

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Covariance, Correlation & Graphic Displays of Statistics

Lecture 4 · Wednesday 23 September 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–5 min) Recap; the question of two attributes
2. (5–20 min) Covariance from zero, by hand on 5 students
3. (20–35 min) Pearson correlation r: build, interpret, apply to 500 farms
4. (35–45 min) Correlation is not causation: confounders
5. (45–65 min) Graphic displays: histogram, boxplot, quantile, Q-Q and scatter plots
6. (65–70 min) Correlation at scale: one pass
7. (70–75 min) Exercise and summary

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Compute	covariance and Pearson's r by hand, step by step
Interpret	the sign and size of r, and spot when r misleads
Explain	why correlation is not causation, naming a confounder
Choose	the right plot: histogram, boxplot, quantile, Q-Q or scatter
Explain	how r can be computed in one pass over a billion rows

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

→ **Recap and the Question: Do Two Attributes Move Together?**

Covariance from Zero

Pearson Correlation r

Correlation Is Not Causation

Graphic Displays: Seeing the Data

Correlation at Scale

In-Class Exercise

Recap: One Attribute at a Time

IIT Madras Zanzibar

1. **Centre**
   farm size: mean 1.46 ha, median 0.96 ha

2. **Spread**
   rainfall: std. dev. 243 mm

3. **Shape**
   mean ≫ median → right skew → report the median

Today's question: when one attribute goes up, does the other go up, go down, or not care?

Examples: price ↔ revenue · rainfall ↔ yield · study hours ↔ quiz score · airtime top-ups ↔ mobile-money use

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

Recap and the Question: Do Two Attributes Move Together?

→ **Covariance from Zero**

Pearson Correlation r

Correlation Is Not Causation

Graphic Displays: Seeing the Data

Correlation at Scale

In-Class Exercise

A Small Example Everyone Knows: Study Hours

Student	Hours x	Score y
1	1	40
2	3	50
3	5	70
4	7	60
5	9	80
mean	**$\bar{x} = 5$**	**$\bar{y} = 60$**

1. $\bar{x}$
   $(1 + 3 + 5 + 7 + 9) \div 5 = 25 \div 5 = 5$ hours

2. $\bar{y}$
   $(40 + 50 + 70 + 60 + 80) \div 5 = 300 \div 5 = 60$ marks

3. **Question**
   one number for 'how strongly do x and y move together?'

The Idea: Above Average Together?

IIT Madras Zanzibar

below $\bar{x}$, above $\bar{y}$

$(-) \times (+) = -$

above $\bar{x}$, above $\bar{y}$

$(+) \times (+) = +$

below $\bar{x}$, below $\bar{y}$

$(-) \times (-) = +$

above $\bar{x}$, below $\bar{y}$

$(+) \times (-) = -$

Covariance = the average of $(x - \bar{x}) \times (y - \bar{y})$. Mostly positive products → positive covariance → they move together.

Covariance Step by Step

$$\mathrm{Cov}(x, y) = \Sigma\,(x_i - \bar{x})(y_i - \bar{y}) \div n$$

some texts divide by n − 1; the sign and r are unchanged

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	product
1	40	−4	−20	80
3	50	−2	−10	20
5	70	0	10	0
7	60	2	0	0
9	80	4	20	80
			sum	**180**

1. **Divide by n = 5**
   Cov = 180 ÷ 5 = 36

2. **Sign +**
   they move together

Problem: units are hours × marks. Measure in minutes and Cov = 2,160 — nothing real changed.

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

Recap and the Question: Do Two Attributes Move Together?

Covariance from Zero

→ **Pearson Correlation r**

Correlation Is Not Causation

Graphic Displays: Seeing the Data

Correlation at Scale

In-Class Exercise

Pearson's r: Covariance Without Units

$$r = \mathrm{Cov}(x, y) \div (\sigma_x \times \sigma_y)$$

σ = standard deviation (divide by n, matching the covariance); −1 ≤ r ≤ +1

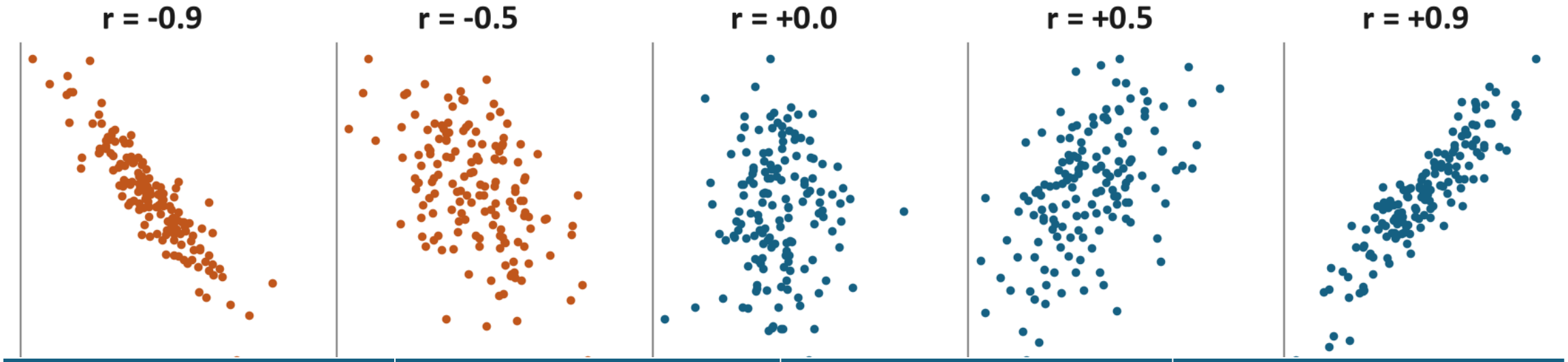
1. **σ_x**
   √[(16 + 4 + 0 + 4 + 16) ÷ 5] = √8 = 2.83 hours

2. **σ_y**
   √[(400 + 100 + 100 + 0 + 400) ÷ 5] = √200 = 14.14 marks

3. **r**
   36 ÷ (2.83 × 14.14) = 36 ÷ 40 = 0.90

4. **In minutes**
   2,160 ÷ (169.7 × 14.14) = 0.90 — unchanged

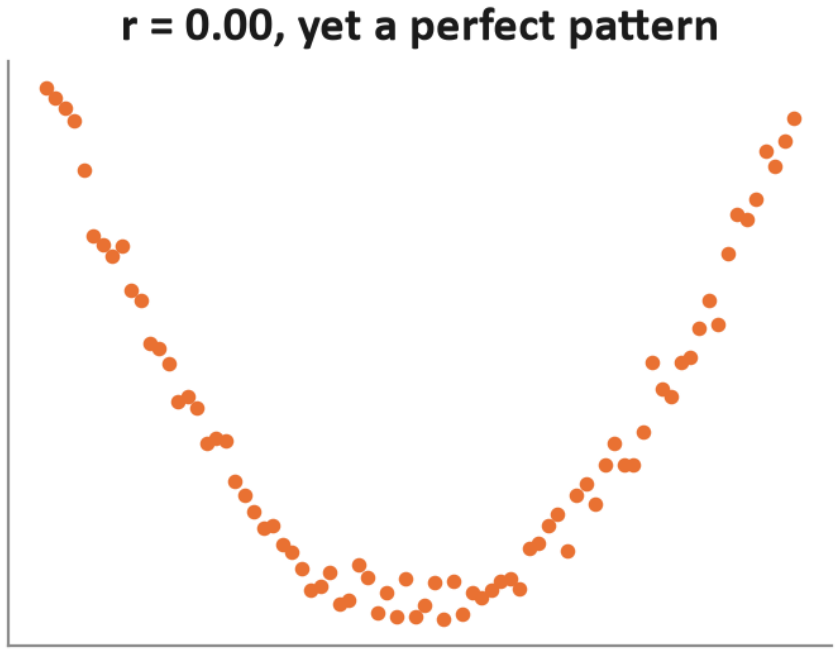
Reading r: Sign and Strength



|r|	below 0.3	0.3 – 0.7	above 0.7
Strength (rough guide)	weak	moderate	strong

Sign: + → up together; − → one up, the other down. Size: how tightly the points hug a straight line.

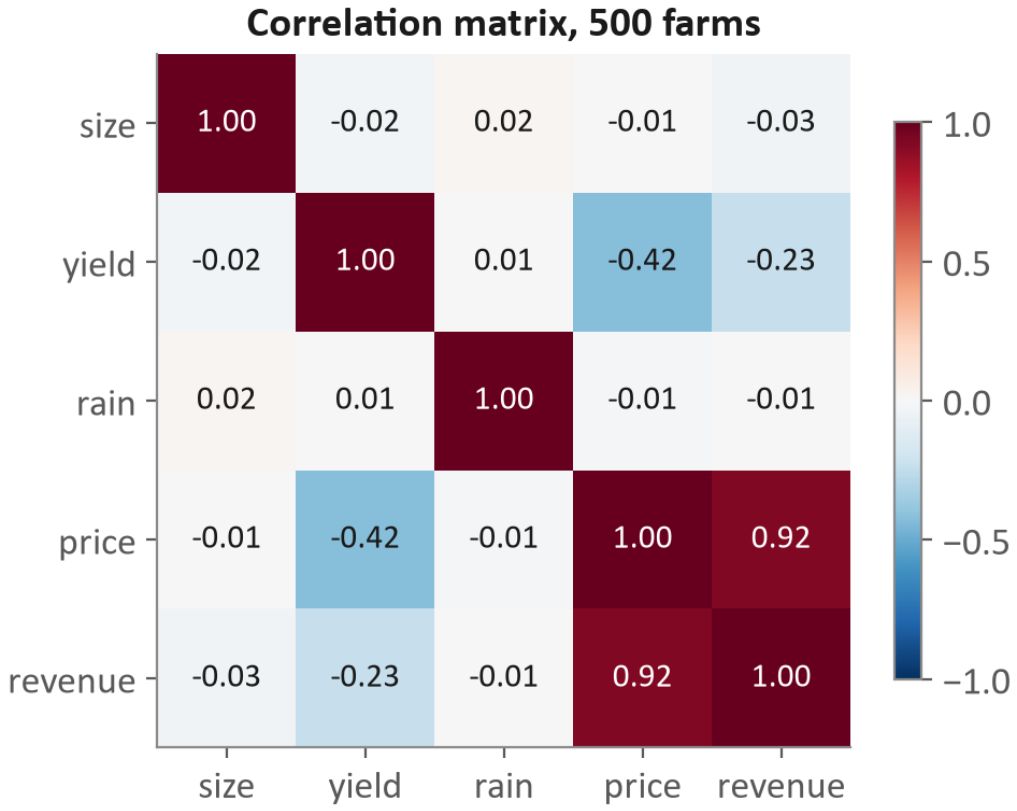
The Trap: r Measures Only Straight Lines



1. **Perfect pattern**
   $y = x^2$ — fully predictable
2. **$r \approx 0$**
   positive and negative products cancel
3. **Real case**
   yield vs temperature: too cold and too hot both hurt
4. **Rule**
   always plot before trusting r (Anscombe, 1973)

r = 0 means: no STRAIGHT-LINE relationship — not 'no relationship'.

r on Our 500 Farms: Three Stories



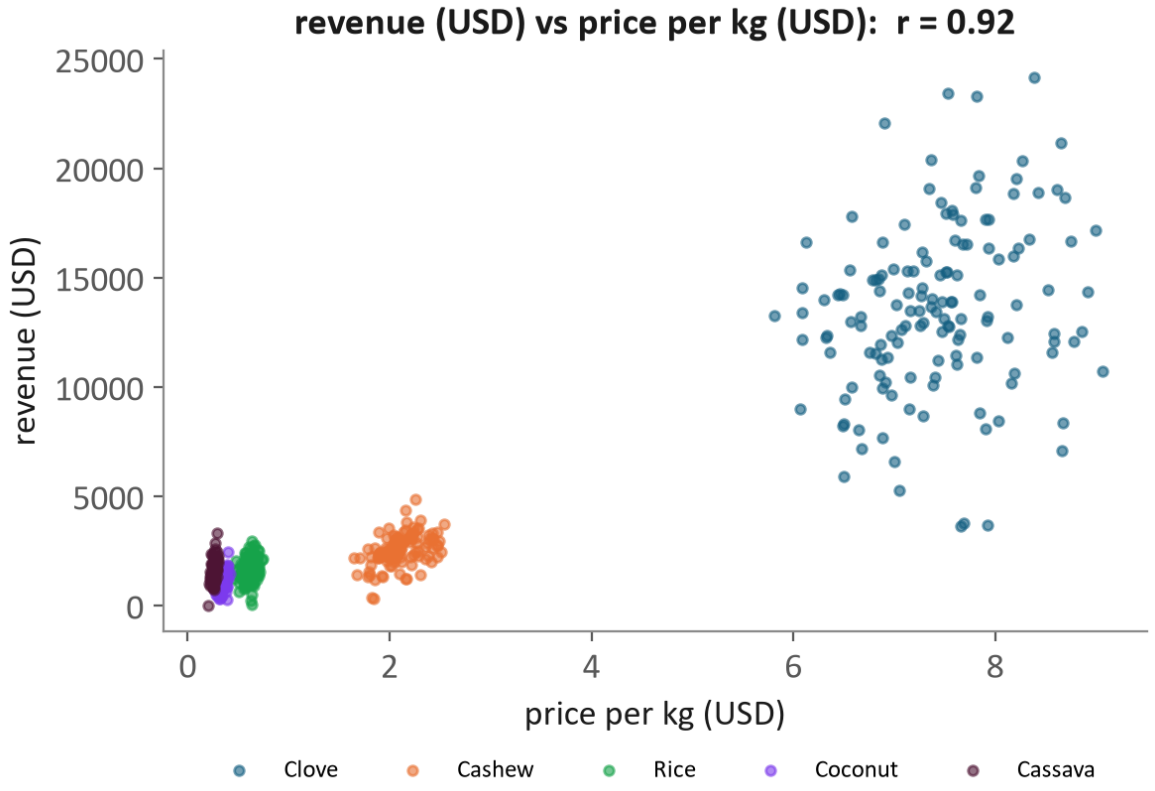
1. **price ↔ revenue**
   r = 0.92: strong positive

2. **price ↔ yield**
   r = -0.42: moderate negative

3. **rainfall ↔ yield**
   r = 0.01: no linear link here

Read the matrix: symmetric (r(x, y) = r(y, x)); diagonal = 1; colour = sign and strength.

Seeing the Strong One: Price and Revenue



1 **Each dot**
one farm, coloured by crop

2 **Two clouds**
clove ($6–9/kg) forms the high cloud; other crops sit bottom-left

3 **Warning**
r is driven by differences between crops

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

Recap and the Question: Do Two Attributes Move Together?

Covariance from Zero

Pearson Correlation r

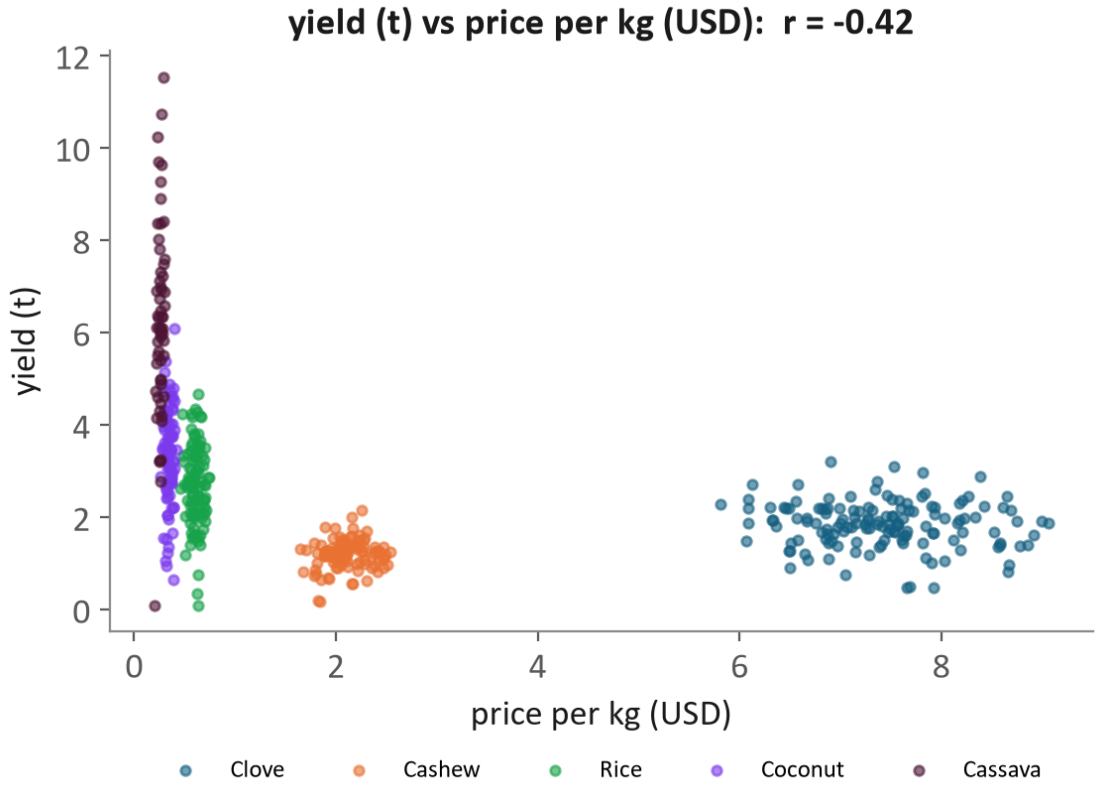
→ **Correlation Is Not Causation**

Graphic Displays: Seeing the Data

Correlation at Scale

In-Class Exercise

Price and Yield: Is the Negative r 'Real'?



1. **Observed**
   r = -0.42: higher price, lower yield

2. **Confounder**
   crop decides both price and tonnes

3. **Causation?**
   no — raising the price would not shrink the harvest

Confounder: a third attribute that drives both — always ask 'what else could explain this?'

Correlation Is Not Causation

IIT Madras Zanzibar

1 Ice cream & drownings

both rise in hot weather — the confounder

2 Clinics & malaria cases

both larger where population is larger

3 Fire trucks & damage

big fires bring more trucks and more damage

4 Price & yield (our data)

crop drives both

5 Reverse cause

sick people visit doctors more; doctors do not cause sickness

6 Pure chance

search 1,000 pairs and some correlate by luck

Common thread: to claim cause, you need an experiment or a strong argument that rules out confounders.

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

Recap and the Question: Do Two Attributes Move Together?

Covariance from Zero

Pearson Correlation r

Correlation Is Not Causation

→ **Graphic Displays: Seeing the Data**

Correlation at Scale

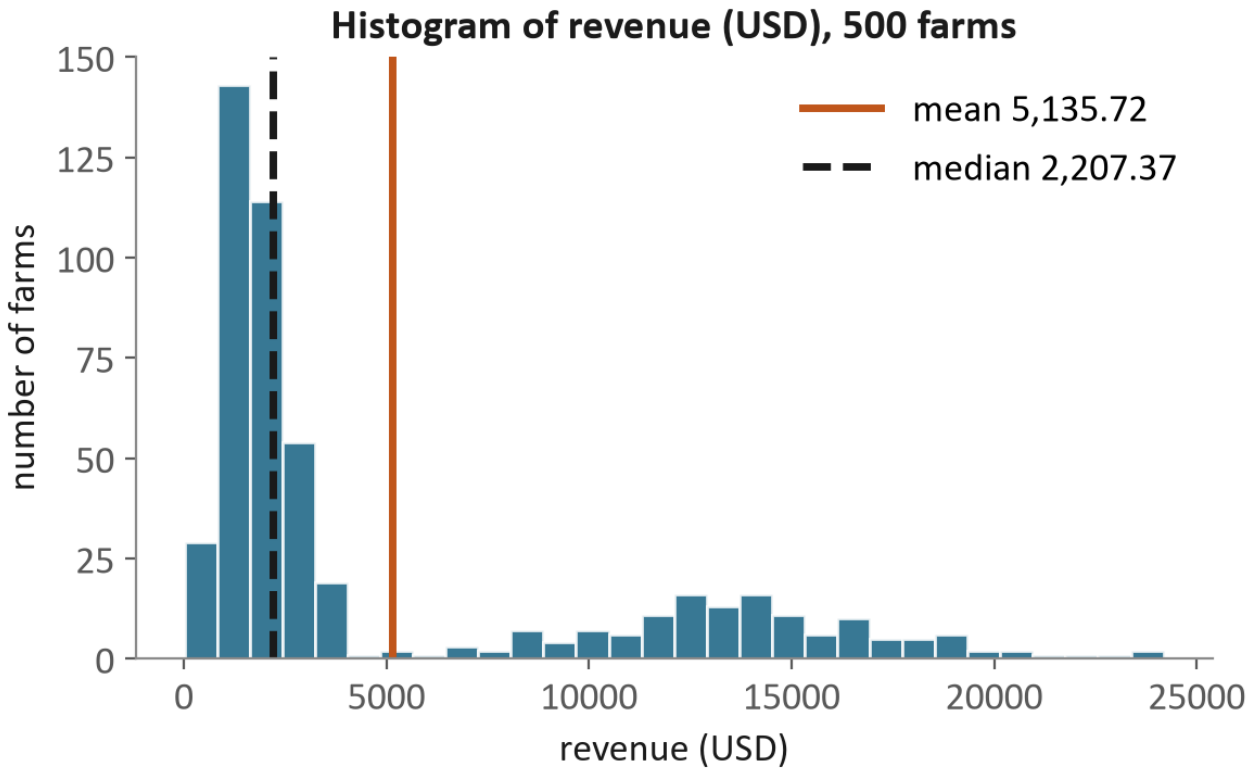
In-Class Exercise

Why Plot? Five Standard Displays

IIT Madras Zanzibar

Display	Shows	Use it when
Histogram	shape of one numeric attribute	is it symmetric or skewed? gaps?
Boxplot	five-number summary + outliers	comparing groups (crop, region)
Quantile plot	each value against its rank (f-value)	reading percentiles directly
Q-Q plot	quantiles of one distribution vs another	are two groups or a group and a normal curve alike?
Scatter plot	two numeric attributes together	before computing or trusting r
Bar chart	counts of a nominal attribute	how many farms per crop

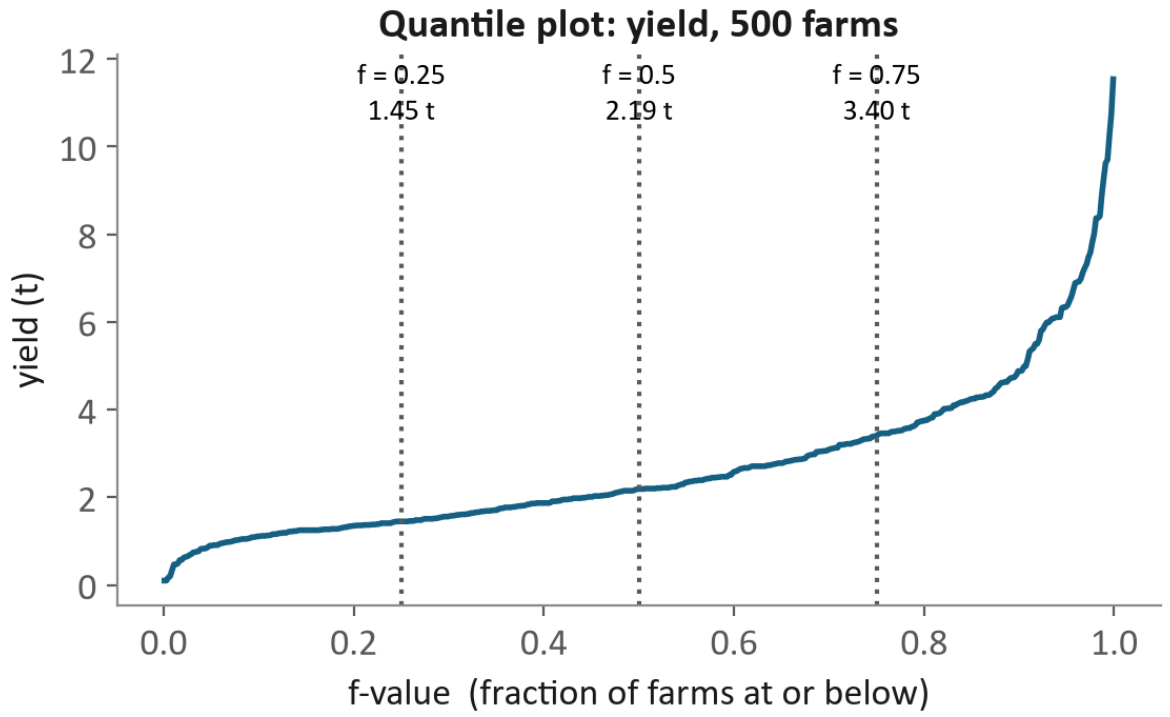
Histogram: The Shape of One Attribute



1. **Bins**
   cut the range into equal widths
2. **Count**
   how many values fall in each bin
3. **Draw**
   bar height = count; bars touch
4. **Read**
   two humps: other crops vs clove

Bin width: too few bins hide the shape; too many show noise. Try several.

Quantile Plot: Every Value Against Its Rank



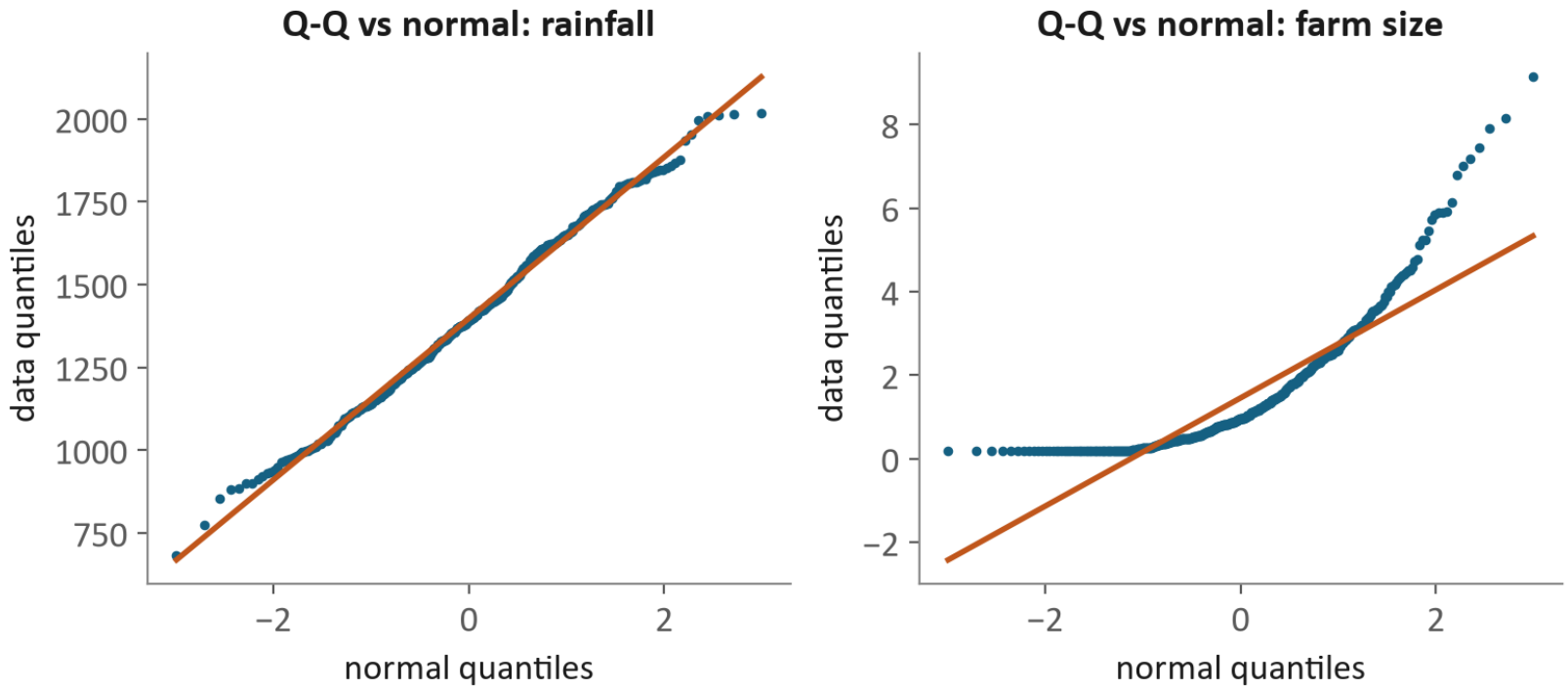
$$f_i = (i - 0.5) \div n$$

i = rank after sorting; n = 500

f	yield
0.25	1.45 t (Q1)
0.50	2.19 t (median)
0.75	3.40 t (Q3)

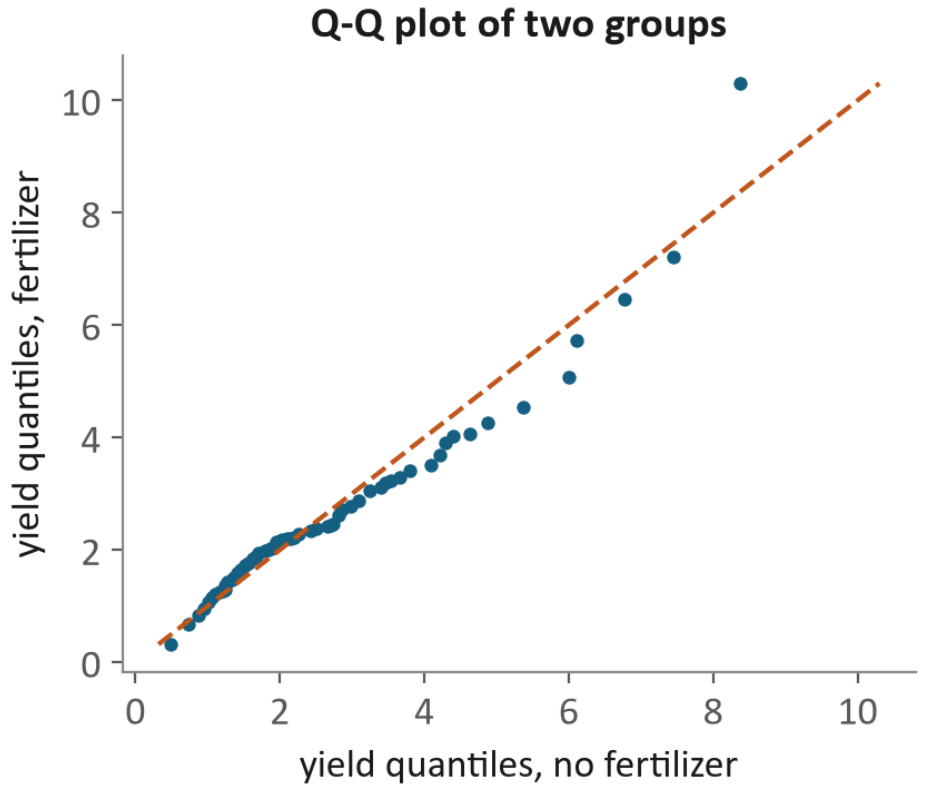
Read: a steep climb at the right = a few very high yields (mostly cassava). Percentiles can be read directly.

Q-Q Plot: Does the Data Look Normal?



Reading: points on the line → normal-like (rainfall). Points bending up at the right → right skew (farm size).

Q-Q Plot of Two Groups: Fertilizer or Not



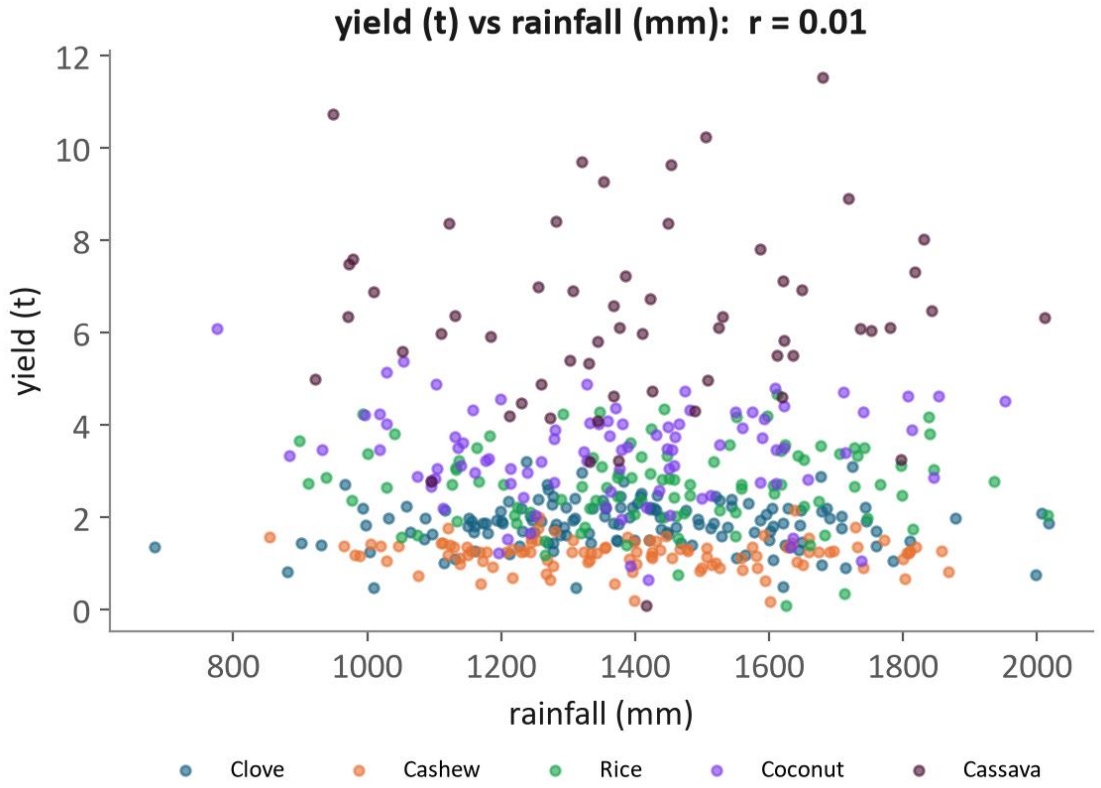
1. **Axes**
   quantiles of group A vs group B

2. **On y = x**
   the two groups are distributed alike

3. **Our farms**
   medians 2.13 t (no) vs 2.21 t (yes)

4. **Next step**
   investigate: which crop? which fertilizer?

Scatter Plot: Look Before You Compute



1 **Trend?**
none: r ≈ 0

2 **Groups?**
horizontal bands by crop

3 **Checklist**
trend · straight or curved · outliers · groups

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

Recap and the Question: Do Two Attributes Move Together?

Covariance from Zero

Pearson Correlation r

Correlation Is Not Causation

Graphic Displays: Seeing the Data

→ **Correlation at Scale**

In-Class Exercise

Correlation in One Pass

$$\mathrm{Cov}(x, y) = \Sigma xy \div n - \bar{x}\,\bar{y}$$

needs only n, Σx, Σy, Σxy (and Σx^2, Σy^2 for the σ's)

Student	x	y	x × y
1	1	40	40
2	3	50	150
3	5	70	350
4	7	60	420
5	9	80	720
		Σxy	**1,680**

1 **Check**
1,680 ÷ 5 − 5 × 60 = 336 − 300 = 36 ✓

2 **At scale**
6 running sums; sums from many machines add

Matrix of d attributes: about $d^2/2$ sums — 1,000 attributes → ~500,000 numbers, still cheap.

Covariance, Correlation & Graphic Displays of Statistics

IIT Madras Zanzibar

Recap and the Question: Do Two Attributes Move Together?

Covariance from Zero

Pearson Correlation r

Correlation Is Not Causation

Graphic Displays: Seeing the Data

Correlation at Scale

→ **In-Class Exercise**

In-Class Exercise (6 min, pairs)

IIT Madras Zanzibar

IN-CLASS TASK

Four Stone Town shops: weekly airtime sales x (hundred TZS) and mobile-money deposits y (thousand TZS): (2, 10), (4, 14), (6, 12), (8, 20)

1. Compute $\bar{x}$, $\bar{y}$ and the covariance (divide by n).
2. Compute σ_x, σ_y and Pearson's r. Describe the strength and direction.
3. Check the covariance using $\Sigma xy \div n - \bar{x}\bar{y}$.
4. A manager says: 'selling more airtime causes more deposits'. Name a possible confounder.

Solution

IIT Madras Zanzibar

1 **Means and Cov**

$\bar{x} = 5$, $\bar{y} = 14$; products $(-3)(-4) + (-1)(0) + (1)(-2) + (3)(6) = 12 + 0 - 2 + 18 = 28$ → Cov = $28 \div 4 = 7$

2 **r**

$\sigma_x = \sqrt{(20 \div 4)} = 2.24$; $\sigma_y = \sqrt{(56 \div 4)} = 3.74$; $r = 7 \div (2.24 \times 3.74) = 0.84$ → strong positive

3 **Check**

$\Sigma xy = 20 + 56 + 72 + 160 = 308$; $308 \div 4 - 5 \times 14 = 77 - 70 = 7$ ✓

4 **Confounder**

customer footfall (shop location): busy shops sell more of everything

Summary & Exam Checkpoints

IIT Madras Zanzibar

Covariance

average of $(x - \bar{x})(y - \bar{y})$: sign = direction, but size depends on units

Pearson's r

Cov ÷ $(\sigma_x \sigma_y)$, between −1 and +1; measures straight-line association only

Not causation

look for confounders (crop drives price and yield), reverse cause, chance

Plot first

histogram, boxplot, quantile, Q-Q, scatter; r in one pass from 6 running sums

What's Next

IIT Madras Zanzibar

Thu **Lecture 5 (24 Sep): Similarity & Distance**
Euclidean, Manhattan, Minkowski, Jaccard, cosine, mixed types

Tut **Tutorial 2 (Thu 4:45 pm)**
correlation, plots, distance and similarity on the 500 farms

Fri **Lecture 6 (25 Sep): Data Quality & Cleaning**
missing values, outliers, integration, χ^2 for redundancy

Practise: df.corr(numeric_only=True) and pd.plotting.scatter_matrix(df) on the farm file.

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed., Sections 2.2.2–2.2.3 and 3.3.2. Morgan Kaufmann.
- Pearson, K. (1895). Notes on Regression and Inheritance in the Case of Two Parents. Proc. Royal Society of London 58.
- Anscombe, F. J. (1973). Graphs in Statistical Analysis. The American Statistician 27(1).
- Wilk, M. B. & Gnanadesikan, R. (1968). Probability Plotting Methods for the Analysis of Data. Biometrika 55(1).
- Leskovec, J., Rajaraman, A. & Ullman, J. (2020). Mining of Massive Datasets, 3rd ed. mmds.org.