

# Statistics and Data Analysis

## Unit 02 – Lecture 03: Correlation, Skewness, Kurtosis

Tofik Ali

School of Computer Science, UPES Dehradun

February 14, 2026

<https://github.com/tali7c/Statistics-and-Data-Analysis>

Overview  
oo

Correlation  
oooooooooooo

Skewness  
ooooooo

Kurtosis  
ooooooo

Demo  
o

Summary  
o

# Quick Links

Correlation

Skewness

Kurtosis

Demo

Summary

# Agenda

1 Overview

2 Correlation

3 Skewness

4 Kurtosis

5 Demo

6 Summary

## Learning Outcomes

- Explain correlation and compute Pearson correlation  $r$

## Learning Outcomes

- Explain correlation and compute Pearson correlation  $r$
  - Relate covariance and correlation, and explain why correlation is scale-free

## Learning Outcomes

- Explain correlation and compute Pearson correlation  $r$
  - Relate covariance and correlation, and explain why correlation is scale-free
  - Identify common pitfalls: outliers, non-linearity, and correlation vs causation

## Learning Outcomes

- Explain correlation and compute Pearson correlation  $r$
  - Relate covariance and correlation, and explain why correlation is scale-free
  - Identify common pitfalls: outliers, non-linearity, and correlation vs causation
  - Interpret skewness (right/left skew) and kurtosis (tail heaviness)

# From Covariance to Correlation (Recap)

- Covariance tells direction of joint variation, but it depends on units

## From Covariance to Correlation (Recap)

- Covariance tells direction of joint variation, but it depends on units
- Correlation standardizes covariance to a unitless number in  $[-1, 1]$

# From Covariance to Correlation (Recap)

- Covariance tells direction of joint variation, but it depends on units
- Correlation standardizes covariance to a unitless number in  $[-1, 1]$
- That makes correlation easier to compare across different datasets

# What is Correlation?

Correlation measures **linear association** between two variables.

- $r > 0$ : as  $x$  increases,  $y$  tends to increase

# What is Correlation?

Correlation measures **linear association** between two variables.

- $r > 0$ : as  $x$  increases,  $y$  tends to increase
- $r < 0$ : as  $x$  increases,  $y$  tends to decrease

# What is Correlation?

Correlation measures **linear association** between two variables.

- $r > 0$ : as  $x$  increases,  $y$  tends to increase
- $r < 0$ : as  $x$  increases,  $y$  tends to decrease
- $r \approx 0$ : no strong *linear* pattern (could still be non-linear)

# Pearson Correlation (Formula)

For paired data  $(x_i, y_i)$ :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Always between  $-1$  and  $1$

# Pearson Correlation (Formula)

For paired data  $(x_i, y_i)$ :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Always between  $-1$  and  $1$
- Unitless (no units)

# Pearson Correlation (Formula)

For paired data  $(x_i, y_i)$ :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Always between  $-1$  and  $1$
- Unitless (no units)
- Sensitive to outliers

# Correlation vs Covariance

$$r = \frac{s_{xy}}{s_x s_y}$$

- $s_{xy}$ : sample covariance (Lecture 02)

# Correlation vs Covariance

$$r = \frac{s_{xy}}{s_x s_y}$$

- $s_{xy}$ : sample covariance (Lecture 02)
- $s_x, s_y$ : sample standard deviations

# Correlation vs Covariance

$$r = \frac{s_{xy}}{s_x s_y}$$

- $s_{xy}$ : sample covariance (Lecture 02)
- $s_x, s_y$ : sample standard deviations
- Scaling a variable (changing units) does **not** change  $r$

# Interpreting $r$ (Rule of Thumb)

$ r $ range	Common description
0.00–0.19	very weak
0.20–0.39	weak
0.40–0.59	moderate
0.60–0.79	strong
0.80–1.00	very strong

*Always confirm with a scatter plot.*

## Exercise 1: Pearson Correlation (Positive)

Hours studied vs Score:

Hours (x)	1	2	3	4	5
Score (y)	52	55	60	65	68

Given:  $\bar{x} = 3$ ,  $\bar{y} = 60$ ,  $\sum(x - \bar{x})(y - \bar{y}) = 42$ ,  $\sum(x - \bar{x})^2 = 10$ ,  $\sum(y - \bar{y})^2 = 178$ .

**Task:** Compute  $r$  and interpret it.

# Solution 1

$$r = \frac{42}{\sqrt{10}\sqrt{178}} = \frac{42}{\sqrt{1780}} \approx 0.9955$$

**Interpretation:** Very strong positive linear association between hours and score.

## Exercise 2: Pearson Correlation (Negative)

Price vs Demand:

Price (x)	1	2	3	4	5
Demand (y)	80	70	60	50	40

**Task:** Compute  $r$ . What does the sign mean?

## Solution 2

Here  $y = 90 - 10x$  is a perfect decreasing line.

$$r = -1$$

**Interpretation:** Perfect negative linear relationship.

## Exercise 3: $r = 0$ Does Not Mean “No Relationship”

Consider:

$$x = [-2, -1, 0, 1, 2], \quad y = x^2 = [4, 1, 0, 1, 4]$$

**Task:** Compute  $r$ . Is there a relationship between  $x$  and  $y$ ?

# Solution 3

$\bar{x} = 0$ ,  $\bar{y} = 2$ . The numerator  $\sum(x - \bar{x})(y - \bar{y}) = 0$ , so:

$$r = 0$$

**Key point:**  $r = 0$  means no *linear* association; here the relationship is strong but non-linear.

# Correlation $\neq$ Causation

- Correlation only says “they move together” (linearly)

# Correlation $\neq$ Causation

- Correlation only says “they move together” (linearly)
- A third variable can cause both (confounding)

# Correlation $\neq$ Causation

- Correlation only says “they move together” (linearly)
- A third variable can cause both (confounding)
- Sometimes correlation is accidental (spurious)

# Correlation $\neq$ Causation

- Correlation only says “they move together” (linearly)
- A third variable can cause both (confounding)
- Sometimes correlation is accidental (spurious)
- Use domain knowledge + experiments/causal reasoning to claim causation

## Exercise 4: Interpret a Correlation Claim

"Ice cream sales and drowning incidents are positively correlated."

Which statement is most correct?

- 1** Ice cream causes drowning.
- 2** Drowning causes ice cream sales.
- 3** Both may increase due to a third factor (e.g., temperature/season).

# Solution 4

**Correct: (3).** A confounder like hot weather can increase both swimming (risk) and ice cream sales.

# Skewness (Distribution Asymmetry)

Skewness describes the **direction of the tail**.

- **Right-skewed (positive):** long tail to the right (few very large values)

# Skewness (Distribution Asymmetry)

Skewness describes the **direction of the tail**.

- **Right-skewed (positive):** long tail to the right (few very large values)
- **Left-skewed (negative):** long tail to the left (few very small values)

# Skewness (Distribution Asymmetry)

Skewness describes the **direction of the tail**.

- **Right-skewed (positive):** long tail to the right (few very large values)
- **Left-skewed (negative):** long tail to the left (few very small values)
- Symmetric: tails are similar on both sides

# Mean vs Median vs Mode (Heuristic)

- Right-skewed: mean > median > mode

*Reason: the mean is pulled toward the long tail.*

# Mean vs Median vs Mode (Heuristic)

- Right-skewed: mean > median > mode
- Left-skewed: mean < median < mode

*Reason: the mean is pulled toward the long tail.*

# Mean vs Median vs Mode (Heuristic)

- Right-skewed: mean > median > mode
- Left-skewed: mean < median < mode
- Symmetric: mean  $\approx$  median  $\approx$  mode

*Reason: the mean is pulled toward the long tail.*

# Moment Skewness (One Common Formula)

Let  $m_k$  be the  $k$ th central moment (divide by  $n$ ):

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Moment skewness:

$$g_1 = \frac{m_3}{m_2^{3/2}}$$

- $g_1 > 0$  right-skewed,  $g_1 < 0$  left-skewed

# Moment Skewness (One Common Formula)

Let  $m_k$  be the  $k$ th central moment (divide by  $n$ ):

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Moment skewness:

$$g_1 = \frac{m_3}{m_2^{3/2}}$$

- $g_1 > 0$  right-skewed,  $g_1 < 0$  left-skewed
- Different software may use small-sample corrections

## Exercise 5: Identify Skewness Direction

Dataset A (Income, INR thousands):

20    22    23    24    25    26    27    28    60

Dataset B (Scores):

50    80    85    88    90    92    93    94    95    96

**Task:** For each dataset, decide if it is right-skewed or left-skewed.  
Predict whether  $\text{mean} > \text{median}$  or  $\text{mean} < \text{median}$ .

# Solution 5

- Dataset A: one large value (60) creates a right tail  $\Rightarrow$  right-skewed,  $\text{mean} > \text{median}$
- Dataset B: one small value (50) creates a left tail  $\Rightarrow$  left-skewed,  $\text{mean} < \text{median}$

## Exercise 6: Compute Moment Skewness

For Dataset A (income), suppose:

$$\bar{x} = 28.33, \quad m_2 = 130.89, \quad m_3 = 3404.07$$

**Task:** Compute  $g_1 = \frac{m_3}{m_2^{3/2}}$  and interpret the sign.

# Solution 6

$$g_1 = \frac{3404.07}{(130.89)^{3/2}} \approx 2.27$$

**Interpretation:** Positive and large  $\Rightarrow$  strongly right-skewed distribution.

# Kurtosis (Tail Heaviness)

Kurtosis summarizes how heavy the tails are (and how often extreme values appear).

- Often reported as **excess kurtosis** = kurtosis – 3

# Kurtosis (Tail Heaviness)

Kurtosis summarizes how heavy the tails are (and how often extreme values appear).

- Often reported as **excess kurtosis** = kurtosis - 3
- Normal distribution has excess kurtosis 0

# Kurtosis (Tail Heaviness)

Kurtosis summarizes how heavy the tails are (and how often extreme values appear).

- Often reported as **excess kurtosis** = kurtosis – 3
- Normal distribution has excess kurtosis 0
- Positive excess: heavier tails; negative excess: lighter tails

# Moment Kurtosis (One Common Formula)

Moment kurtosis:

$$g_2 = \frac{m_4}{m_2^2}$$

Excess kurtosis:

$$\text{Excess} = g_2 - 3$$

- If excess  $> 0$ , more extreme values than normal (heavy tails)

# Moment Kurtosis (One Common Formula)

Moment kurtosis:

$$g_2 = \frac{m_4}{m_2^2}$$

Excess kurtosis:

$$\text{Excess} = g_2 - 3$$

- If excess  $> 0$ , more extreme values than normal (heavy tails)
- If excess  $< 0$ , fewer extremes than normal (light tails)

## Exercise 7: Excess Kurtosis (Small Symmetric Data)

Dataset: 1, 2, 3, 4, 5.

Mean = 3, deviations:  $[-2, -1, 0, 1, 2]$ .

**Task:** Compute  $m_2$ ,  $m_4$ , then  $g_2$  and excess kurtosis.

# Solution 7

$$m_2 = \frac{4 + 1 + 0 + 1 + 4}{5} = 2$$

$$m_4 = \frac{16 + 1 + 0 + 1 + 16}{5} = \frac{34}{5} = 6.8$$

$$g_2 = \frac{6.8}{2^2} = 1.7, \quad \text{excess} = 1.7 - 3 = -1.3$$

**Interpretation:** Negative excess  $\Rightarrow$  lighter tails than normal (platykurtic).

## Exercise 8: Excess Kurtosis (Income Example)

For the income dataset (Exercise 6), suppose:

$$m_2 = 130.89, \quad m_4 = 112590.30$$

**Task:** Compute  $g_2 = \frac{m_4}{m_2^2}$  and excess kurtosis. Interpret.

# Solution 8

$$g_2 \approx \frac{112590.30}{(130.89)^2} \approx 6.57, \quad \text{excess} \approx 3.57$$

**Interpretation:** Large positive excess  $\Rightarrow$  heavy tails / extreme values (outliers).

# Common Pitfalls (Skewness & Kurtosis)

- Small samples can give unstable skewness/kurtosis values

# Common Pitfalls (Skewness & Kurtosis)

- Small samples can give unstable skewness/kurtosis values
- Different formulas exist (bias corrections), so values may differ across tools

# Common Pitfalls (Skewness & Kurtosis)

- Small samples can give unstable skewness/kurtosis values
- Different formulas exist (bias corrections), so values may differ across tools
- Always verify with plots (histogram/boxplot) and context

# Mini Demo (Python)

Run:

```
python demo/correlation_skew_kurt_demo.py
```

What it does:

- Computes Pearson correlation for three paired datasets
- Prints correlation matrix for data/student\_metrics.csv
- Computes moment skewness and excess kurtosis for example distributions
- (Optional) Saves plots to images/ if matplotlib is installed

# Summary

- Correlation standardizes covariance to  $[-1, 1]$  and measures linear association

**Exit question:** Give one real-life example where correlation might be misleading and explain why.

# Summary

- Correlation standardizes covariance to  $[-1, 1]$  and measures linear association
- $r = 0$  does not mean independence; it only indicates no linear relation

**Exit question:** Give one real-life example where correlation might be misleading and explain why.

# Summary

- Correlation standardizes covariance to  $[-1, 1]$  and measures linear association
- $r = 0$  does not mean independence; it only indicates no linear relation
- Skewness describes tail direction; mean is pulled toward the tail

**Exit question:** Give one real-life example where correlation might be misleading and explain why.

# Summary

- Correlation standardizes covariance to  $[-1, 1]$  and measures linear association
- $r = 0$  does not mean independence; it only indicates no linear relation
- Skewness describes tail direction; mean is pulled toward the tail
- Excess kurtosis relates to tail heaviness and outliers

**Exit question:** Give one real-life example where correlation might be misleading and explain why.