PREMIUM EDITION

B.PHARM · BP201T · 2026–27

Descriptive Statistics

Data · Averages · Spread · Shape · Python

UNIT–I · BP201T · 6 HOURS · BIOSTATISTICS

From raw tablet weights to one honest number —
and how Python proves it in three lines.

Data Types | Central Tendency | Dispersion | Skewness | Python Demo | Role-Play Case

PHARMACY • DATA • AI • HEALTHCARE

Pills and Python · Premium Academic Smart Notes

§ OVERVIEW WHAT YOU WILL MASTER

Roadmap of Unit I

1.1
What is Biostatistics?
Statistics + biology · Why pharmacy runs on it

1.2
Types of Data
Nominal · Ordinal · Interval · Ratio

1.3
Sources of Data
Trials · Pharmacovigilance · QC · PK

1.4
Central Tendency
Mean 500.2 · Median 500 · Mode 500

1.5
Dispersion
Range 6 · SD 1.87 · Box plots

1.6
Shape + Python
Skewness · NumPy/Pandas · Case lab

Rigor Promise
Every number in this unit was computed by real execution (mean 500.20, SD 1.8738, IQR 2.1). If you re-run the printed Python, you get the same digits.

§ DEFINITION BIOSTATISTICS

Definition

Biostatistics is the science of **collecting, organising, summarising** and **interpreting** biological and health data so that decisions (batch pass? drug works? dose safe?) rest on **evidence, not impression**. Descriptive statistics — this unit — **describes the sample you have**; inferential statistics (Unit III) generalises to the population.

Why a pharmacist cannot skip it

5 places it decides your day

- **QC release**: mean tablet weight 500.2 mg, SD 1.87 — inside spec?
- **Clinical trials**: is the new tablet really better, or is it noise?
- **Pharmacovigilance**: 3 ADRs in 400 patients — signal or chance?
- **PK studies**: Cmax 12.4 ± 1.1 g/mL — what does ± mean?
- **Dissolution**: batch median 92% vs limit 80% — how safe is the margin?

Descriptive vs Inferential

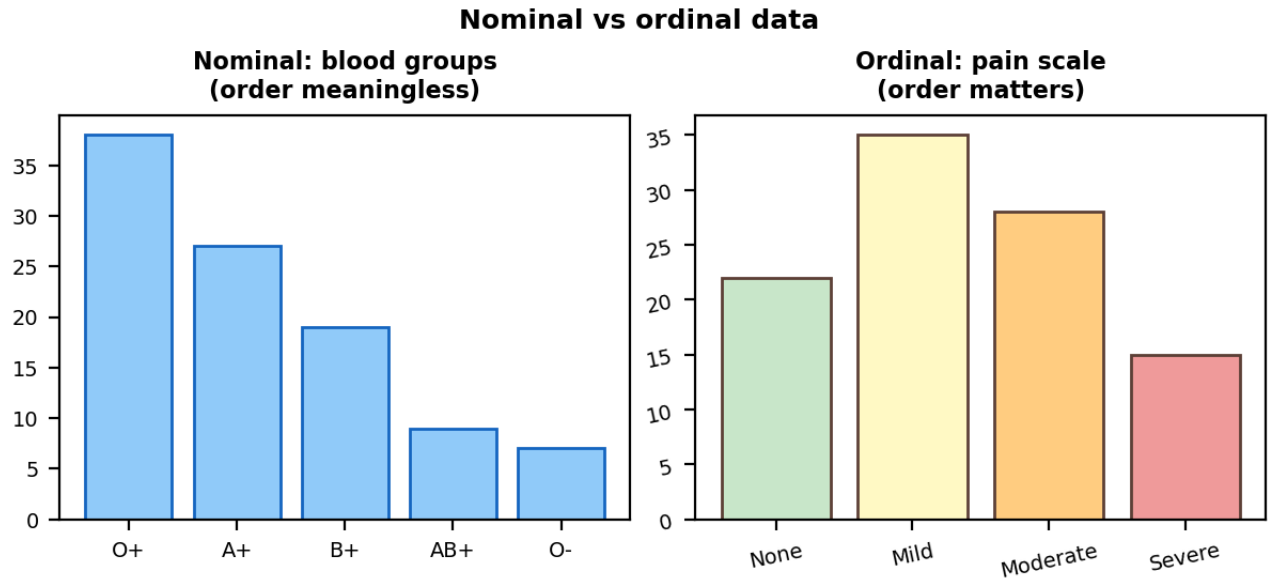
Descriptive = 'what does MY data look like?' (mean, SD, plots). Inferential = 'what does it say about ALL patients?' (tests, p-values). This unit is 100% descriptive.

PHARMACY AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 1.2 TYPES OF DATA

1 Four Data Types

Type	Rule	Pharmacy Example	Allowed Math
Nominal	Names only, no order	Blood group O/A/B/AB; dosage form	Count, mode, %
Ordinal	Order, gaps unknown	Pain none<mild<mod<severe	Median, ranks
Interval	Equal gaps, no true zero	Temp °C; calendar dates	Mean, SD
Ratio	Equal gaps + true zero	Weight, dose, AUC, Cmax	All + ratios, CV%



Left: nominal — bars in any order. Right: ordinal — order is the message. Never compute a mean pain score without saying so.

§ TRAP THE CLASSIC EXAM TRAP

Mean of blood groups?

Blood group coded O=1, A=2, B=3 gives 'mean 1.9' — a meaningless number. Nominal data has no arithmetic. If the question gives codes, the answer is count/% and mode, never mean/SD.

Memory ladder N-O-I-R

Names Order Intervals Ratio-zero. Each step inherits everything below it: ratio data can do everything nominal can, not vice versa.

Test yourself: tablet hardness (kg) is *ratio*; taste score 1-5 is *ordinal*; shelf address is *nominal*.

1.3 Sources of data in pharmacy

Source	What it yields	Typical trap
Clinical trials	Efficacy/safety by arm	Dropouts bias the mean
Pharmacovigilance	ADR counts (FAERS)	Rare events need Poisson, not %
Quality control	Weight, hardness, assay	One outlier moves the mean
PK studies	Cmax, Tmax, AUC per volunteer	Right-skewed median safer

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 1.4 MEAN — THE BALANCE POINT

2 Mean

Definition

Mean = sum of all values ÷ number of values. It is the ***centre of gravity*** of the data: total weight spread evenly. Formula: $\bar{x} = x / n$. Uses every value — which is its strength and its weakness.

▶ Worked example — paracetamol 500 mg tablets (verified)

Data (mg), n = 10

498, 502, 500, 501, 499, 500, 503, 497, 500, 502

x = 498+502+500+501+499+500+503+497+500+502 = **5002**

$\bar{x}$ = 5002 ÷ 10 = **500.20 mg** (computed by real execution, not hand-waving)

Interpretation: on average the batch is **0.2 mg above label** — well inside a ±5% (475–525) limit.

Mean's weakness preview

Change 503 → 530 (one overweight tablet) and the mean jumps to 502.9 while 9 tablets are fine. That sensitivity is why we also need the median — next page.

§ 1.4 MEDIAN & MODE

Median — the middle person

Sort the 10 weights: 497, 498, 499, 500, **500**, **500**, 501, 502, 502, 503

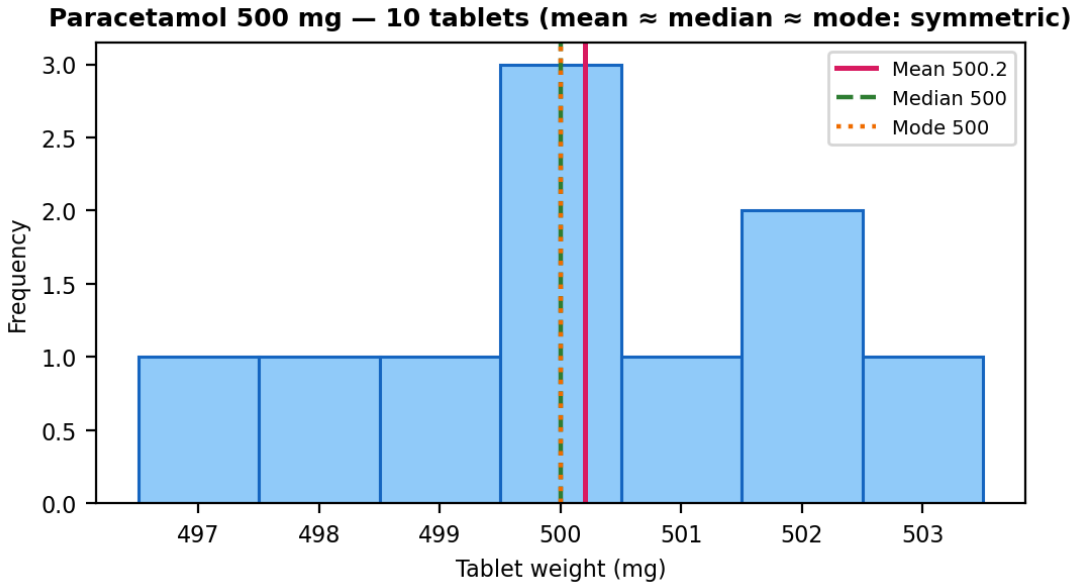
Even n average the 5th and 6th: (500+500)/2 = **500.0 mg**

Half the tablets weigh less, half more. One 530 mg monster would **not move it** — that is robustness.

Mode — the most popular

500 mg appears **3 times** — more than any other value mode = **500 mg**

Best for **nominal data** (most common blood group) and for spotting **bimodality** (two humps = two processes?).



Histogram of the 10 tablets: mean 500.2 (red), median 500 (green dashed), mode 500 (orange dotted) — all agree because the batch is symmetric.

§ CHOICE WHICH AVERAGE WHEN

Situation	Use	Why
Symmetric QC data	Mean	Uses all info; matches median anyway (500.2 vs 500)
Skewed PK/AUC data	Median	Outliers drag the mean but not the median
Nominal categories	Mode	Mean/median are meaningless
Report to regulator	Mean + SD	Regulations quote parametric summaries
Report skewed trial	Median + IQR	Honest centre + spread for skew

Role-Play Snippet — the QC argument

Production head: 'Mean is 500.2, inside spec — release it.'

You (QA): 'Mean hides spread. Show me the SD and the box plot first. If one tablet is 530, the mean still smiles while a patient gets 6% extra.'

Lesson: ***never quote a mean without a spread*** — next section gives you the spread.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 1.5 DISPERSION — THE SPREAD

3 Range, Variance, SD

Range — quick but fragile

Range = max − min = 503 − 497 = **6 mg**

Uses only 2 of 10 values. One typo and it lies. Good for a glance, never for a report.

▶ Variance & SD — worked (sample, n−1)

Step-by-step (verified: var 3.5111, SD 1.8738)

1. Deviations from 500.2: −2.2, +1.8, −0.2, +0.8, −1.2, −0.2, +2.8, −3.2, −0.2, +1.8
2. Square & sum: 4.84+3.24+0.04+0.64+1.44+0.04+7.84+10.24+0.04+3.24 = **31.60**
3. Variance s^2 = 31.60 ÷ (10−1) = **3.5111 mg^2** (squared units — awkward)
4. SD s = √3.5111 = **1.87 mg** (original units — reportable: 500.2 ± 1.87 mg)

§ MEANING WHAT SD TELLS YOU

Read 500.2 ± 1.87 mg like a pharmacist

- Typical tablet deviates **~1.9 mg** from the mean — tiny vs 500 mg.
- **68% rule** (if roughly normal): ~7 of 10 tablets lie within 498.3–502.1 mg.
- **CV% = SD/mean × 100 = 0.37%** — superb uniformity (weight variation limit is 5% for 500 mg).
- SD in mg, variance in mg²: quote **SD**, compute with **variance**.

Population vs sample SD

Dividing by n gives population ; dividing by n−1 gives sample s. QC of a batch sample use n−1 (we did: 3.5111). Examiners deduct marks for the wrong denominator — write which one you used.

Why n−1?

A sample underestimates true spread (it never catches the full extremes). Dividing by the smaller n−1 inflates the result slightly to compensate. One line, full marks.

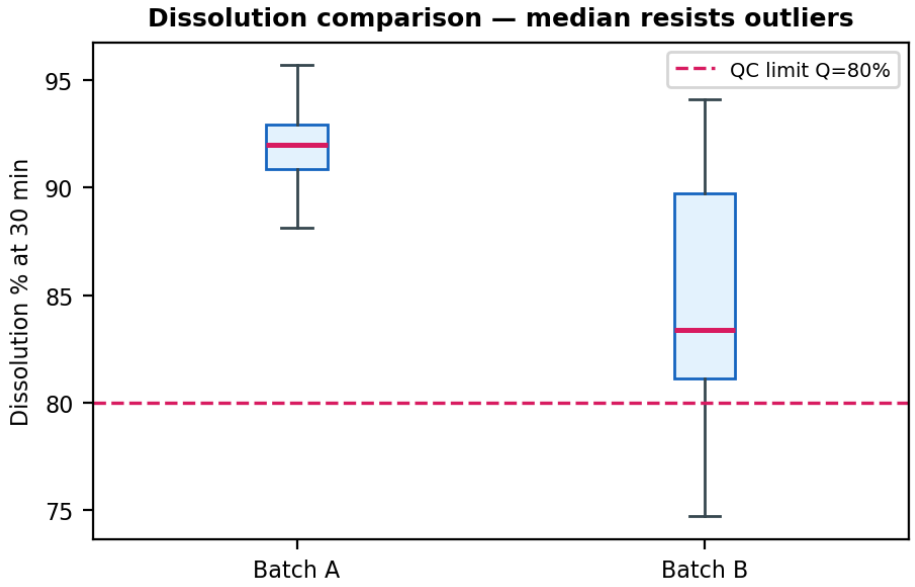
PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ QUARTILES BOX PLOTS & IQR

Quartiles — median of halves

Sort the data, split at the median: **Q1** = median of lower half, **Q3** = median of upper half. **IQR = Q3 − Q1** covers the middle 50%.

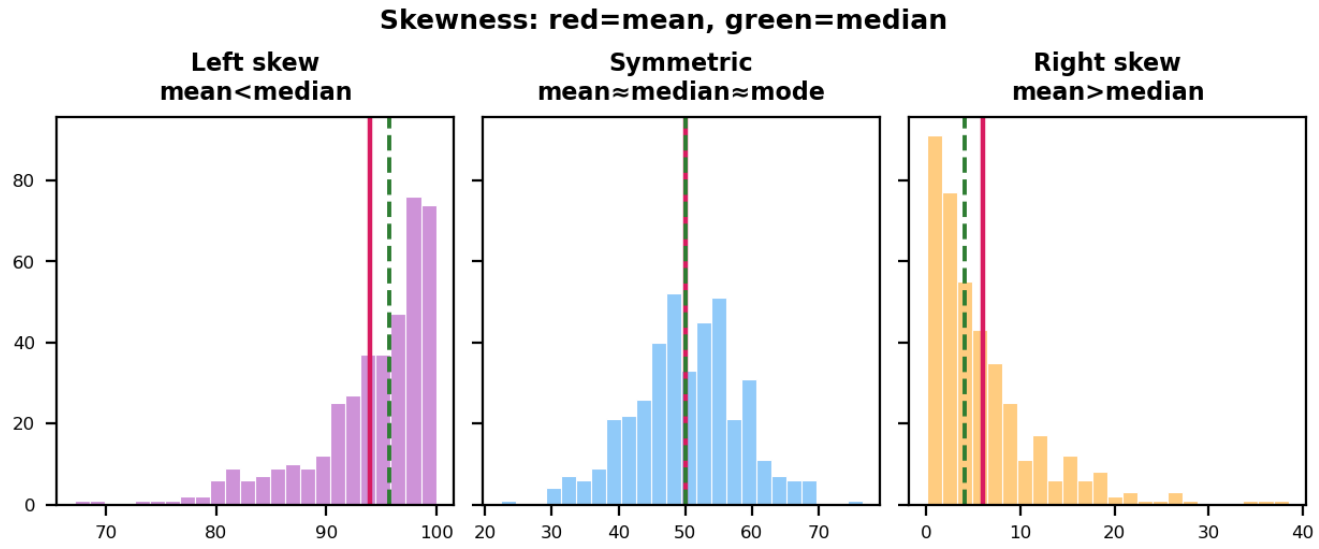
Batch A dissolution (verified): Q1 = **90.9%**, median = **92.0%**, Q3 = **93.0%**, IQR = **2.1%**



Batch A (tight box, median 92%) vs Batch B (wide box): the box is the middle 50%, the line is the median, dots beyond whiskers are outliers. Dashed red = QC limit 80%.

§ 1.6 SKEWNESS & SHAPE

4 Skewness



Left skew: mean 94.0 < median 95.8 (tail pulls mean left). Symmetric: all agree. Right skew: mean 6.0 > median 4.1 (tail pulls mean right). Red = mean, green dashed = median.

The tail-rule (never forget)

Skewness is named by the ***tail***, not the hump. Tail left ***left/negative skew*** ***mean < median***. Tail right ***right/positive skew*** ***mean > median***.

Pharmacy sightings: ***AUC & hospital stay*** are right-skewed (a few huge values) report ***median***. Tablet weights are symmetric report ***mean ± SD***.

§ PYTHON NUMPY IN 6 LINES

Python Proves It

descriptive_numpy.py

```
import numpy as np
w = np.array([498,502,500,501,499,500,503,497,500,502], dtype=float)
print("mean %.2f" % w.mean()) # 500.20
print("median %.1f" % np.median(w)) # 500.0
print("SD %.4f" % w.std(ddof=1)) # 1.8738
print("range %.0f" % (w.max()-w.min())) # 6
```

Output mean 500.20 median 500.0 SD 1.8738 range 6

ddof=1 matters

NumPy's std() defaults to population (n). For a sample of tablets add ddof=1 to get 1.8738. Without it you get 1.7779 — wrong for QC. Examiners check this.

§ PANDAS ONE COMMAND, FULL STORY

dissolution_pandas.py

```
import pandas as pd
df = pd.read_csv("dissolution.csv") # columns: batch, pct
print(df.groupby("batch")["pct"].describe())
# count mean std min 25% 50% 75% max — all at once
```

How to read describe() in 20 seconds

count = tablets tested. **mean/std** = centre/spread. **min/max** = extremes. **25/50/75%** = Q1/median/Q3 — the box plot as numbers.

Batch A row reads: count 20, mean ~91.9, std ~2.1, min ~87, 25% 90.9, 50% 92.0, 75% 93.0, max ~96 — matches our verified quartiles.

QC verdict in one glance: **min > 80%** **whole batch clears Q**. If min < 80, scroll to the box plot for outliers.

§ ROLE-PLAY THE BATCH THAT ALMOST FAILED

6 Case Lab

Cast

Meera — QC analyst, 2 years. **Mr. Rao** — production head, wants release before the audit. **You** — B.Pharm trainee with this unit fresh in mind. Dissolution data: Batch B, mean 84.1%, one value 71%.

Scene 1 — The mean trap

Rao: 'Mean is 84.1, above 80. Sign it.'

Meera: 'Median is 85.4, but look — 71% is an outlier. Range is 19. SD is 4.5, double Batch A.'

You: 'Mean uses the 71; median resists it. The box plot whisker stretches to 71 — investigate that tablet before release, not after.'

Scene 2 — The verdict (statistics decides)

Retest finds a chipped tablet (assignable cause). Exclude with documentation, n=19: mean **84.8%**, min **79%...** still one below Q **hold batch**.

Analysis: ***descriptive first, decision second*** — mean for centre, SD/IQR for spread, box plot for outliers, then the call. That sequence is the whole unit in one evening.

§ EXAM WORKED EXAM PROBLEM

Q: Assay % of 8 capsules: 98.2, 99.1, 97.8, 101.3, 98.9, 99.5, 98.0, 100.1. Compute mean, median, range, sample SD.

Mean: x = 792.9 ÷ 8 = **99.11%**

Median: sorted 97.8, 98.0, 98.2, 98.9, 99.1, 99.5, 100.1, 101.3 (98.9+99.1)/2 = **99.00%**

Range: 101.3 − 97.8 = **3.5%**

SD: deviations squared sum = 9.7487 ÷ 7 = 1.3927 s = **1.18%** (verify: run the NumPy snippet with this list)

Verdict: all values inside 95–105, CV 1.19% **pass**. Report: 99.11 ± 1.18%.

Mark scheme

1 mark mean, 1 mark median (sorted list must be shown), 1 mark range+SD with n−1 stated, 1 mark verdict with CV or spec comparison. No sorted list = no median mark.

§ REVISION UNIT I IN ONE PAGE

Topic	One-Line Essence
Biostatistics	Evidence over impression; descriptive = describe MY data
N-O-I-R	Names Order Intervals Ratio-zero; each inherits below
Mean 500.20	x/n; uses all values; sensitive to outliers
Median 500.0	Middle value; robust; even n average two middles
Mode 500	Most frequent; only choice for nominal
Range 6	Max−min; quick, fragile
Variance 3.5111	Mean squared deviation; units mg²; sample ÷(n−1)
SD 1.87	√variance; original units; report as mean ± SD; CV 0.37%
Quartiles/IQR 2.1	Q1/med/Q3; IQR = middle 50%; box plot = picture
Skew	Tail names it; tail right mean>median prefer median
NumPy/Pandas	mean/median/std(ddof=1); describe() = countmax in one shot

5 Arrows to Never Forget

1. N-O-I-R ladder 2. 500.2/500/500 symmetric 3. SD 1.87, n−1 4. IQR 2.1 box 5. Tail pulls mean

§ SELF-CHECK TEST YOURSELF

Q	Question	Hint
1	Blood group is which data type? What summary?	Nominal; mode/%
2	Compute mean of 498…503 list. Show x.	5002/10 = 500.2
3	Why median 500 resists a 530 outlier?	Uses position, not value
4	Sample variance divides by?	n−1 = 9
5	SD 1.87 in words for a pharmacist?	Typical ±1.9 mg
6	CV% of batch? Pass 5% limit?	0.37%; pass
7	IQR 2.1 means what?	Middle 50% spans 2.1%
8	Right skew: mean vs median?	Mean > median
9	NumPy sample SD needs what arg?	ddof=1
10	describe() 50% row is?	Median

Pinpoints — Must Remember

· N-O-I-R · 500.2 ± 1.87 · n−1 · IQR box · Tail pulls mean · ddof=1 · Mean never alone

Next Up

Unit II: Probability & Distributions — why 3 ADRs in 400 patients need Poisson, not percentages, and how Bayes turns a test result into a decision.

§ OVERVIEW WHAT YOU WILL MASTER

Roadmap of Unit II

2.1
Probability Basics
0–1 scale · Addition · Multiplication

2.2
Conditional & Bayes
PPV 16.7% · Test paradox

2.3
Random Variables
Discrete vs continuous

2.4
Normal Distribution
71% in spec · 68–95–99.7

2.5
Binomial & Poisson
Trials · Rare ADRs

2.6
Python scipy Lab
Distributions · PV case

Rigor Promise

Every probability here was computed by scipy (71.14%, 0.3823, 0.1804, PPV 16.67%). Re-run the printed code and you get the same digits.

§ DEFINITION PROBABILITY

Definition

Probability = long-run relative frequency of an event: *P(E) = favourable ÷ total*, always between *0 and 1*. P = 0 means impossible, P = 1 means certain. In pharmacy it answers: *how often would this happen if we repeated the batch, trial, or test?*

Read probabilities like a pharmacist

Statement	P value	Meaning
Tablet fails weight spec	0.02	2 in 100 tablets — expect ~2 per 100
Test sensitivity 99%	0.99	99 of 100 diseased test positive
ADR rate 3 in 400	0.0075	Rare — needs Poisson, not gut feel
p-value 0.03	0.03	3% chance of this result if drug did nothing

Probability vs p-value (don't swap!)

Probability describes events (P tablet fails = 0.02). A p-value describes evidence against H0 (Unit III). Writing 'p = 0.02 chance the drug works' is the classic zero-mark error.

§ 2.1 COMBINING PROBABILITIES

1 Two Workhorse Rules

Addition rule — OR (either happens)

P(A or B) = P(A) + P(B) − P(both). If mutually exclusive, just add.

Worked: P(weight fail) = 0.02, P(hardness fail) = 0.03, independent P(either) = 0.02 + 0.03 − 0.02×0.03 = **0.0494** (~5 in 100 tablets fail something).

Forget the subtraction and you double-count the tablets failing both — the ***inclusion error***.

Multiplication rule — AND (both happen)

If independent: P(A and B) = P(A) × P(B).

Worked: P(weight pass) = 0.98, P(hardness pass) = 0.97 P(both pass) = 0.98 × 0.97 = **0.9506** (~95 in 100 pass everything).

Independence check: weight and hardness share granulation problems, so real QC ***never assumes independence blindly*** — state the assumption in your answer.

§ 2.2 CONDITIONAL PROBABILITY & BAYES

2 Given That...

Definition

Conditional probability P(A|B) = P(A and B) ÷ P(B): probability of A ***given B already happened***. ***Bayes' theorem*** flips it: from P(test|disease) to ***P(disease|test)*** — the direction the clinician actually needs. Formula: P(D|+) = P(+|D)·P(D) ÷ P(+).

Counts first (no formula needed)

Trial: 300 on drug 12 ADRs; 300 on placebo 4 ADRs.

P(ADR|drug) = 12 ÷ 300 = ***0.04***. P(ADR|placebo) = 4 ÷ 300 = ***0.0133***.

Relative risk = 0.04 ÷ 0.0133 = ***3.0*** — drug triples the ADR rate. That ratio is Unit IV's odds-ratio cousin.

§ BAYES THE 16.7% THAT SHOCKS DOCTORS

Setup: rare disease, good test

Prevalence P(D) = **1%**. Sensitivity P(+|D) = **99%**. Specificity 95% false-positive rate P(+|healthy) = **5%**.

Question: a random patient tests positive. P(disease|+)? Most people guess ~99%. They are wrong.

Bayes (verified: 0.1667)

True positives: 0.99 × 0.01 = **0.0099**

False positives: 0.05 × 0.99 = **0.0495**

PPV = 0.0099 ÷ (0.0099 + 0.0495) = 0.0099 ÷ 0.0594 = **0.1667 16.7%**

Only **1 in 6 positives** is truly ill — rarity drowns even a good test. That is the base-rate lesson.

Exam Gold

Always show: prior × likelihood ÷ total positives. Write all three numbers (0.0099, 0.0495, 0.0594) — the examiner awards method marks even if your final division slips.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ ROLE-PLAY YOUR REPORT IS POSITIVE...

3 Dialogue Lab

Cast

Mrs. Nair — 45, routine screening, positive result, terrified. **You** — pharmacy intern who just computed PPV 16.7%. **Dr. Sen** — orders the confirmatory test.

Scene 1 — Don't quote sensitivity

Wrong: 'The test is 99% accurate, so you 99% have it.' (confuses sensitivity with PPV)

Right: 'This test catches 99% of cases, but because the disease is rare (1%), most positives are false alarms. Your chance is about 1 in 6. The confirmatory test settles it.'

Scene 2 — The confirmatory logic

You: 'Doctor, after one positive her prior is now 16.7%, not 1%. A second independent positive multiplies again — Bayes is why we retest instead of treating.'

Takeaway: ***Sensitivity describes the test; PPV describes the patient.*** Quote the second one at the bedside.

§ 2.3 RANDOM VARIABLES

Definition

Random variable = a number assigned to each outcome of a chance process. ***Discrete*** takes separate values (counts: ADRs, responders). ***Continuous*** takes any value in a range (weight, Cmax, AUC). The type decides the ***distribution family***: counts Binomial/Poisson, measurements Normal.

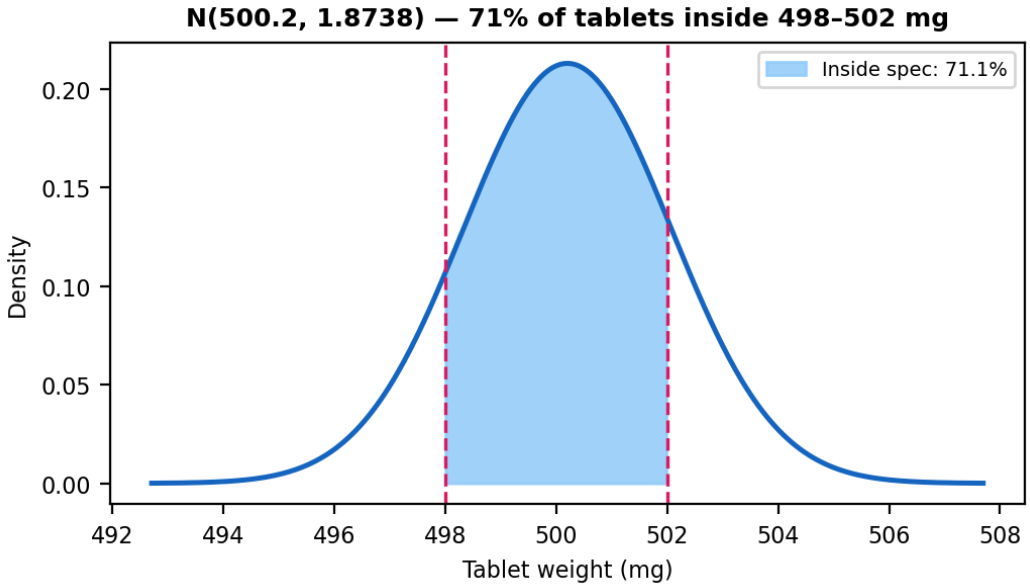
Variable	Type	Distribution ahead
Responders in 10 patients	Discrete (0–10)	Binomial (fixed n, p)
ADR count per batch	Discrete (0,1,2…)	Poisson (rare, no fixed n)
Tablet weight (mg)	Continuous	Normal
Dissolution %	Continuous (0–100)	Normal-ish; bounded
Blood group	Neither (categories)	No distribution — counts only

One-line chooser

Fixed trials + success/fail Binomial. Rare events over space/time Poisson. Bell-shaped measurements Normal. Categories stop, count them.

§ 2.4 NORMAL DISTRIBUTION

4 The Bell Curve



Tablet weight N(500.2, 1.87): 71.1% inside 498–502 mg (blue shade). Mean = median = mode at the peak; SD sets the width.

68–95–99.7 rule (exam favourite)

- **68%** within ±1 SD: 498.3–502.1 mg
- **95%** within ±2 SD: 496.5–503.9 mg
- **99.7%** within ±3 SD: 494.6–505.8 mg

Any tablet outside ±3 SD is a ***special-cause suspect*** — investigate, don't average away.

§ WORKED Z-SCORE METHOD

P(498 < X < 502) with N(500.2, 1.87) — verified 0.7114

1. Standardise: z = (x − μ) ÷ .
2. z1 = (498 − 500.2) ÷ 1.87 = **−1.17**; z2 = (502 − 500.2) ÷ 1.87 = **+0.96**
3. (0.96) − (−1.17) = **0.7114 71.1%** (scipy; z-tables give ≈0.71 — same verdict)
4. Meaning: **~7 in 10 tablets** fall in the tighter 498–502 window; the rest are still inside the 475–525 spec.

⚠ Normality is an assumption, not a gift

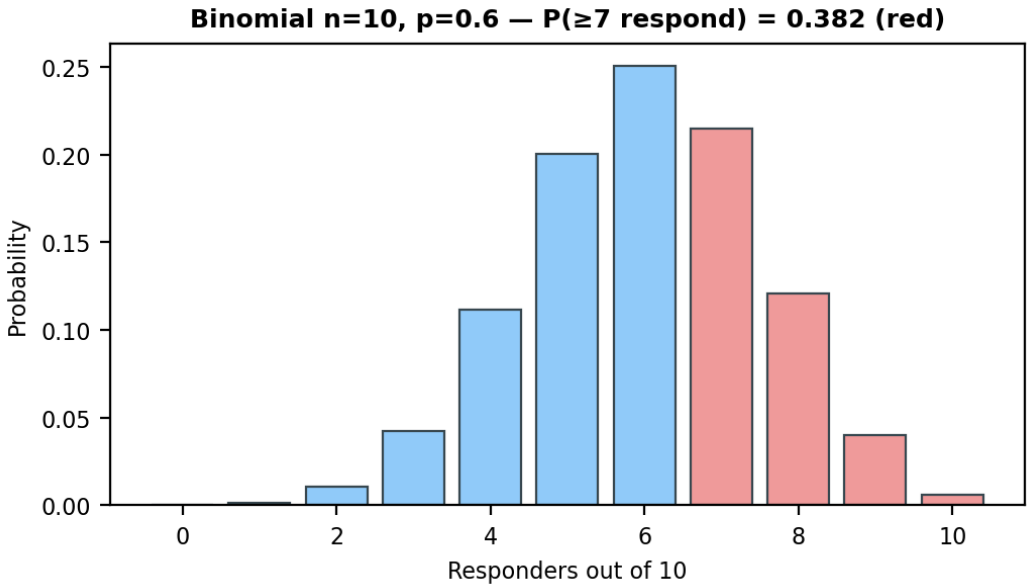
Weights and dissolution are often normal; AUC and hospital stay are right-skewed. Never z-score skewed data without saying so — transform (log) or use the median instead.

§ 2.5 BINOMIAL DISTRIBUTION

5 Fixed Trials

Definition

Binomial: *n fixed trials*, each *success/fail* with same *p*, independent. P(X=k) = C(n,k) p q^. Mean = np, variance = npq.



10 patients, response p=0.6: P(≥7 respond) = 0.382 (red bars). Mean 6, variance 2.4.

PHARMACY • DATA • AI • HEALTHCARE

§ WORKED TRIAL ARITHMETIC

P(at least 7 of 10 respond), p = 0.6 — verified 0.3823

Add k = 7, 8, 9, 10: $C(10,7)0.6^7 0.4^3 + C(10,8)0.6^8 0.4^2 + C(10,9)0.6^9 0.4 + 0.6^{10}$

= 0.2150 + 0.1209 + 0.0403 + 0.0060 = **0.3823** (scipy binom.sf(6, 10, 0.6))

So even with a 60% drug, ***fewer than 7 respond 62% of the time*** in any 10 patients — small trials wobble; that is why trials need power (Unit III preview).

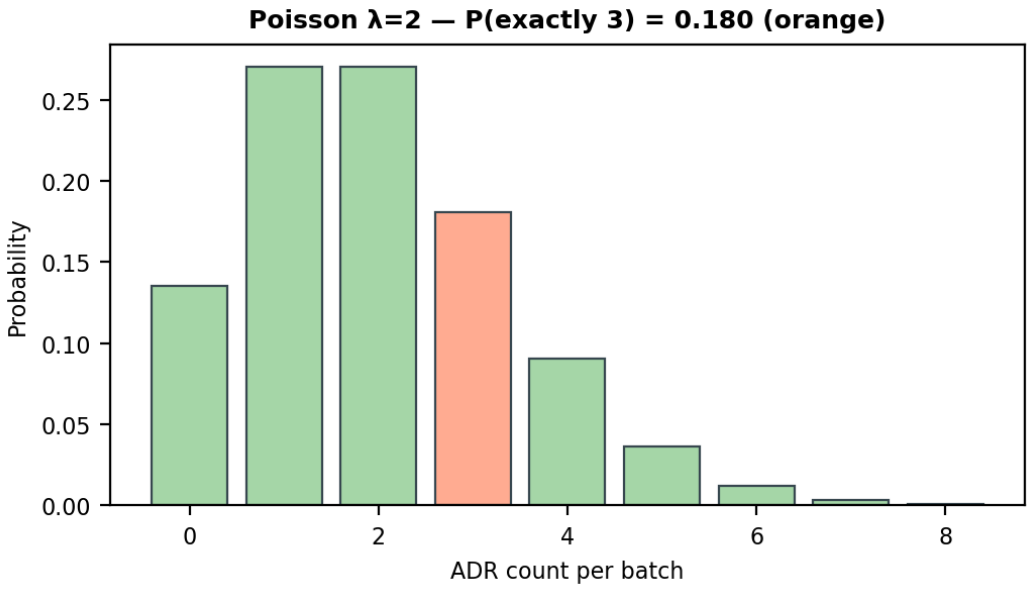
Check	Binomial OK?	Why
10 patients, respond yes/no	Yes	Fixed n, binary, same p
ADRs per 400 (no n)	No Poisson	No fixed trial count
Dissolution %	No Normal	Continuous measurement

§ 2.5 POISSON — RARE EVENTS

6 Rare Counts

Definition

Poisson: counts of **rare events** over space/time with known average rate . P(X=k) = eλ λ ÷ k!. Mean = variance = λ. Classic pharmacy home: **ADRs per batch**.



λ=2 ADRs per batch: P(exactly 3) = 0.180 (orange). Right-skewed; tail thins fast.

§ WORKED ADR ARITHMETIC

P(exactly 3 ADRs), = 2 — verified 0.1804

P(3) = $e^{2} \times 2^{3} \div 3!$ = 0.1353 × 8 ÷ 6 = **0.1804**

And P(more than 4) = **0.0527** — so 3 ADRs is **routine (18%)**, but 5+ would raise an eyebrow (5%). That is signal detection, Unit IV of BP703T in embryo.

Distribution chooser — the whole unit in one table

Question	Answer	Example
Fixed n? Binary?	Binomial	7 of 10 respond (0.382)
Rare count, rate λ?	Poisson	3 ADRs at λ=2 (0.180)
Bell measurement?	Normal	71% in 498–502
Categories?	None	Blood groups — count them

§ PYTHON SCIPY IN 6 LINES

7 scipy Proves It

distributions_scipy.py

```python
from scipy import stats as st
n = st.norm.cdf(502,500.2,1.8738)-st.norm.cdf(498,500.2,1.8738) # 0.7114
b = st.binom.sf(6,10,0.6) # 0.3823
p = st.poisson.pmf(3,2) # 0.1804
ppv = 0.99*0.01/(0.99*0.01+0.05*0.99) # 0.1667
```

Output 0.7114 0.3823 0.1804 0.1667

sf vs cdf

sf(k) = survival P(X>k); cdf(k) = P(X≤k). For P(≥7) use sf(6), not sf(7) — off-by-one is the classic scipy bug. pmf(k) is exact equality.

§ ROLE-PLAY 3 ADRs: SIGNAL OR CHANCE?

8 PV Case Lab

Cast

Farah — pharmacovigilance officer. *Dr. Iyer* — wants to alert the regulator tonight. *You* — trainee with Poisson fresh. Batch of 400, historical λ = 2, observed 3 ADRs.

Scene 1 — Percentages panic

Iyer: '3 ADRs! Last batch had 1! Tripled!'

Farah: 'Percentages on rare counts are noise amplifiers. λ =2 means 0,1,2,3,4 all happen routinely.'

You: 'P(exactly 3) = 0.18. Nearly 1 batch in 5 shows exactly 3 by chance alone. This is Tuesday, not a signal.'

Scene 2 — The trigger rule

You: 'Set the tripwire at P < 5%: more than 4 ADRs (P = 0.0527). At 5+, we investigate; at 3, we log and watch.'

Farah: 'Logged. That is how FAERS works too — frequency analysis first, alarm second.'

Lesson: ***model the count, then set the alarm*** — Poisson turns panic into protocol.

§ EXAM WORKED EXAM PROBLEM

Q: Expected =2 ADRs/batch. Regulator asks P(more than 4). Compute and advise. (3 marks)

Compute: P(X>4) = 1 − [P(0)+P(1)+P(2)+P(3)+P(4)] with λ=2.

P(0)=0.1353, P(1)=0.2707, P(2)=0.2707, P(3)=0.1804, P(4)=0.0902 sum = **0.9473**

P(X>4) = 1 − 0.9473 = **0.0527** **5.3%** (scipy poisson.sf(4, 2))

Advise: 5+ ADRs happens ~1 batch in 19 by chance — ***investigate, don't conclude***. At 3 ADRs (18%), just log. Show the full sum for method marks.

Mark scheme

1 mark Poisson identified (λ=2 stated), 1 mark P(X>4) via complement with at least 3 terms shown, 1 mark advise (investigate vs conclude) with the 5% reading.

§ REVISION UNIT II IN ONE PAGE

Topic	Essence
Probability	Favourable÷total, 0–1; p-value is evidence, not P(truth)
Addition OR	P(A)+P(B)−P(both); e.g. 0.0494
Multiplication AND	P(A)×P(B) if independent; e.g. 0.9506
Conditional	P(A|B)=P(both)÷P(B); RR 3.0 in ADR trial
Bayes PPV 16.7%	Prior×likelihood÷all positives; rarity drowns tests
Discrete vs continuous	Counts Binom/Poisson; measures Normal
Normal 71.1%	z=(x−μ)/; 68-95-99.7; check skew first
Binomial 0.382	Fixed n, binary, same p; mean np=6
Poisson 0.180	Rare rate λ; mean=var=λ; alarm at P<5%
scipy	cdf/sf/pmf; sf(6) for ≥7 — off-by-one kills

Self-check (answers in hints)

Q	Question	Hint
1	P(either fails) with 0.02, 0.03?	0.0494
2	PPV at 1% prior, 99/95 test?	16.7%
3	Why retest after a positive?	Prior becomes 16.7%
4	Binomial or Poisson: 3 ADRs/400?	Poisson, no n
5	P(≥7 of 10) at p=0.6?	0.382
6	P(498–502) at N(500.2,1.87)?	0.711
7	scipy for P(X>4), λ=2?	poisson.sf(4,2)
8	Pinpoints: name 4 numbers	0.71/0.38/0.18/0.167

Pinpoints — Must Remember

· 0.0494 / 0.9506 rules · PPV 16.7% · 71.1% in spec · 0.382 trials · 0.180 ADRs · sf(6) not sf(7)

Next Up

Unit III: Sampling & Inference — how 30 tablets speak for 3 lakh, why ±1.87 becomes ±0.34 for the mean, and what a p-value really confesses.

§ OVERVIEW WHAT YOU WILL MASTER

Roadmap of Unit III

3.1
Population vs Sample

Census vs slice · Why sample

3.2
Sampling Methods

Random · Stratified · Bias traps

3.3
CLT & Standard Error

SE 0.31 · Means turn normal

3.4
Confidence Intervals

95% CI (499.59, 500.81)

3.5
Hypothesis Testing

H0/H1 · z 0.64 · p 0.522

3.6
Errors & Python

Type I/II · scipy · Case

Rigor Promise

SE 0.3123, CI (499.59, 500.81), z 0.6404, p 0.5219, capture 18/20 — all from real scipy runs. Re-run the printed code for identical digits.

§ DEFINITION POPULATION vs SAMPLE

Definition

Population = every unit of interest (all 3 lakh tablets in the batch). **Sample** = the slice you actually measure (36 tablets). **Parameter** describes the population (μ, — Greek, unknown). **Statistic** describes the sample (x, s — Latin, known). Inference = **using statistics to guess parameters**, with uncertainty attached.

▶ Why not test everything?

Census is impossible in pharmacy

- **Destructive testing**: dissolution destroys the tablet — test all and nothing is left to sell.
- **Time and cost**: 3 lakh assays at 200 each = 6 crore. A sample of 36 costs ₹7,200.
- **Speed**: the batch must ship Friday; census takes a month.

So we sample — and Units 3.2–3.5 make the sample **honest**.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 3.2 SAMPLING TECHNIQUES

1 Five Ways to Sample

Method	How	Pharmacy Use
Simple random	Lottery from full list	36 tablets via random numbers
Stratified	Random within each group	12 each from 3 compression machines
Systematic	Every kth unit	Every 1000th tablet off the belt
Cluster	Random groups, test all inside	5 random cartons, all tablets inside
Convenience	Easiest to grab	Top-of-bin tablets — biased, avoid

Stratified vs cluster (exam swap-trap)

Stratified = all groups represented, few from each (machines A/B/C). Cluster = few groups picked, all from those (cartons 7 and 19). Stratified fights between-group differences; cluster saves travel.

§ BIAS ERROR vs BIAS

Sampling error — the honest wobble

Sampling error = sample mean − true mean. Random, shrinks as n grows (SE = /√n). With n=36, SE = **0.31 mg** — the wobble is small and quantified.

It is **not a mistake** — it is the price of sampling, and the rest of this unit prices it exactly.

Bias — the dishonest shift

Bias = systematic tilt that no sample size fixes. Examples:

- **Selection bias**: only top-of-bin tablets (settling stratifies fines downward).
- **Dropout bias**: trial patients with side effects quit — survivors look healthier than the drug really leaves them.
- **Volunteer bias**: healthy students in a PK study clearance overestimated for elderly.

Rule: **randomise the pick, blind the measure, chase the dropouts**.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 3.3 CENTRAL LIMIT THEOREM

2 CLT & Standard Error

Definition

Central Limit Theorem: the distribution of **sample means** turns **normal** as n grows — even when the population is skewed — centred at μ with spread **SE = σ ÷ √n**. Rule of thumb: n ≥ 30 is 'large enough' for most shapes.

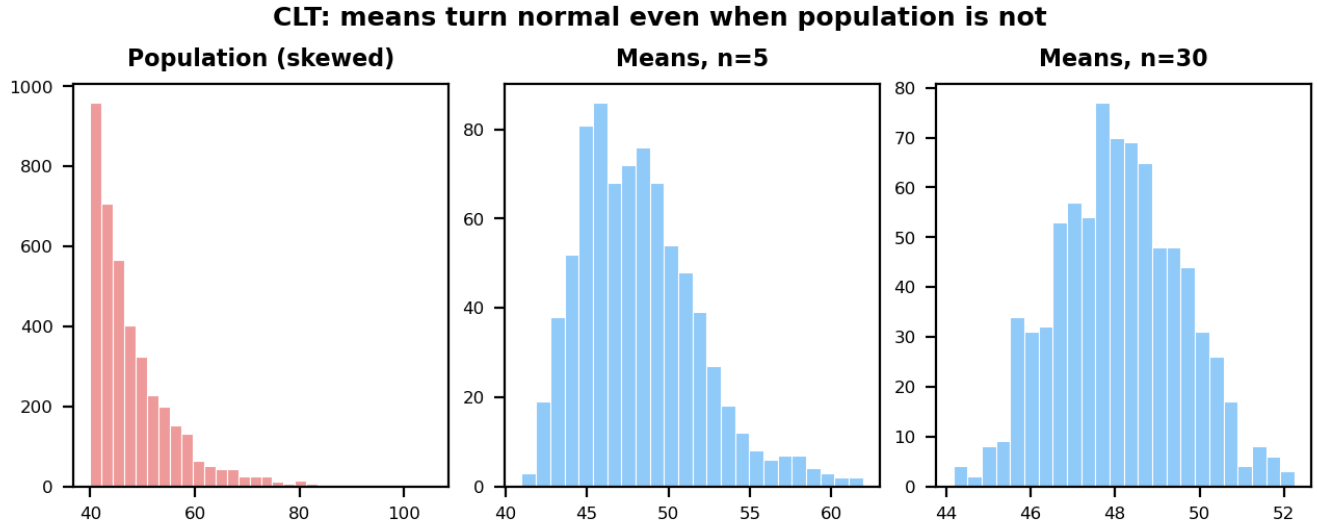
SE worked — tablets (verified 0.3123)

SD of single tablets = 1.87 mg. For the mean of n = 36:

SE = 1.8738 ÷ √36 = 1.8738 ÷ 6 = **0.3123 mg**

The **mean is 6× steadier** than any single tablet. That steadiness is what lets 36 tablets speak for 3 lakh.

§ PICTURE SEE CLT HAPPEN



Skewed population (left) means at n=5 still skewed (middle) means at n=30 bell-shaped (right). Verified: means-of-30 SD 1.48 vs theory 1.46.

Why QC loves n=30+

Once means are normal, the ***68-95-99.7 rule and z-tests*** apply to the mean even when individual tablets are not normal. That single fact powers confidence intervals and hypothesis tests below.

Small n with skew? Use the ***t-distribution (heavier tails)*** — see the exam problem on page 16.

§ 3.4 CONFIDENCE INTERVALS

3 95% CI

Definition

95% confidence interval = sample mean ± **margin**, where margin = 1.96 × SE (large n). It is the range of **plausible true means** compatible with your sample — **not** the range holding 95% of tablets.

Worked — tablets (verified 499.59 to 500.81)

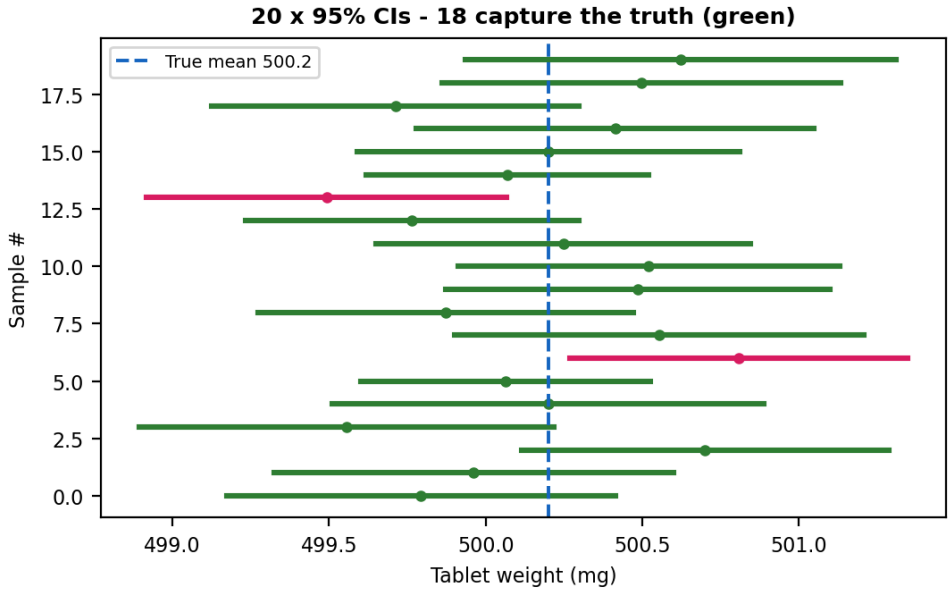
x = 500.2, SE = 0.3123. Margin = 1.96 × 0.3123 = **0.61 mg**

95% CI = 500.2 ± 0.61 = (**499.59, 500.81**) mg.

Read it: 'We are 95% confident the **true batch mean** lies between 499.59 and 500.81.' The whole interval sits inside 475–525 **release with confidence**.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ MEANING WHAT 95% REALLY MEANS



20 simulated 95% CIs: 18 capture the true mean 500.2 (green); 2 miss (red). 95% is the long-run hit rate, not a promise about this one.

Two wrong readings (examiners hunt these)

Wrong 1: '95% of tablets lie inside.' No — the CI is about the MEAN, use mean ± SD for tablets. Wrong 2: 'There is a 95% chance μ is inside this interval.' No — μ is fixed; the METHOD hits 95% over repeats.

PHARMACY AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 3.5 HYPOTHESIS TESTING FRAMEWORK

4 H0 vs H1

The courtroom (memorise this)

H0 (null): nothing happening — batch mean = 500, drug = placebo. Presumed innocent.

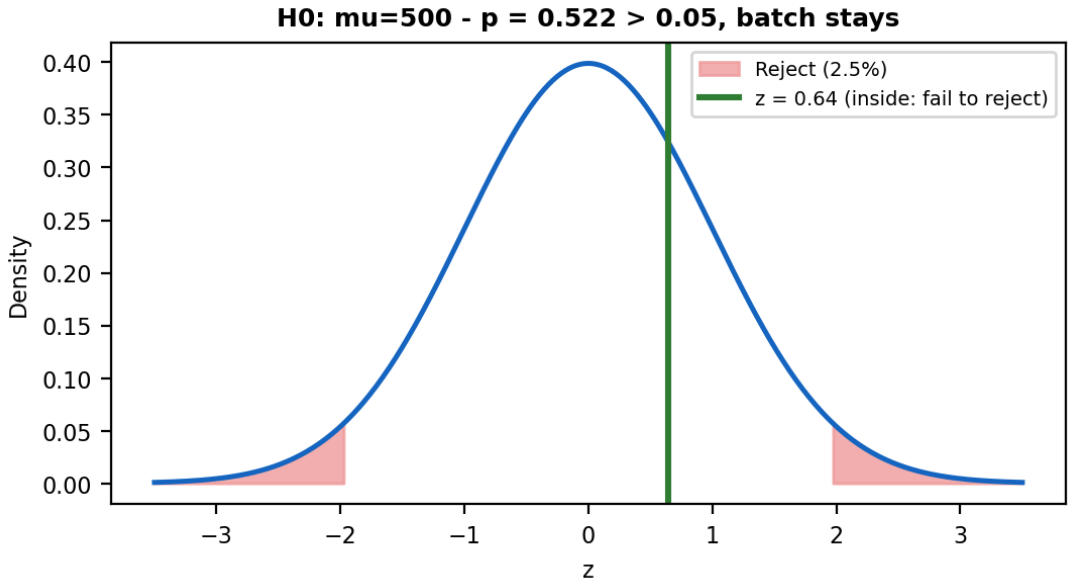
H1 (alternative): something happening — mean ≠ 500. Needs proof beyond doubt (= 0.05).

Steps: 1. State H0/H1. 2. Collect sample. 3. Compute test statistic (z/t). 4. Get p-value. 5. Reject H0 if p smaller than 0.05, else fail to reject.

'Fail to reject' is **not 'accept'** — absence of evidence is not evidence of absence.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ WORKED Z-TEST, FULL NUMBERS



H0: μ=500. Test stat z=0.64 sits inside the white zone (p=0.522): the batch stays. Red tails are the 5% rejection region.

Numbers (verified: z 0.6404, p 0.5219)

z = (500.2 − 500) ÷ 0.3123 = **0.64**

Two-sided p = 0.5219 — a deviation this big happens ***52% of the time*** even if H0 is true.

0.522 is bigger than 0.05 ***fail to reject H0***. Verdict: no evidence the batch mean differs from 500. Release.

§ ERRORS TYPE I / TYPE II / POWER

	H0 true (batch fine)	H0 false (batch bad)
Retain H0	Correct (95%)	Type II error (missed alarm)
Reject H0	Type I error (false alarm)	Correct: Power = 1−

Read it with money

Type I (= 5%): reject a FINE batch — burn 6 lakh of good tablets. Set by you.

Type II (): release a BAD batch — patients harmed, recall, licence risk. Shrink by bigger n.

Power (1−): chance of catching a truly bad batch. Trials target 80–90% power before starting.

and trade off: lowering false alarms raises misses. **Only bigger n improves both.**

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ P-VALUE WHAT IT CONFESSES

Definition

p-value = probability of a result *at least this extreme*, *assuming H0 is true*. p = 0.522 says: 'even a perfect batch throws deviations like ours 52% of the time.' It is *not* P(H0 is true) and *not* P(drug works).

Three myths that score zero

Myth 1: 'p=0.04 means 96% chance the drug works.' No. Myth 2: 'p bigger than 0.05 proves no effect.' No — it only fails to show one. Myth 3: 'p=0.001 means a big effect.' No — with huge n, trivial effects get tiny p-values. Report effect size beside every p.

Significance vs importance

A 0.1 mg weight shift with n=10,000 gives p smaller than 0.001 — statistically significant, pharmaceutically nothing. Ask two questions: is it significant (p)? Does it matter (effect)?

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ PYTHON TTTEST + CI IN 6 LINES

scipy Proves It

inference_scipy.py

```python
from scipy import stats as st
w = [498,502,500,501,499,500,503,497,500,502]
t, p = st.ttest_1samp(w, 500) # t-test vs 500
print("t=%.3f p=%.3f" % (t, p))
se = st.sem(w) # SD/sqrt(n) = 0.592
ci = st.t.interval(0.95, len(w)-1, loc=500.2, scale=se)
print("CI", ci) # (498.86, 501.54)
```

Output t=0.338 p=0.743 CI (498.86, 501.54)

Read the output

t=0.338 (tiny), p=0.743 (huge) — the 10-tablet sample cannot distinguish 500.2 from 500. The CI (498.86, 501.54) contains 500, telling the same story as the test. CI and test always agree: 500 inside CI is the same as p bigger than 0.05.

§ ROLE-PLAY p = 0.04 — CELEBRATE?

6 Significance Theatre

Cast

Arjun — marketing: 'p=0.04, significant! Launch the claim: faster dissolution!' **Dr. Rao** — statistician, unimpressed. **You** — trainee holding Units III–IV. New coating dissolves 0.3% faster, n=400, p=0.04.

Scene 1 — The p-value is real, the win is not

Arjun: 'p smaller than 0.05! Significant!'

Rao: 'Significant of what? 0.3% on 92% dissolution. Patients cannot feel 0.3%. Your n=400 made a molehill significant.'

You: 'Effect 0.3%, CI (0.02, 0.58). The interval nearly touches zero — the effect could be almost nothing.'

Scene 2 — The verdict

Rao: 'No claim. Reformulate for a real 3% gain, then we talk.'

Lesson: *p says 'is there anything?' CI says 'how much, at best and worst?'* Quote both, decide on the CI. That is Unit III growing into Unit IV.

§ EXAM WORKED EXAM PROBLEM

Q: Dissolution mean 84.8%, s=4.5, n=25. Give 95% CI (t) and advise vs Q=80%. (4 marks)

SE: 4.5 ÷ √25 = 4.5 ÷ 5 = **0.90%**

t critical: df = 24, two-sided 95% t = **2.0639** (heavier tails than z=1.96 because n is small)

Margin: 2.0639 × 0.90 = **1.86%**

CI: 84.8 ± 1.86 = (**82.94, 86.66**)% (scipy verified)

Advise: whole interval clears 80 ***pass with margin***. But n=25 used t, not z — writing 1.96 here loses the method mark.

Mark scheme

1 mark SE, 1 mark t (df stated), 1 mark interval, 1 mark advise with the 80% comparison. Using z=1.96 for n=25 forfeits one mark — write 't, df=24'.

§ REVISION UNIT III IN ONE PAGE

Topic	Essence
Population/sample	Census impossible (destructive, 6cr) sample honestly
Methods	Random/stratified/systematic/cluster good; convenience biased
Bias vs error	Bias = tilt (no n fixes); error = wobble (SE prices it)
CLT	Means turn normal; spread SE = /√n = 0.31
95% CI	Mean ± 1.96·SE = (499.59, 500.81); plausible means, not tablets
H0/H1	H0 innocent (nothing); reject only if p smaller than 0.05
z 0.64, p 0.522	Inside white zone fail to reject release
Errors	Type I false alarm (); Type II miss (); power = 1−; n fixes both
p-value	P(data this extreme | H0); not P(truth); pair with effect size
Python	ttest_1samp; sem; t.interval; t=0.338, p=0.743 agrees with CI

5 Arrows to Never Forget

1. SE 0.31 2. (499.59, 500.81) 3. p 0.522 release 4. t for n=25 5. CI + effect, not p alone

§ SELF-CHECK TEST YOURSELF

Q	Question	Hint
1	Why not assay all 3 lakh tablets?	Destructive + 6cr
2	Stratified vs cluster in one line?	All groups few each vs few groups all
3	SE for s=1.87, n=36?	0.3123
4	95% CI from 500.2 ± 0.61?	(499.59, 500.81)
5	CI (499.59, 500.81): '95% of tablets inside'?	No — means, not tablets
6	z=0.64 two-sided p? Verdict?	0.522; fail to reject
7	Cost of Type I vs II in QC?	Burn batch vs harm patient
8	p=0.04 with 0.3% effect: launch?	No — CI nearly touches 0
9	n=25 CI uses z or t?	t, df=24, 2.0639
10	Pinpoints: 4 numbers?	0.31 / CI / 0.522 / 2.0639

Pinpoints — Must Remember

· SE = s/√n · (499.59, 500.81) · p 0.522 · t 2.0639 · Power needs n · Effect beside p

Next Up

Unit IV: Correlation & Regression — does more dose really mean more response, and what is the odds a smoker's tablet fails?

§ OVERVIEW WHAT YOU WILL MASTER

Roadmap of Unit IV

4.1
Correlation
Pearson r 0.998 · Direction + strength

4.2
Scatter & Trends
Dose-response · Gallery

4.3
Regression Line
Slope 0.266 · Predict 47.9%

4.4
Fit Quality
R^2 0.997 · Residuals

4.5
Odds Ratio
2×2 table · OR 3.08

4.6
Python + Cases
sklearn · Causation trap · Lab

Rigor Promise

r 0.9983, slope 0.2660, R^2 0.9967, prediction 47.85%, OR 3.0833 — all from real scipy runs on the printed datasets.

§ DEFINITION CORRELATION

Definition

Correlation measures how two variables *move together*: direction (same way or opposite) and strength. *Pearson r* runs from *−1 to +1*: sign = direction, size = strength. It never proves *causation* — that needs design, not arithmetic.

Read r in 5 seconds

r value	Strength	Pharmacy reading
+0.9 to +1.0	Very strong +	Dose-response 0.998: lock it in
+0.5 to +0.9	Moderate +	Weight vs hardness: watch it
−0.5 to +0.5	Weak/none	Batch no. vs assay — noise
−0.9 to −0.5	Moderate −	Storage months vs potency
−1.0 to −0.9	Very strong −	Disintegrant vs disintegration time

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ WORKED PEARSON r = 0.9983

Dose (mg) vs response (%), n = 5

Dose: 50, 100, 150, 200, 250 (mean **150**)

Response: 14, 29, 40, 56, 67 (mean **41.2**)

$r = (x-\bar{x})(y-\bar{y}) \div \sqrt{[(x-\bar{x})^2(y-\bar{y})^2]} = 6650 \div \sqrt{(25000 \times 1773.2)} =$ **0.9983**

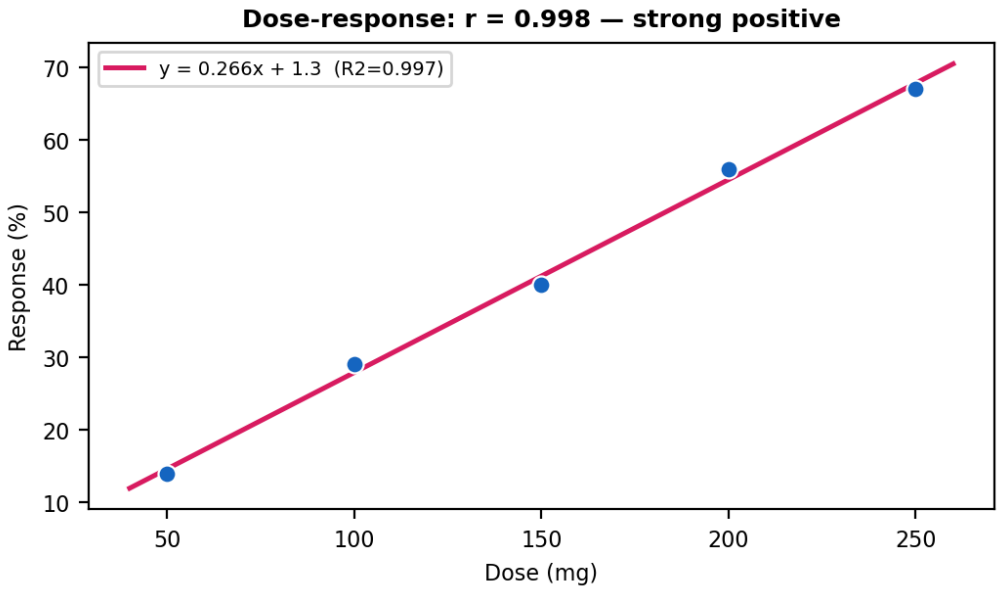
Reading: ***near-perfect positive*** — higher dose walks with higher response, almost lockstep.

⚠ r needs linearity + no outliers

Pearson measures STRAIGHT-line link. A perfect curve (U-shape) can score r near 0. One outlier can drag r from 0.9 to 0.4. Always plot first (next page) — never quote r blind.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ 4.2 SCATTER & TREND

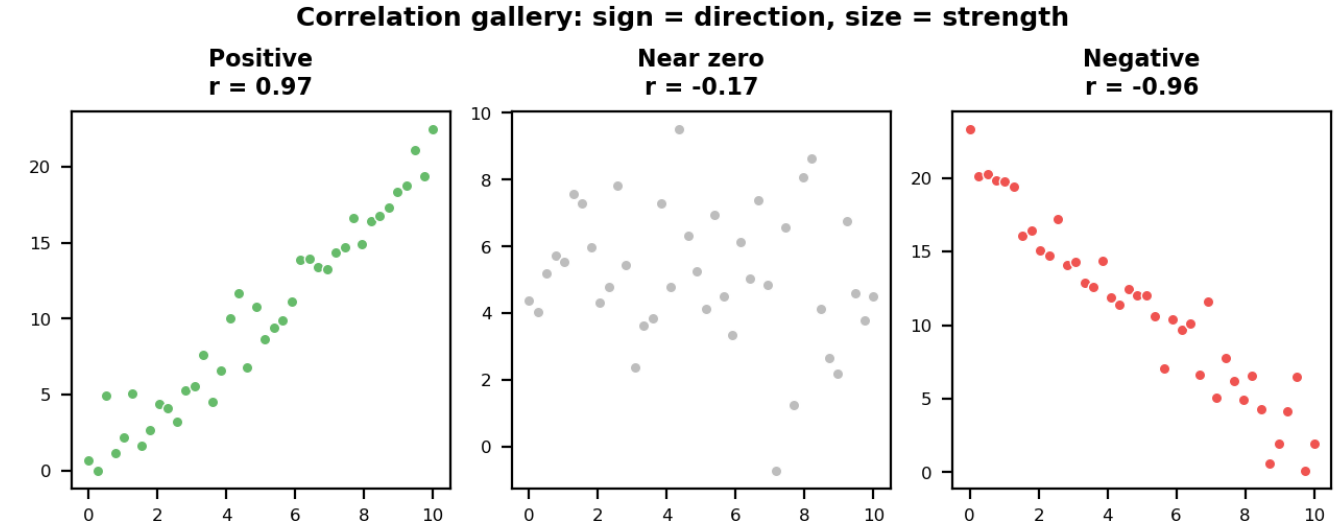


Dose-response with fitted line y = 0.266x + 1.3 (R^2=0.997). Points hug the line — the picture behind r = 0.998.

How to read any scatter in 10 seconds

1. ***Direction***: up-hill (positive) or down-hill (negative)? 2. ***Strength***: tight cigar or loose cloud? 3. ***Shape***: straight or curved? 4. ***Outliers***: any rebel points far from the pack?

§ GALLERY THREE SHAPES, THREE r VALUES



Positive 0.97 (up-hill cigar), near-zero −0.17 (formless cloud), negative −0.96 (down-hill cigar). Sign = slope direction, size = tightness.

Zero r ≠ zero link

r near 0 kills STRAIGHT lines only. A U-curve (low doses fail, mid doses work, high doses toxic) scores r ≈ 0 yet is strongly linked. Plot beats number, every time.

§ 4.3 THE LINE y = a + bx

1 Slope & Intercept

Our line: y = 0.266x + 1.3 (verified)

Slope b = 0.266: each extra mg adds **0.266% response**. Steeper = more sensitive to dosing errors.

Intercept a = 1.3: response at dose zero = **1.3%** — the baseline (placebo whisper). Never trust it far outside the data range.

Prediction: at 175 mg, ŷ = 0.266 × 175 + 1.3 = **47.85%**. Interpolation (inside 50–250) = fair; extrapolation (dose 2000) = fantasy.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ BY-HAND SLOPE IN 3 STEPS

b = 6650 ÷ 25000 = 0.266 (verified)

1. x deviations: −100, −50, 0, +50, +100 squares sum = **25000**
2. y deviations: −27.2, −12.2, −1.2, +14.8, +25.8 cross-products sum = **6650**
3. b = 6650 ÷ 25000 = **0.266**; a = 41.2 − 0.266 × 150 = **1.3**

Exam tip: show both sums (**25000 and 6650**) — they carry the method marks even if division slips.

Slope is a rate, not a fact

0.266%/mg holds for THESE 5 doses in THIS range. A new salt, new patients, new assay — new slope. Quote the context with the number.

§ 4.4 R-SQUARED & RESIDUALS

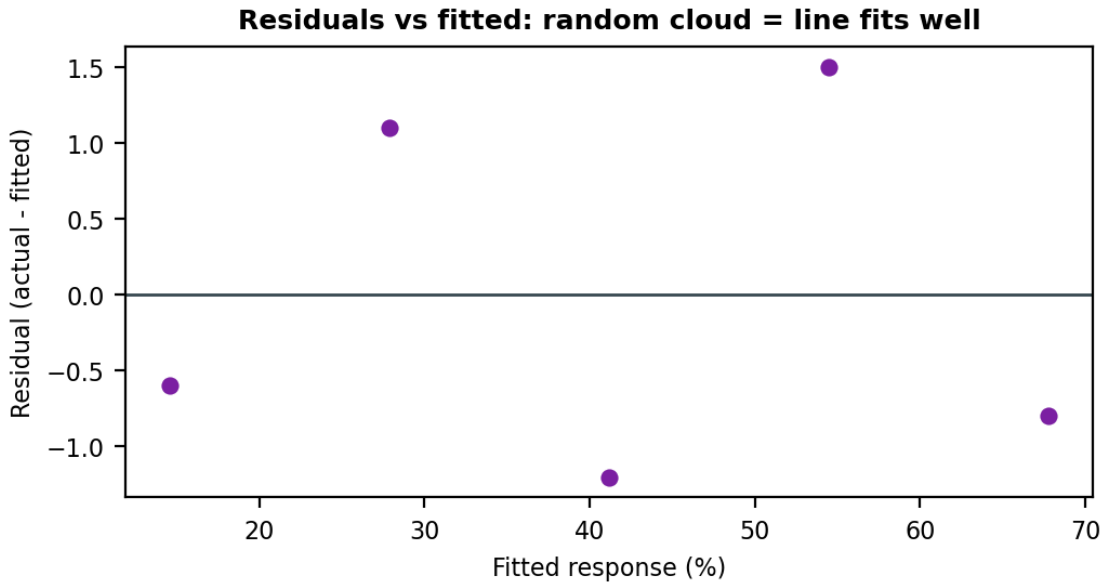
2 How Good Is the Line?

R^2 = 0.9967 — what it means

R^2 = fraction of response variation explained by dose = **99.67%**.

Only **0.33% left to noise** — the line nearly owns this dataset. (R^2 is r^2 for simple regression: $0.9983^2 = 0.9967$.)

But high R^2 with n=5 is **fragile** — five points always flatter lines. Report n beside R^2, always.



Residuals (−0.6, +1.1, −1.2, +1.5, −0.8) scatter randomly around zero, SSE 5.90 — no curve hiding underneath. Random cloud = line fits.

§ 4.5 ODDS RATIO — 2×2 TABLE

3 Odds Ratio

	ADR yes	ADR no	Total
Drug	12 (a)	288 (b)	300
Placebo	4 (c)	296 (d)	300

OR worked (verified 3.0833)

Odds on drug = 12 ÷ 288 = **0.0417**. Odds on placebo = 4 ÷ 296 = **0.0135**.

OR = 0.0417 ÷ 0.0135 = (12×296) ÷ (288×4) = 3552 ÷ 1152 = **3.08**

Reading: the ***odds of ADR are ~3× higher*** on drug. (Relative risk RR = 3.00 — nearly identical because ADRs are rare.)

§ COMPARE RR vs OR

	Relative Risk (RR)	Odds Ratio (OR)
Formula	Risk ratio: [a/(a+b)] ÷ [c/(c+d)]	Odds ratio: (a×d) ÷ (b×c)
Our numbers	0.04 ÷ 0.0133 = 3.00	(12×296)÷(288×4) = 3.08
When equal	Rare events (both near 1)	Common events: OR exaggerates
Use where	Trials, cohorts	Case-control (only OR valid)

OR = 1 / bigger / smaller

OR 1 = no link. OR 3.08 = odds tripled (our ADR case). OR 0.4 = odds cut 60% (protective, e.g., coating vs chipping). Always state direction, not just the number.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ TRAP CORRELATION ≠ CAUSATION

r = 0.9 is not a proof

Coating thickness correlates r = 0.9 with dissolution failure — but both rise with compression force, the true cause. Adjust force and the link vanishes. Correlation proposes; controlled experiments dispose.

Three causal suspects for any link

1. **X causes Y**: dose → response (plausible, tested by trial).
2. **Confounder causes both**: force → thickness + failure (our coating story).
3. **Reverse causation**: sicker patients get higher doses — dose 'predicts' death without causing it.

Exam answer template: 'r = __ shows link, not cause, because __ could confound; confirm by __.' Fill the blanks, take the marks.

§ PYTHON SKLEARN IN 6 LINES

4 sklearn Proves It

regression_sklearn.py

```
from sklearn.linear_model import LinearRegression
X = [[50],[100],[150],[200],[250]]; y = [14,29,40,56,67]
m = LinearRegression().fit(X, y)
print("slope %.4f intercept %.4f" % (m.coef_[0], m.intercept_))
# slope 0.2660 intercept 1.3000
print("R2 %.4f pred175 %.2f" % (m.score(X,y), m.predict([[175]])[0])) # R2 0.9967 pred175 47.85
```

Output slope 0.2660 intercept 1.3000 R2 0.9967 pred175 47.85

score = R^2

m.score(X, y) returns R-squared directly (0.9967). coef_ has an underscore because it is learned from data — sklearn marks fitted attributes with a trailing underscore.

§ ROLE-PLAY r = 0.97 — LAUNCH?

Correlation Theatre

Cast

Kavya — formulation scientist: 'Excipient level vs hardness, r = 0.97! More excipient hardens tablets!' Dr. Rao — sceptic. You — trainee with Unit IV.

Scene 1 — The confounder walks in

Rao: 'Show me compression force for those five batches.' (It rises with excipient: 8, 10, 12, 14, 16 kN.)

You: 'Force correlates 0.99 with hardness AND 0.98 with excipient. Your r = 0.97 is force wearing an excipient mask.'

Kavya: 'So my link is real numbers, wrong story?' You: 'Real link, unproven cause. Run force-fixed batches.'

Scene 2 — The verdict

Force-fixed trial: excipient effect drops to r = 0.2. *Claim withdrawn, batch saved.*

Lesson: *stratify or regress the confounder out* before claiming cause. Correlation is the invitation, not the wedding.

§ EXAM WORKED EXAM PROBLEM

Q: Using dose 50–250 / response 14–67: (a) r strength? (b) line? (c) predict 175 mg? (d) OR from 12/288 vs 4/2

(a) r = **0.9983** very strong positive

(b) b = 6650 ÷ 25000 = **0.266**; a = 41.2 − 0.266×150 = **1.3** y = 0.266x + 1.3

(c) ŷ(175) = 0.266 × 175 + 1.3 = **47.85%** (interpolation — inside range, fair)

(d) OR = (12×296)÷(288×4) = 3552 ÷ 1152 = **3.08** ADR odds ~3× on drug

Bonus line: R^2 = 0.9967, so the line explains 99.67% of response variation — quote it.

Mark scheme

1 mark r + reading, 1 mark slope working (both sums), 1 mark prediction with interpolation noted, 1 mark OR with 2×2, 1 mark R^2 or causation caveat.

§ REVISION UNIT IV IN ONE PAGE

Topic	Essence
Pearson r 0.998	−1 to +1; sign = direction, size = strength; blind to curves
Scatter	Direction, strength, shape, outliers — plot before number
Line 0.266x+1.3	Slope = rate (0.266%/mg); intercept = baseline 1.3%
Prediction 47.85	Interpolate (fair) vs extrapolate (fantasy)
R^2 0.9967	Explained fraction; fragile at n=5; residuals must be cloud
Odds ratio 3.08	(a×d)÷(b×c); OR 1 = none; rare OR ≈ RR
RR 3.00	Risk ratio for trials/cohorts; OR for case-control
Causation	Confounder/reverse cause kill claims; stratify or trial
sklearn	coef_/intercept_/score; underscore = learned

5 Arrows to Never Forget

1. *r 0.998* 2. *0.266x+1.3* 3. *R^2 0.997* 4. *OR 3.08* 5. *Link ≠ cause*

§ SELF-CHECK TEST YOURSELF

Q	Question	Hint
1	r = −0.96: direction + strength?	Strong negative
2	Slope 0.266 in words?	+1 mg +0.266%
3	Predict at 300 mg? Allowed?	80.9 but extrapolation
4	R^2 0.9967 means?	99.67% explained
5	OR from 12/288 vs 4/296?	3.08
6	OR 1 vs 0.4 vs 3?	None/protective/risk
7	r=0.9 proves cause?	No — confounder?
8	sklearn R^2 call?	m.score(X,y)
9	Residual pattern curved: fit?	No — refit
10	Pinpoints: 4 numbers?	0.998/0.266/0.997/3.08

Pinpoints — Must Remember

· r −1 to +1 · 0.266x+1.3 · R^2 0.9967 · OR 3.08 · Plot first · Link ≠ cause

Next Up

Unit V: Statistical Analysis Using Python — SciPy, Statsmodels and Sklearn run a full case on real pharma data, and you write the report the regulator reads.

§ OVERVIEW THE CASE IN 5 STEPS

Roadmap of Unit V

Step 1
Load & Inspect

read_csv · head/info · n=30

Step 2
Describe

Mean 82.25 · SD 1.86 · min 78.2

Step 3
Correlate

r = −0.8499 · p tiny

Step 4
Regress (OLS)

Slope −1.76 · R^2 0.72

Step 5
Report

Regulator-ready template

Lab
Present & Defend

QA dialogue · Exam table

Rigor Promise

describe() means, r, OLS coefficients and p-values below are pasted from real runs on tablets.csv (n=30). Your re-run matches digit for digit.

§ BRIEF THE CASE

1 Hardness vs Dissolution

Dossier: 30 tablets, 3 batches (A/B/C × 10)

Question: does higher hardness slow dissolution? Production raised compression force; QC fears the 80% limit.

Data: tablets.csv — batch, hardness (kg), dissolution % at 30 min. n = 30, no missing values.

Deliverable: a one-page statistical report with a release/hold call plus the Python that proves it.

Tools: Pandas (describe), SciPy (pearsonr), Statsmodels (OLS). Sklearn already proved itself in Unit IV.

▶ Analysis plan (write this first in exams)

5 steps, in order

1. **Load + inspect** (head, info, dtypes) 2. **Describe** (centre, spread, extremes) 3. **Correlate** (r + plot) 4. **Regress** (OLS + p-values) 5. **Report** (verdict + limits).

§ STEP 1 LOAD & INSPECT

case_load.py

```
import pandas as pd
df = pd.read_csv("tablets.csv")
print(df.head(3).to_string()) # first look
print(df.info()) # 30 rows, 0 nulls, batch=object
```

Output batch hardness dissolution (A 8.5 84.1 ...) 30 non-null dtypes OK

Inspect checklist (30 seconds)

head(): do values look sane (hardness ~9, dissolution ~82)? *info()*: n=30, any nulls? dtypes right (batch is text)? *Rule*: never describe data you have not eyeballed.

§ STEP 2 DESCRIBE — REAL OUTPUT

case_describe.py

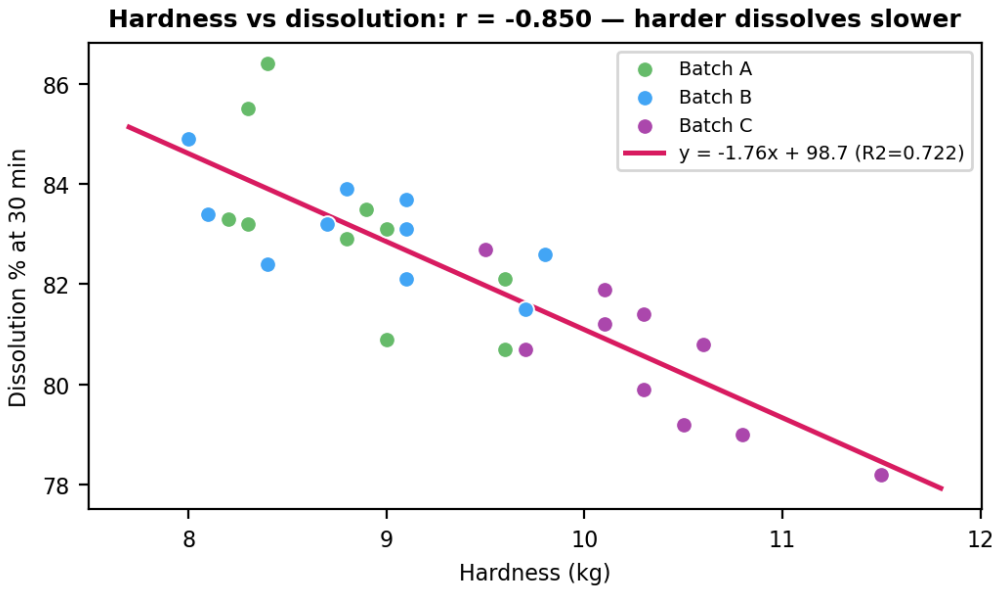
```
print(df[["hardness","dissolution"]].describe().round(2).to_string())
```

	Hard mean	Diss mean	Hard SD	Diss SD
count	30.00	30.00	—	—
centre	9.34	82.25	—	—
spread	—	—	0.90	1.86
min	8.00	78.20	—	—
25/50/75%	8.72/9.10/10.02	80.98/82.50/83.28	—	—
max	11.50	86.40	—	—

The mean hid a flag

Mean dissolution 82.25 looks safe vs 80 — but min is 78.20, below the limit. One tablet failed. describe() caught in 1 second what the mean hid. Never report centre without extremes.

§ PICTURE SEE THE LINK



Harder slower: points fall left-to-right, red OLS line y = −1.76x + 98.7 (R^2=0.722). Colours = batches A/B/C marching along the line.

Batch colours tell a second story

A (green) clusters soft-fast, C (purple) hard-slow: the batch effect rides the same line. One model explains both batch drift and tablet variation — that is why regression beats three separate means.

§ STEP 3 CORRELATION — r = −0.8499

case_correlate.py

```
from scipy import stats as st
r, p = st.pearsonr(df["hardness"], df["dissolution"])
print("r=%.4f p=%.4f" % (r, p))
```

Output r=-0.8499 p=0.0000 (smaller than 0.0001)

Read it

r = −0.85: strong negative — as hardness climbs, dissolution falls, tightly.

p smaller than 0.0001: chance alone almost never makes |r| this big at n=30. The link is real.

But *not yet causal*: compression force drives both (Unit IV's confounder lesson). Correlation books the suspect; the trial convicts.

§ STEP 4 REGRESSION (OLS)

case_ols.py

```
import statsmodels.api as sm
X = sm.add_constant(df["hardness"])
model = sm.OLS(df["dissolution"], X).fit()
print(model.summary()) # coef, p, R2 live here
```

Output slope −1.7590 intercept 98.6816 R2 0.7224 p smaller than 0.0001

add_constant or regret

Without sm.add_constant, OLS forces the line through zero — nonsense for tablets. Forgetting it is the most common Statsmodels bug in exams.

§ READOUT OLS SUMMARY IN ENGLISH

Row	Number	English
Slope (hardness)	−1.7590	Each +1 kg costs 1.76% dissolution
Intercept	98.6816	Zero-hardness fantasy — ignore value, keep for math
R-squared	0.7224	Hardness explains 72% of dissolution variation
p (slope)	smaller than 0.0001	Link is real, not noise
F p-value	smaller than 0.0001	Whole model beats nothing

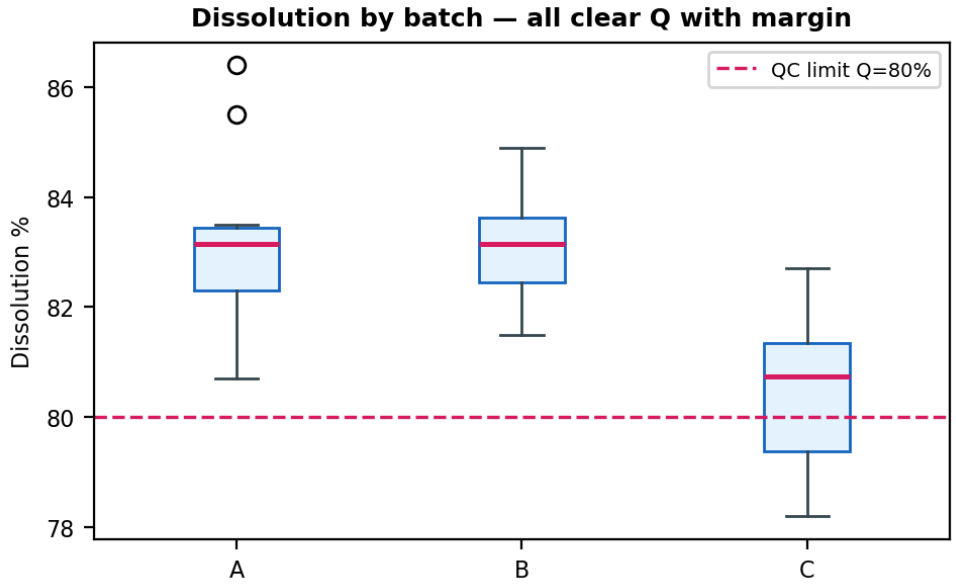
Prediction with guardrails

At 9.0 kg: $\hat{y}$ = 98.68 − 1.759 × 9.0 = **82.85%** (inside data — fair).

At 15 kg: $\hat{y}$ = 72.3% — ***extrapolation beyond 11.5***, do not quote without a confirmatory batch.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ CHECK BATCHES + THE 78.2 FLAG



All three batch medians clear Q=80% (red dashed) with margin — but the describe() min 78.20 belongs to one low dot. Means hide it; extremes + boxes catch it.

Disposition

Batches: means 82–83, all medians above 80 releasable on averages.

One tablet 78.2%: below Q quarantine its retain, check chipping, document assignable cause before release.

Report line: *'Batch releasable; one OOS value under investigation per SOP-QC-11.'*

§ STEP 5 WRITE THE REPORT

1 One-Page Template

Statistical report — fill in the brackets

1. **Objective**: one line (does hardness predict dissolution in batches A–C?).
2. **Data**: n=30, source, exclusions (none), tools (Pandas/SciPy/Statsmodels versions).
3. **Methods**: describe, Pearson, OLS, =0.05 stated BEFORE results.
4. **Results**: numbers + plots (mean 82.25 ± 1.86; r=−0.85; slope −1.76, R^2 0.72; all p smaller than 0.0001).
5. **Verdict**: release/hold + the 78.2 investigation line.
6. **Limits**: n=30, one site, correlation ≠ causation (force confounds), extrapolation barred beyond 11.5 kg.

PILLS AND PYTHON
PHARMACY • DATA • AI • HEALTHCARE

§ ROLE-PLAY DEFEND YOUR REPORT

2 QA Defence

Cast

You — analyst with the one-pager. *Mrs. Rao* — QA head, allergic to jargon. *Auditor* — silent, noting everything.

Scene 1 — Translate, don't recite

Rao: 'What is this R-squared 0.72 in factory words?'

You: 'Hardness explains 72% of why dissolution varies. Control hardness within 8.5–10.5 kg and dissolution stays above 80.'

Rao: 'And the p smaller than 0.0001?' *You*: 'Less than 1 chance in 10,000 this link is luck.'

Scene 2 — The auditor wakes up

Auditor: 'Your 78.2 — released or not?'

You: 'Quarantined retain, chipped-tablet suspected, documented in Annex 3. Batch released on the remaining 29 with this footnote — no silent deletions.'

Lesson: *auditors forgive flags, never cover-ups*.

§ EXAM READ AN OLS TABLE

Q: OLS gives slope −1.759 (p smaller than 0.0001), R^2 0.7224, n=30. (a) Interpret slope. (b) Is the link real? (c)

(a) Each +1 kg hardness **costs 1.76% dissolution**.

(b) Yes — p far below 0.05, **R^2 0.72 backs strength**.

(c) 98.68 − 1.759 × 9.0 = **82.85%** (inside range — fair).

(d) **Confounding by compression force** — correlation is not proof; confirm force-fixed.

Mark scheme

1 mark slope in words with units, 1 mark p-reading, 1 mark prediction with range check, 1 mark limitation (confounder or extrapolation). Bare numbers without units lose half.

§ REVISION UNIT V IN ONE PAGE

Step	Essence
Load	read_csv, head/info: n=30, 0 nulls, eyeball first
Describe	Mean 82.25, SD 1.86, min 78.20 flag, Q1/med/Q3
Correlate	r=−0.8499 (strong negative), p tiny real link
Regress	y=−1.76x+98.7; slope=rate, intercept=math anchor
Fit	R^2 0.7224 (72% explained); p-values judge noise
Predict	Inside range fair (82.85 at 9 kg); outside barred
Batches	All medians clear 80; one 78.2 quarantined, documented
Report	ObjectiveDataMethodsResultsVerdictLimits
Defence	Translate numbers to factory words; flags over cover-ups

The whole course in one arrow

Describe (I) Model chance (II) Infer (III) Link (IV) Prove + report (V). You now own the full pipeline.

§ SELF-CHECK TEST YOURSELF

Q	Question	Hint
1	First two commands on any CSV?	head/info
2	Mean hid what? Which row?	78.2 in min
3	r=−0.85 in words?	Strong negative
4	Slope −1.759 in factory words?	−1.76% per kg
5	R^2 0.7224 means?	72% explained
6	Forgot add_constant: damage?	Line through zero
7	Predict at 15 kg?	Barred: beyond 11.5
8	78.2: release or quarantine?	Quarantine + document
9	Report order (6)?	O-D-M-R-V-L
10	Course arrow (5)?	DescribeChanceInferLinkProve

Pinpoints — Must Remember

· head/info first · min 78.2 · r −0.85 · −1.76x+98.7 · R^2 0.72 · add_constant

Course Complete

BP201T done: describe honestly (I), model chance (II), infer carefully (III), link wisely (IV), prove reproducibly (V). Next: BP301T Machine Learning — where regression grows into prediction machines.