s = \sqrt{[\ \Sigma f(x - \bar{x})^2 / (n - 1)\ ]}$, where x is the class mid-point and f the frequency.

WORKED EXAMPLE 12: SD FROM THE GROUPED TABLE (EXAMPLE 1)

$\bar{x}$ = 506.25. Deviations ($x - \bar{x}$): −21.25, −11.25, −1.25, 8.75, 18.75. Frequencies 3, 8, 15, 9, 5.

$\Sigma f(x - \bar{x})^2$ = 3(451.56) + 8(126.56) + 15(1.56) + 9(76.56) + 5(351.56) = 1354.7 + 1012.5 + 23.4 + 689.1 + 1757.8 = **4837.5**.

s^2 = 4837.5 / 39 = 124.04 mg^2; s = 11.14 mg. (Using n instead of n − 1 gives σ = 11.00 mg.)

Properties of standard deviation

- SD is **never negative**; it is **zero** only when all values are identical.
- **Adding or subtracting** a constant to every value **does not change** SD. **Multiplying** every value by a constant multiplies SD by that constant (e.g. converting mg to g divides SD by 1000).
- It uses all observations and is the basis of the normal distribution, confidence intervals and hypothesis tests.
- It is sensitive to outliers, because deviations are squared.

The empirical rule (68–95–99.7 rule)

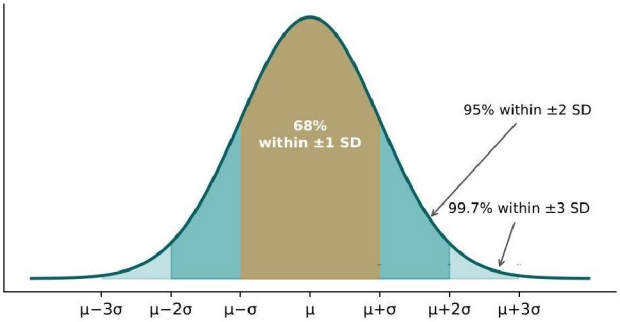

Figure 5: For bell-shaped (normal) data about 68%, 95% and 99.7% of values lie within 1, 2 and 3 SD of the mean

Range	Expected % of data	For tablets (mean 500.8, s = 3.25)	Observed (20 tablets)
$\bar{x}$ ± 1 SD	About 68%	497.5 to 504.1 mg	14 of 20 = 70%
$\bar{x}$ ± 2 SD	About 95%	494.3 to 507.3 mg	20 of 20 = 100%
$\bar{x}$ ± 3 SD	About 99.7%	491.0 to 510.6 mg	20 of 20 = 100%

The rule works well only for **roughly symmetrical, bell-shaped** data. A useful 'rough estimate' is that **SD ≈ Range / 4** for a reasonably large sample.

6. Standard error of the mean (SEM)

- **SEM = s / √n.** It measures how precisely the **sample mean** estimates the true mean. For the tablets: 3.25 / √20 = **0.73 mg**.

- **SD describes the spread of individual observations; SEM describes the uncertainty of the mean.** SEM becomes smaller as the sample gets larger, SD does not.
- Use **SD** to describe data and **SEM or confidence interval** to show the precision of the mean (Unit II/III).

7. Coefficient of variation (CV) / Relative standard deviation (RSD)

FORMULA

CV (%) = (SD / Mean) × 100

CV has **no units**, so it allows comparison of variability between data sets with **different units or different means**.

WORKED EXAMPLE 13: COMPARING TWO PRODUCTS

Product A: assay mean 99.5% label claim, SD = 1.2%. CV = (1.2 / 99.5) × 100 = **1.21%**.

Product B: mean content 50 mg, SD = 1.0 mg. CV = (1.0 / 50) × 100 = **2.0%**.

Although B has the smaller SD (1.0 versus 1.2), its **relative variability is higher**, so B is the less uniform product. The SDs could not be compared directly because the units and means differ.

- **Uses in pharmacy:** precision of an analytical method (RSD of replicate injections in HPLC is usually required to be small, often not more than 2%); variability in bioequivalence (**intra-subject CV**); batch consistency; comparing variability of weight vs hardness.
- **Rough guide:** CV below 5% means very consistent; 5–15% moderate; above 20–30% highly variable (acceptable limits depend on the test).
- CV is valid only for **ratio-scale data** with a true zero and positive mean; it is meaningless for temperature in °C.

Summary of dispersion measures

Measure	Best used with	Affected by outliers?	Remark
Range	Quick look, QC charts	**Yes, very much**	Uses only two values
IQR	Median, skewed data	No	Middle 50% of data; used in box plots
Variance	Statistical theory	Yes	Unit is squared
SD	**Mean, symmetrical data**	Yes	Most common; same unit as data
SEM	Precision of the mean	Yes	Falls as n rises
CV%	Comparing variability across groups or units	Yes	No unit; ratio data only

REPORTING RULE

Report **mean ± SD** for symmetrical data and **median (IQR)** or **median (min–max)** for skewed or ordinal data. Example: 'Age 50.0 ± 8.0 years'; 'Pain score median 3 (IQR 2–4)'.

PART D: SKEWNESS AND SHAPE OF BIOLOGICAL DATA

10. Skewness, Kurtosis and Shape of Distributions

The **shape** of a distribution (seen in a histogram) tells whether the mean and SD are good summaries. Shape is described by **symmetry (skewness)** and **peakedness (kurtosis)**.

The normal (Gaussian) distribution

- A **symmetrical, bell-shaped curve** fully described by its **mean (μ)** and **SD (σ)**. Mean = Median = Mode at the centre.
- About **68%, 95% and 99.7%** of values lie within ±1, ±2 and ±3 SD (empirical rule, Figure 5). The total area under the curve is 1 (100%).
- **Examples of near-normal biological data:** height, tablet weight in a well-controlled process, haemoglobin in healthy adults, systolic BP in a population.
- The **z-score** = (x − mean) / SD shows how many SD a value is from the mean; |z| > 3 flags a possible outlier.
- Many statistical tests (Units II to IV) assume that data are approximately normal, so checking shape is an important first step.

Skewness

Skewness measures the **lack of symmetry**. The side with the **long tail** gives the name of the skew.

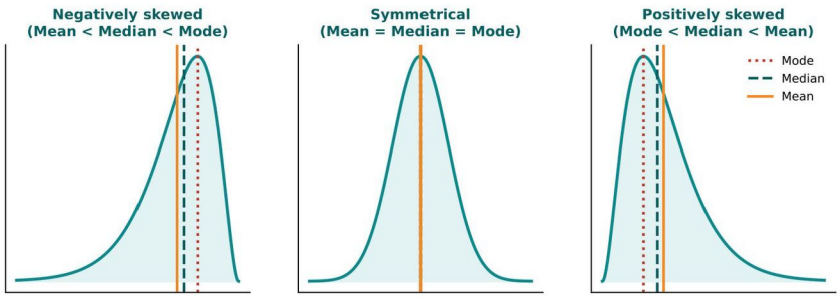

Figure 7: Position of mean, median and mode in negatively skewed, symmetrical and positively skewed distributions

Feature	Positive (right) skew	Symmetrical	Negative (left) skew
Long tail	On the **right** (high values)	None	On the **left** (low values)
Order of averages	**Mode < Median < Mean**	Mean = Median = Mode	**Mean < Median < Mode**
Skewness value	Greater than 0	About 0	Less than 0
Mean is pulled…	Towards the high tail	–	Towards the low tail
Biological examples	Cmax, AUC, drug concentrations, hospital stay, bacterial counts, triglycerides, income, survival time, ADR counts	Height, tablet weight, Hb	Age at death in developed countries, score of an easy exam, % assay close to an upper cap

EASY WAY TO REMEMBER

The **mean is always dragged towards the tail**. In positive skew, the tail is on the right, so mean > median. Because the mean is unstable with skew, the **median** is a safer summary.

Measures of skewness

Measure	Formula	Remarks
Karl Pearson's first coefficient	Sk = (Mean − Mode) / SD	Use when a clear mode exists
Karl Pearson's second coefficient	Sk = 3 × (Mean − Median) / SD	Used when mode is ill-defined; value lies between −3 and +3
Bowley's (quartile) coefficient	Sk = (Q3 + Q1 − 2 × Median) / (Q3 − Q1)	Based on quartiles; not affected by outliers; lies between −1 and +1
Moment coefficient (g_1)	$g_1 = (1/n)\,\Sigma[(x - \bar{x})/s]^3$ (adjusted form divides by (n−1)(n−2))	What software (NumPy, Pandas, SciPy) reports

WORKED EXAMPLE 17: SKEWNESS OF THE 20 TABLET WEIGHTS

Mean = 500.8, Median = 500.5, SD = 3.254, Q1 = 498.75, Q3 = 503.0.

Pearson (second): 3 × (500.8 − 500.5) / 3.254 = **0.28**. **Bowley:** (503.0 + 498.75 − 2 × 500.5) / 4.25 = **0.18**. **Moment (Pandas): 0.16**.

All the values are small and positive, so the distribution is **almost symmetrical** (a very slight right tail). Mean ≈ median confirms this. The mean and SD are suitable summaries.

How to interpret the skewness value from software

Skewness	Interpretation	What to do
Between −0.5 and +0.5	Approximately symmetrical	Mean and SD are fine
−1 to −0.5 or +0.5 to +1	Moderately skewed	Check histogram; consider median or a transformation
Below −1 or above +1	Highly skewed	Prefer median (IQR) or log-transform; avoid mean ± SD alone

These cut-offs are **rules of thumb**, not strict laws. With a small sample (below about 30), skewness values are unstable, so always look at the graph as well.

Kurtosis

Kurtosis describes the **peakedness and tail weight** of a distribution compared with the normal curve. Software usually reports **excess kurtosis** (= kurtosis − 3), so the normal curve has 0.

Type	Excess kurtosis	Shape	Meaning
Mesokurtic	≈ 0	Normal bell curve	Standard peak and tails
Leptokurtic	> 0	Tall, thin peak with heavy tails	More extreme values and more concentration near the centre
Platykurtic	< 0	Flat top, light tails	Values are spread more evenly; fewer extremes

For the tablet weights the excess kurtosis is −0.68 (slightly platykurtic), a normal finding for a small sample.

11. Distribution Shape in Biological Measurements

Log-normal data and the geometric mean in pharmacokinetics

- **Cmax and AUC** are positively skewed and roughly **log-normal**. Bioequivalence guidelines therefore **log-transform** these parameters (ln Cmax, ln AUC), analyse them, and then report the **geometric mean and the ratio of geometric means** (test/reference), with the accepted 90% confidence limit of **80–125%**.
- **Procedure:** take ln of each value → compute mean and SD of ln values → **GM = e^(mean of ln values)**.
- Log transformation often turns a right-skewed histogram into a nearly symmetrical one and also **stabilises the variance**.

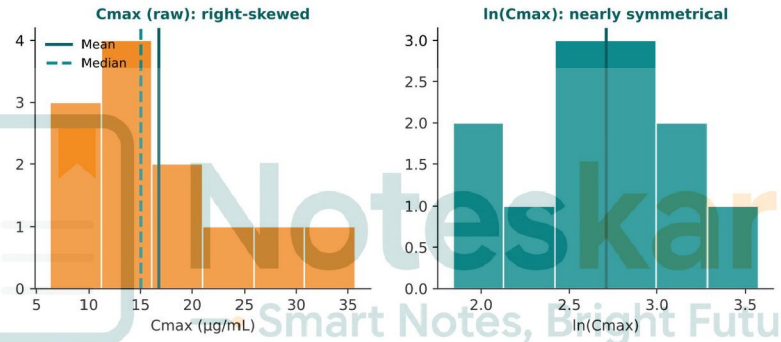

Figure 8: Cmax values of 12 volunteers: right-skewed on the original scale, nearly symmetrical after log transformation

WORKED EXAMPLE 18: SKEWED CMAX DATA

Cmax (µg/mL) of 12 volunteers: 35.7, 17.6, 10.0, 26.7, 6.3, 8.4, 12.8, 15.8, 22.5, 14.2, 11.4, 19.7.

Mean = 16.76; Median = 15.0; Skewness = 1.05 (positive, high). The mean is above the median because of the largest value (35.7).

After log transformation, skewness = **0.00** and the **geometric mean = 14.99 µg/mL**, close to the median.

Conclusion: the data are right-skewed and log-normal; the **median or geometric mean** represents a typical volunteer better than the arithmetic mean.

Outliers and the box-and-whisker plot

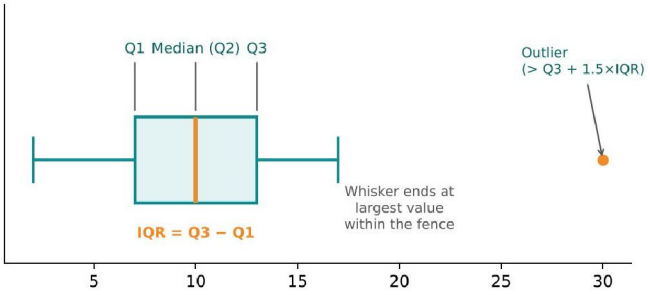

Figure 9: Anatomy of a box-and-whisker plot

- The **box** spans **Q1 to Q3** (IQR); the line inside is the **median**; the **whiskers** reach the most extreme values **inside the fences**; points beyond the fences are plotted separately as **outliers**.
- **Tukey's rule (1.5 × IQR):** lower fence = Q1 − 1.5 × IQR; upper fence = Q3 + 1.5 × IQR. A value outside the fences is a **possible outlier**.

- **Reading a box plot:** the **median position inside the box** shows skew (median closer to Q1 with a longer upper whisker = right skew); a **long box** = large spread; **two boxes** side by side compare groups.

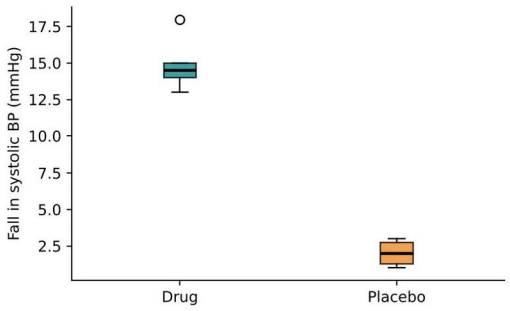

Figure 10: Fall in systolic BP in the Drug and Placebo groups (12 patients): the drug group shows a larger fall and little overlap

WORKED EXAMPLE 19: FINDING OUTLIERS

For the 12 Cmax values: Q1 = 11.1 and Q3 = 20.4 (Pandas method), so IQR = 9.3. Upper fence = 20.4 + 1.5 × 9.3 ≈ **34.4**. The largest value, **35.7**, lies above the fence and is flagged as a possible outlier.

Action: check the original record. It may be a true high absorber (keep it, report it) or a data error (correct it). **Never delete a value only to improve the result.**

Choosing summary statistics by shape

Shape of data	Centre	Spread	Graph
Symmetrical (normal-like)	Mean	SD (or CV)	Histogram, box plot
Skewed or with outliers	Median	IQR (or range)	Box plot, histogram
Log-normal (PK data)	Geometric mean	Geometric CV or SD of ln values	Histogram of ln values
Ordinal	Median, mode	IQR	Bar chart
Nominal	Mode	Percentages	Bar chart, pie chart

PART E: DESCRIPTIVE STATISTICS USING PYTHON (NUMPY AND PANDAS)

Hand calculation is useful for learning, but real pharmaceutical data sets have hundreds or thousands of rows. **Python** is a free, beginner-friendly programming language, and its libraries **NumPy** (fast numerical arrays) and **Pandas** (tables called DataFrames) calculate descriptive statistics in one line. The most important skill is not writing code but **interpreting the output**.

12. Getting Started with Python

The main libraries

Library	Import statement	Use in descriptive statistics
NumPy	import numpy as np	Arrays and fast mathematical functions: np.mean, np.median, np.std, np.var, np.percentile
Pandas	import pandas as pd	DataFrames and Series; reading CSV/Excel; .describe(), .groupby(), .value_counts(), .skew()
SciPy (stats)	from scipy import stats	stats.skew, stats.kurtosis, stats.mode, many tests (Unit III)
Matplotlib	import matplotlib.pyplot as plt	Histogram, box plot and other graphs
statistics (built in)	import statistics as st	st.mean, st.median, st.multimode, st.stdev for small lists

PYTHON BASICS IN FIVE LINES

starts a comment (ignored by Python). **print()** shows output. **=** stores a value in a variable. **[ ]** makes a list. **Indexing starts at 0**: weights[0] is the first value. Use **round(value, 2)** to limit decimal places. Python is **case-sensitive** (Mean and mean are different).

13. Descriptive Statistics with NumPy

A NumPy **array** holds a list of numbers. Methods such as .mean() or functions such as np.median() calculate statistics for the whole array. The program below analyses the 20 tablet weights used throughout this unit.

PYTHON CODE

```
import numpy as np

weights = np.array([498, 502, 505, 497, 500, 503, 499, 501, 496, 504,
                    500, 502, 498, 507, 495, 501, 499, 503, 500, 506])

values, counts = np.unique(weights, return_counts=True)

print("n (tablets)      :", weights.size)
print("Mean (mg)        :", weights.mean())
print("Median (mg)      :", np.median(weights))
print("Mode (mg)        :", values[counts.argmax()])
print("Range (mg)       :", weights.max() - weights.min())
print("Variance (n-1)   :", round(np.var(weights, ddof=1), 3))
print("Std deviation    :", round(np.std(weights, ddof=1), 3))
print("CV (%)           :", round(np.std(weights, ddof=1) / weights.mean() * 100, 3))
q1, q3 = np.percentile(weights, [25, 75])
print("Q1, Q3           :", q1, q3)
print("IQR              :", q3 - q1)
```

OUTPUT

```
n (tablets)      : 20
Mean (mg)        : 500.8
Median (mg)      : 500.5
Mode (mg)        : 500
Range (mg)       : 12
Variance (n-1)   : 10.589
Std deviation    : 3.254
CV (%)           : 0.65
Q1, Q3           : 498.75 503.0
IQR              : 4.25
```

Noteskarts — Smart Notes, Bright Future

Reading the output

Result	How to interpret
Mean 500.8 and Median 500.5	Almost equal, so the distribution is nearly symmetrical; average tablet is on target (500 mg).
Mode 500	Most frequent weight is 500 mg (appears 3 times).
Range 12; SD 3.254	Weights vary within a narrow band; a typical tablet is about 3.3 mg from the mean.
CV 0.65%	Very low relative variability: excellent uniformity.
Q1 498.75, Q3 503.0, IQR 4.25	The middle half of the tablets weigh between 498.75 and 503 mg.

IMPORTANT: DDOF (THE N − 1 ISSUE)

np.var() and np.std() divide by n by default (ddof = 0, population formula). For sample data you must write **ddof=1** to divide by n − 1, as done in the program.

Pandas .std() and .var() use n − 1 by default. This difference is the most common reason for mismatching answers.

np.var(weights) gives 10.06 (÷ n), but np.var(weights, ddof=1) gives **10.589** (÷ n − 1).

14. Descriptive Statistics with Pandas

A **DataFrame** is a table in which each **row** is a patient or sample and each **column** is a variable. Data are normally read from a file.

PYTHON CODE

```
import pandas as pd

# Reading data from a file (CSV or Excel) saved on your computer or Colab
df = pd.read_csv("trial_data.csv")        # for Excel: pd.read_excel("trial_data.xlsx")

df.head()          # first 5 rows
df.shape           # (number of rows, number of columns)
df.info()          # column names, data types, missing values
df.describe()      # count, mean, SD, min, quartiles, max of numeric columns
```

Example: a small clinical trial data set

The data below are for 12 hypertensive patients: 6 received the drug and 6 received placebo. The code finds counts, summary statistics and group-wise results.

PYTHON CODE

```
import pandas as pd

data = {
    "patient":  list(range(1, 13)),
    "group":    ["Drug"] * 6 + ["Placebo"] * 6,
    "age":      [45, 52, 38, 60, 49, 55, 47, 51, 42, 58, 63, 40],
    "sbp_before": [152, 160, 148, 165, 155, 158, 154, 159, 150, 162, 166, 149],
    "sbp_after":  [138, 142, 135, 150, 140, 144, 152, 156, 149, 160, 163, 148],
    "adverse_event": ["No", "Yes", "No", "No", "Yes", "No",
                      "No", "No", "Yes", "No", "No", "No"],
}
df = pd.DataFrame(data)
df["sbp_fall"] = df["sbp_before"] - df["sbp_after"]

print(df.head(3))
print()
print(df["adverse_event"].value_counts())
print()
print(df[["age", "sbp_before", "sbp_after", "sbp_fall"]].describe().round(2))
print()
print(df.groupby("group")["sbp_fall"].agg(["count", "mean", "median", "std"]).round(2))
```

OUTPUT

```
   patient group  age  sbp_before  sbp_after adverse_event  sbp_fall
0        1  Drug   45         152        138            No        14
1        2  Drug   52         160        142           Yes        18
2        3  Drug   38         148        135            No        13

adverse_event
No     9
Yes    3
Name: count, dtype: int64

         age  sbp_before  sbp_after  sbp_fall
count  12.00       12.00      12.00     12.00
mean   50.00      156.50     148.08      8.42
std     8.01        6.13       8.73      6.83
min    38.00      148.00     135.00      1.00
25%    44.25      151.50     141.50      2.00
50%    50.00      156.50     148.50      8.00
75%    55.75      160.50     153.00     14.25
max    63.00      166.00     163.00     18.00
```

```
         count   mean  median   std
group
Drug         6  14.83    14.5  1.72
Placebo      6   2.00     2.0  0.89
```

Reading the output

Output	Meaning and interpretation
value_counts() for adverse_event	9 patients had no adverse event and 3 had one: **25% reported an adverse event** (nominal data: use counts and percentages, not the mean).
describe(): count 12	No missing values (all four columns have 12 observations).
mean 50.00, std 8.01 for age	Average age is 50 years; ages are spread about 8 years around the mean.
min 38, 25% 44.25, 50% 50.0, 75% 55.75, max 63	Youngest 38, oldest 63; the middle half are aged 44 to 56. Median = mean = 50, so age is symmetrical.
sbp_before mean 156.5	All patients were hypertensive at the start.
sbp_fall: mean 8.42, median 8.00, std 6.83	Overall mean fall is 8.4 mmHg, but the large SD hints that the two groups behave differently.
groupby: Drug mean 14.83, Placebo mean 2.00	Drug group had a **much larger fall** (14.8 mmHg versus 2.0 mmHg). SDs (1.72 and 0.89) are small, so the response in each group is consistent.

CAUTION

Descriptive statistics only **describe** these 12 patients. A **large difference in means does not prove** that the drug works in all patients: that conclusion needs a hypothesis test such as the t-test (Unit III).

15. Skewness, Outliers and Graphs in Python

Skewness and log transformation

The program below uses the 12 Cmax values to check skewness, calculate the geometric mean and find outliers with Tukey's 1.5 × IQR rule.

PYTHON CODE

```
import numpy as np
import pandas as pd

cmax = pd.Series([35.7, 17.6, 10.0, 26.7, 6.3, 8.4, 12.8, 15.8, 22.5, 14.2, 11.4, 19.7])

print("Mean              :", round(cmax.mean(), 2))
print("Median            :", round(cmax.median(), 2))
print("Skewness          :", round(cmax.skew(), 2))

log_cmax = np.log(cmax)
print("Skewness (log)    :", round(log_cmax.skew(), 2))
print("Geometric mean    :", round(np.exp(log_cmax.mean()), 2))

q1, q3 = cmax.quantile([0.25, 0.75])
upper_fence = q3 + 1.5 * (q3 - q1)
print("Upper fence       :", round(upper_fence, 2))
print("Possible outliers :", cmax[cmax > upper_fence].tolist())
```

OUTPUT

```
Mean              : 16.76
Median            : 15.0
Skewness          : 1.05
Skewness (log)    : 0.0
Geometric mean    : 14.99
```

```
Upper fence       : 34.42
Possible outliers : [35.7]
```

Output	Interpretation
Mean 16.76 > Median 15.0	Mean is pulled up by a long right tail: **positive skew**.
Skewness 1.05	Above +1: **highly right-skewed**; the arithmetic mean alone is misleading.
Skewness (log) 0.00	After taking the natural log the data are symmetrical: the data are **log-normal**.
Geometric mean 14.99	Close to the median; the best single value for a 'typical' Cmax.
Upper fence 34.42; outlier [35.7]	One high value exceeds the fence: check the record; it may be a genuine high absorber.

Drawing graphs

PYTHON CODE

```
import matplotlib.pyplot as plt

plt.hist(weights, bins=7, color="teal", edgecolor="white")
plt.axvline(weights.mean(), color="orange", label="Mean")
plt.xlabel("Tablet weight (mg)")
plt.ylabel("Number of tablets")
plt.legend()
plt.show()

df.boxplot(column="sbp_fall", by="group")   # compare Drug and Placebo
plt.ylabel("Fall in systolic BP (mmHg)")
plt.show()
```

These commands produce the histogram of tablet weights (Figure 4) and the box plot comparing Drug and Placebo (Figure 10). Always look at the graph together with the numbers: a histogram can reveal bimodality or outliers that a mean and SD hide.

16. Case Studies: Python Analysis of Pharmaceutical Data

Case study 1: batch-to-batch uniformity (quality control)

Six tablets from each of three batches (A, B, C) were assayed for drug content as % of label claim. The task is to compare the batches using the mean, SD and CV.

PYTHON CODE

```
import pandas as pd

batches = pd.DataFrame({
    "batch": ["A"] * 6 + ["B"] * 6 + ["C"] * 6,
    "assay": [99.2, 100.1, 99.8, 100.4, 99.5, 100.0,
              98.5, 101.8, 97.9, 102.3, 99.0, 100.7,
              96.8, 97.5, 97.1, 96.9, 97.4, 97.3],
})

summary = batches.groupby("batch")["assay"].agg(["count", "mean", "median", "std", "min",
"max"])
summary["CV_%"] = summary["std"] / summary["mean"] * 100
print(summary.round(2))
```

OUTPUT

```
       count    mean  median   std   min    max  CV_%
batch
A          6   99.83   99.90  0.43  99.2  100.4  0.43
B          6  100.03   99.85  1.83  97.9  102.3  1.83
```

```
C          6   97.17   97.20  0.28  96.8   97.5  0.29
```

Observation	Interpretation
Mean: A 99.83, B 100.03, C 97.17	A and B are on target (about 100%). **Batch C has a lower mean (97.2%)**: it is consistently under-potent, a **systematic (accuracy) problem**.
SD / CV: A 0.43 (0.43%), B 1.83 (1.83%), C 0.28 (0.29%)	Batch C is the **most uniform** (smallest SD) but off-target; Batch B is on target on average but the **least uniform** (individual tablets from 97.9 to 102.3%), a **precision (mixing or blending) problem**.
Median close to mean	No marked skew in any batch; mean ± SD is a suitable summary.
Overall	**Only Batch A is both accurate and precise.** This shows why the mean and the SD (or CV) must always be read together.

ACCURACY VERSUS PRECISION

Accuracy = closeness of the mean to the true or target value. **Precision** = closeness of repeated values to each other (small SD or CV). A process can be precise but inaccurate (Batch C) or accurate on average but imprecise (Batch B).

PART F: SOLVED NUMERICAL PROBLEMS AND FORMULAE

17. Solved Numerical Problems

PROBLEM 1: TABLET HARDNESS (UNGROUPED DATA)

Hardness (kg/cm²) of 10 tablets: 6.2, 5.8, 6.5, 6.0, 5.9, 6.3, 6.1, 6.0, 6.4, 5.8. Find mean, median, mode, range, SD and CV.

Arranged: 5.8, 5.8, 5.9, 6.0, 6.0, 6.1, 6.2, 6.3, 6.4, 6.5. $\Sigma x = 61.0$, $n = 10$.

Mean = 61.0 / 10 = **6.10**. **Median** = (6.0 + 6.1)/2 = **6.05**. **Mode** = **5.8 and 6.0** (each occurs twice; bimodal). **Range** = 6.5 − 5.8 = **0.7**.

Deviations: −0.3, −0.3, −0.2, −0.1, −0.1, 0, 0.1, 0.2, 0.3, 0.4. Squares add up to **0.54**.

s^2 = 0.54 / 9 = 0.06; s = √0.06 = 0.245; CV = 0.245 / 6.10 × 100 = 4.0%.

Interpretation: mean and median are close, so the data are fairly symmetrical; CV 4% shows consistent compression.

PROBLEM 2: GROUPED DATA (PATIENT AGES)

Ages of 60 patients: **Class** 20–30, 30–40, 40–50, 50–60, 60–70; **f** = 6, 14, 20, 12, 8. Find mean, median, mode and SD.

Mid-points 25, 35, 45, 55, 65. $\Sigma fx = 150 + 490 + 900 + 660 + 520 = 2720$, so **mean = 2720 / 60 = 45.33 years**.

Cumulative frequency: 6, 20, 40, 52, 60. n/2 = 30 falls in class 40–50: **median = 40 + (30 − 20)/20 × 10 = 45 years**.

Mode: $f_1 = 20$, $f_0 = 14$, $f_2 = 12$. Mode = 40 + (20 − 14)/(40 − 14 − 12) × 10 = 40 + 60/14 = **44.3 years**.

SD: $\Sigma f(x - \bar{x})^2 = 8193.3 \rightarrow s^2 = 8193.3 / 59 = 138.9 \rightarrow$ **s = 11.8 years**.

Interpretation: mean ≈ median ≈ mode (45, 45, 44): symmetrical age distribution centred at about 45 years.

PROBLEM 3: WHICH METHOD IS MORE PRECISE?

Method A: mean = 250 µg/mL, SD = 5.0. Method B: mean = 98.0 µg/mL, SD = 2.4.

CV(A) = 5.0/250 × 100 = 2.0%. CV(B) = 2.4/98.0 × 100 = 2.45%.

Method B has the smaller SD, but Method **A is relatively more precise** (lower CV), because the SDs are measured on different scales.

PROBLEM 4: SKEWNESS FROM SUMMARY VALUES

A hospital-stay data set has mean = 62 hours, mode = 55 hours and SD = 14 hours. Find the Pearson coefficient of skewness and the approximate median.

Sk = (Mean − Mode) / SD = (62 − 55) / 14 = +0.5 → positive (right) skew, moderate.

Median ≈ (Mode + 2 × Mean) / 3 = (55 + 124) / 3 = 59.7 hours. Mode < Median < Mean, as expected for positive skew.

PROBLEM 5: USING THE EMPIRICAL RULE

Serum cholesterol in healthy adults is approximately normal with mean 190 mg/dL and SD 20 mg/dL.

68% of people lie between 190 ± 20 = **170 to 210**; **95%** between 190 ± 40 = **150 to 230**; **99.7%** between 190 ± 60 = **130 to 250 mg/dL**.

A person with 250 mg/dL has z = (250 − 190)/20 = **+3**, which is very unusual (only about 0.15% of healthy adults are higher).

PROBLEM 6: EFFECT OF CHANGING UNITS

Tablet weights have mean 500.8 mg and SD 3.254 mg. Convert to grams (divide by 1000).

New mean = 0.5008 g; new SD = 0.003254 g. CV = 0.003254 / 0.5008 × 100 = 0.65%, unchanged.

Mean and SD change with the unit but **CV does not**; this is why CV is used to compare variability across different units.

18. Formulae Used in This Unit

Measure	Formula (sample)
Arithmetic mean	$\bar{x} = \Sigma x / n$; grouped: $\bar{x} = \Sigma fx / n$; short-cut: $\bar{x} = A + h \cdot \Sigma fd / n$
Weighted mean	$\bar{x}w = \Sigma wx / \Sigma w$
Geometric mean / harmonic mean	$GM = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{(1/n)} = \exp(\Sigma \ln x / n)$; $HM = n / \Sigma(1/x)$
Median (grouped)	$L + [(n/2 - cf) / f] \times h$
Mode (grouped)	$L + [(f_1 - f_0) / (2f_1 - f_0 - f_2)] \times h$; Mode ≈ 3 Median − 2 Mean
Range, IQR, quartile deviation	Max − Min; Q3 − Q1; (Q3 − Q1) / 2
Variance and SD	$s^2 = \Sigma(x - \bar{x})^2 / (n - 1)$; $s = \sqrt{s^2}$; grouped: $s^2 = \Sigma f(x - \bar{x})^2 / (n - 1)$
Standard error of the mean	$SEM = s / \sqrt{n}$
Coefficient of variation	$CV\% = (s / \bar{x}) \times 100$
z-score	$z = (x - \bar{x}) / s$
Pearson skewness	Sk = (Mean − Mode) / SD or 3 (Mean − Median) / SD
Bowley skewness	(Q3 + Q1 − 2 Median) / (Q3 − Q1)
Outlier fences (Tukey)	Q1 − 1.5 × IQR and Q3 + 1.5 × IQR

Noteskarts
Smart Notes, Bright Future

STUDY SMARTER WITH NOTESKARTS

Free and premium resources for D.Pharm and B.Pharm students

WHERE TO FIND US

Noteskarts Prime App – complete notes, test series, PYQs and video courses in one place.

Noteskarts.com – free study material, blog articles and exam updates.

YouTube – Noteskarts – lecture videos and revision sessions.

WhatsApp and Telegram communities – daily updates and doubt discussion.

NOTE FOR STUDENTS

These notes are prepared for study and revision. Pharmacopoeial limits and reference values vary between editions and laboratories; always follow your prescribed textbook and the current pharmacopoeia.

Noteskarts
Smart Notes, Bright Future