B.PHARM SEMESTER II · BP201T · UNIT IV

BP201T Applied Biostatistics and Data Analytics for Pharmaceutical Sciences

Unit IV — Correlation & Regression

Final Study Notes

Core Theory · 6 Hours

From-zero explanations · Worked pharmaceutical examples · Python visualisation · Exam revision

Pills & Python

Pharmacy • Data • AI • Healthcare

Chapter 4 — Basics of Correlation & Regression (Unit IV, 6 hours)

4.1 Learning objectives

After studying this chapter, you will be able to:

1. Define correlation in plain words.
2. Interpret the Pearson correlation coefficient's sign and size.
3. Read and draw a scatter plot, including dose-response relationships.
4. Explain the concept of simple linear regression and interpret its coefficients.
5. Define odds ratio and apply it to clinical risk analysis.
6. Warn against the "correlation implies causation" trap.

4.2 What is correlation?

Definition: Correlation measures how strongly two variables move *together* — and in which direction. The **Pearson correlation coefficient (r)** is a number between **−1 and +1**:

- **r = +1:** perfect positive — as one rises, the other rises in lockstep.
- **r = −1:** perfect negative — as one rises, the other falls in lockstep.
- **r = 0:** no linear relationship.

Interpreting the size (rough guide): |r| near 0.1 = weak, near 0.3–0.5 = moderate, above 0.7 = strong. The **sign** tells direction; the **size** tells strength.

> **Worked mini-example (illustrative):** Dose of antihypertensive (mg) vs BP reduction (mmHg): (10, 5), (20, 9), (30, 12), (40, 16), (50, 19).

- Step 1 — means: $\bar{x} = 30$, $\bar{y} = 61 \div 5 = 12.2$.
- Step 2 — deviations and products:

x	y	x − x̄	y − ȳ	(x−x̄)(y−ȳ)	(x−x̄)²	(y−ȳ)²
10	5	−20	−7.2	144	400	51.84
20	9	−10	−3.2	32	100	10.24
30	12	0	−0.2	0	0	0.04
40	16	10	3.8	38	100	14.44
50	19	20	6.8	136	400	46.24
Totals				**350**	**1000**	**122.8**

- Step 3: $r = 350 \div \sqrt{(1000 \times 122.8)} = 350 \div \sqrt{122800} = 350 \div 350.43 \approx$ **0.999**.
- Interpretation: an almost perfect **positive** correlation — higher dose, bigger BP drop, moving together.

> ⚠️ **Terminology alert:** Pearson's r measures **linear** relationships only. A perfect U-shaped curve can have r ≈ 0 despite a strong relationship — always plot your data.

Analogy: Correlation is like *two friends walking*: positive = walking the same direction arm-in-arm; negative = walking opposite directions; zero = wandering randomly, ignoring each other.

Key takeaway: r's sign = direction, r's size = strength, and r only sees straight-line patterns.

4.3 Positive vs negative correlation — pharmacy examples

Pair of variables	Expected sign	Why
Drug dose → BP reduction	Positive (+)	More drug, bigger effect
Patient age → drug clearance rate	Negative (−)	Older kidneys clear slower
Tablet hardness → friability	Negative (−)	Harder tablets chip less
Number of prescribing errors → pharmacist experience	Negative (−)	Experience reduces errors

Key takeaway: Before calculating, *predict* the sign from pharmacology — if the sign surprises you, investigate.

4.4 Scatter plots — see it before you compute it

Definition: A scatter plot puts one variable on the x-axis, the other on the y-axis, one dot per observation. The cloud's shape reveals the relationship instantly.

Dose-response example (illustrative): Plot the five (dose, BP reduction) points from §4.2 — the dots climb steadily up-right: the classic dose-response picture every pharmacologist recognises. A flat cloud = no relationship; a downward cloud = negative correlation; a curved cloud = non-linear (Pearson r will understate it).

Python snippet (standard pattern; illustrative data):

```
import matplotlib.pyplot as plt

dose = [10, 20, 30, 40, 50]
bp_drop = [5, 9, 12, 16, 19]
plt.scatter(dose, bp_drop)          # one dot per (dose, response) pair
plt.xlabel("Dose (mg)")             # x-axis label
plt.ylabel("BP reduction (mmHg)")   # y-axis label
plt.title("Dose-response scatter plot")
plt.show()                          # display the figure
```

Interpretation: The dots rise from bottom-left to top-right — visual proof of positive correlation before any number is computed.

Key takeaway: Plot first, calculate second. The eye catches curves and outliers that r hides.

4.5 Simple linear regression — drawing the best line

Concept: While correlation describes *strength*, **regression** gives you an *equation* for prediction: the "line of best fit" through the scatter cloud. $\hat{y} = a + bx$

- **b (slope):** how much y changes for a 1-unit increase in x.
- **a (intercept):** the predicted y when x = 0.
- The line is chosen to minimise the squared vertical distances (the "least squares" principle — the fairest single line through the cloud).

Worked example (illustrative, same dose data):

- Step 1 — slope: $b = \Sigma(x-\bar{x})(y-\bar{y}) \div \Sigma(x-\bar{x})^2 = 350 \div 1000 =$ **0.35**.
- Step 2 — intercept: $a = \bar{y} - b\cdot\bar{x} = 12.2 - 0.35 \times 30 = 12.2 - 10.5 =$ **1.7**.
- Equation: **BP reduction = 1.7 + 0.35 × dose**.
- Step 3 — interpretation: each extra mg of dose adds ~0.35 mmHg of BP reduction; at zero dose the model predicts 1.7 mmHg (a mathematical anchor, not a real measurement — never extrapolate far beyond your data).
- Step 4 — prediction: at 60 mg, predicted reduction = 1.7 + 0.35 × 60 = 1.7 + 21 = **22.7 mmHg**.

⚠️ **Terminology alert — regression vs correlation:** Correlation (r) is symmetric — r(x,y) = r(y,x). Regression is *directional*: regressing y on x answers "predict y from x" and gives a different line than regressing x on y.

Analogy: Correlation tells you *how tightly* the friends hold hands; regression draws the *actual path* they walk so you can predict where they'll be.

Key takeaway: $\hat{y} = a + bx$ turns "related" into "predictable" — slope = effect per unit, intercept = starting point.

4.6 Odds ratio — risk in clinical analysis

Concept: The **odds ratio (OR)** compares the *odds* of an outcome in an exposed group versus an unexposed group. From a 2×2 table:

	Disease YES	Disease NO
Exposed	a	b
Unexposed	c	d

$$OR = (a \times d)/(b \times c)$$

Definitions: *Odds* = P(event) ÷ P(no event). OR = 1 → no association; OR > 1 → exposure linked to *higher* odds of disease (risk factor); OR < 1 → exposure linked to *lower* odds (protective).

Worked example (illustrative): Smoking and lung disease — smokers: 40 diseased (a), 60 healthy (b); non-smokers: 20 diseased (c), 180 healthy (d).

- Step 1: OR = (40 × 180) ÷ (60 × 20) = 7200 ÷ 1200 = **6.0**.
- Interpretation: smokers have **6 times the odds** of lung disease compared with non-smokers — a strong risk signal.

⚠️ **Terminology alert:** Odds ratio ≠ relative risk, though both exceed 1 for risk factors. OR is the standard measure in case-control studies, where true risks cannot be computed directly.

Key takeaway: OR = (a×d)/(b×c); above 1 = risk factor, below 1 = protective, equals 1 = no link.

4.7 Correlation is not causation — the golden warning

Ice-cream sales correlate with drowning deaths — not because ice cream kills, but because **summer** drives both (a *confounding* variable). In pharmacy: patients on stronger drugs may have worse outcomes *because they were sicker to begin with*, not because the drug harms. Only well-designed studies (like randomised trials) can claim causation.

Mnemonic: "**C**orrelation **C**ries, **C**ausation **C**onvinces" — or the classic: *roosters don't cause sunrises.*

Key takeaway: Treat every correlation as a suspect, not a convict — ask "what third factor could explain this?"

4.8 Pharmacy relevance (Unit IV)

- **Dose-response:** the foundation of pharmacology — regression quantifies "how much effect per mg."
- **PK/PD modelling:** drug concentration vs effect relationships.
- **Epidemiology:** odds ratios from case-control studies drive drug-safety decisions.

4.9 Common confusions

1. **r vs slope:** r (−1 to +1) measures strength; slope (any units) measures steepness. A steep line can have a weak r if data are scattered.
2. **Negative correlation** is still a correlation — "no correlation" means r ≈ 0.
3. **Intercept** is often not physically meaningful (dose = 0 may lie outside the data).
4. **Odds vs probability:** odds of 6:1 ≠ probability of 6.

4.10 Common mistakes

- Claiming causation from correlation.
- Extrapolating the regression line far beyond observed x-values.
- Computing Pearson r on curved relationships without plotting.
- Interpreting OR = 2 as "twice the risk" (it is twice the *odds*).

4.11 One-minute revision (Unit IV)

- **r ∈ [−1, +1]**: sign = direction, |size| = strength; linear only — plot first.
- **Scatter plot**: dose-response clouds; shape reveals what numbers hide.
- **Regression**: $\hat{y} = a + bx$; **b** = change in y per unit x; **a** = y at x = 0; least squares picks the line.
- Worked: b = 350/1000 = **0.35**, a = 12.2 − 10.5 = **1.7** → $\hat{y} = 1.7 + 0.35x$.
- **Odds ratio** = (a×d)/(b×c): >1 risk factor, <1 protective, =1 no association.
- **Correlation ≠ causation** — beware confounders.

4.12 Practice questions (Unit IV)

MCQs

1. r = −0.85 indicates:

a) Strong positive b) Strong negative c) Weak negative d) No correlation

2. In ŷ = 2 + 0.5x, the slope means:

a) y is 2 when x is 0 b) y rises 0.5 per unit x c) r = 0.5 d) x rises 0.5 per unit y

3. OR = 0.4 for a new drug vs disease means the drug is:

a) A risk factor b) Protective c) Unrelated d) Harmful

4. Pearson's r is inappropriate when the relationship is:

a) Strong b) Negative c) Curved/non-linear d) Positive

Short questions

5. Interpret r = +0.92 and r = −0.15 in one line each, with a pharmacy example.
6. Distinguish correlation from regression.
7. Define odds ratio and interpret OR = 3.0.

Long / application-based

8. (Illustrative) Using the dose-response data in §4.2, recompute r, the regression line, and predict BP reduction at 45 mg — showing every step.
9. A study finds coffee drinking correlated with heart disease. Give two confounders and explain why causation cannot be claimed.
10. (Illustrative) In a case-control study: exposed cases 30, exposed controls 90, unexposed cases 10, unexposed controls 150. Compute OR with steps and write a one-line clinical interpretation.

COMBINED REFERENCES

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