Statistics and Data Analysis
Unit 02 – Lecture 03 Notes
Correlation, Skewness, Kurtosis

Tofik Ali

February 17, 2026

What You Will Learn (Beginner-Friendly)

This lecture answers two practical questions:

1. If two variables (like study hours and score) move together, how do we measure the *strength* of that relationship?
2. If we look at one variable (like income), how do we describe the *shape* of its distribution beyond center and spread?

By the end, you should be able to:

- Define **Pearson correlation** and compute it on small paired datasets.
- Explain why correlation is **scale-free** and how it relates to covariance.
- Explain key cautions: **outliers**, **non-linearity**, and **correlation vs causation**.
- Interpret **skewness** (right/left skew) and **kurtosis** (tail heaviness).
- Compute simple **moment skewness** and **excess kurtosis** for small datasets.

1. Correlation

1.1 Intuition

Correlation measures **linear association** between two variables. If a scatter plot looks like an upward sloping line, correlation is positive. If it looks like a downward sloping line, correlation is negative.

1.2 Pearson correlation formula

For paired observations (x_i, y_i), the Pearson correlation is:

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

Key facts:

- $-1 \le r \le 1$.
- The sign (+ or −) shows direction (increase together vs opposite).
- The magnitude $|r|$ shows strength of *linear* relationship.
- Correlation is **unitless** (scale-free).

1.3 Correlation vs covariance

In Lecture 02 we computed **sample covariance**:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Correlation is the standardized version:

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

where s_x and s_y are sample standard deviations.

Why "scale-free" matters. If we change units (hours to minutes, marks to percentage), covariance changes because its units change. But correlation does *not* change, because the scaling affects the numerator and denominator in the same way.

1.4 Common pitfalls

- **Outliers:** a single extreme point can strongly change r.
- **Non-linearity:** you can have a strong relationship but $r \approx 0$ if the pattern is curved.
- **Correlation vs causation:** even a high correlation does not prove that x causes y.

2. Exercises (Correlation)

Exercise 1: Pearson correlation (positive)

Hours studied vs Score:

$$x = [1, 2, 3, 4, 5], \quad y = [52, 55, 60, 65, 68]$$

Given: $\bar{x} = 3$, $\bar{y} = 60$.

Step 1: Compute deviations and sums.

$$x - \bar{x} = [-2, -1, 0, 1, 2]$$

$$y - \bar{y} = [-8, -5, 0, 5, 8]$$

Now compute:

$$\sum (x - \bar{x})(y - \bar{y}) = (16 + 5 + 0 + 5 + 16) = 42$$

$$\sum (x - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

$$\sum (y - \bar{y})^2 = 64 + 25 + 0 + 25 + 64 = 178$$

Step 2: Plug into Pearson formula.

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

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

Exercise 2: Pearson correlation (negative)

Price vs Demand:

$$x = [1, 2, 3, 4, 5], \quad y = [80, 70, 60, 50, 40]$$

Here the relationship is perfectly linear: $y = 90 - 10x$. So:

$$r = -1$$

Interpretation: Perfect negative linear relationship.

Exercise 3: $r = 0$ but strong relationship

Let:

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

Compute means:

$$\bar{x} = 0, \quad \bar{y} = \frac{4 + 1 + 0 + 1 + 4}{5} = 2$$

Compute numerator:

$$\sum (x - \bar{x})(y - \bar{y}) = (-2)(2) + (-1)(-1) + 0(-2) + 1(-1) + 2(2) = 0$$

So:

$$r = 0$$

Interpretation: No *linear* association, but a strong *non-linear* relationship exists (y is determined by x).

Exercise 4: Correlation vs causation

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

Best explanation: A third variable like temperature/season can increase both ice cream sales and swimming activity. This is called **confounding**. Correlation alone cannot prove causation.

3. Skewness

3.1 What skewness means

Skewness describes **asymmetry** in a distribution:

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

3.2 Mean vs median (important intuition)

The mean is pulled toward extreme values more than the median. So:

- Right-skewed: mean > median (high outliers pull mean up).
- Left-skewed: mean < median (low outliers pull mean down).

3.3 Moment skewness (one common definition)

Define central moments (divide by n):

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

Moment skewness:

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

Notes:

- $g_1 > 0$ indicates right skew; $g_1 < 0$ indicates left skew.
- Some software uses bias-corrected formulas; values can differ slightly.

4. Exercises (Skewness)

Exercise 5: Identify skewness direction using mean/median

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.

Dataset A. Median is 25 (middle value). Mean is:

$$\bar{x} = \frac{20+22+23+24+25+26+27+28+60}{9} = \frac{255}{9} \approx 28.33$$

Since mean > median, the data is right-skewed (positive skew).

Dataset B. Median is $(90+92)/2 = 91$. Mean is:

$$\bar{x} = \frac{50+80+85+88+90+92+93+94+95+96}{10} = \frac{863}{10} = 86.3$$

Since mean < median, the data is left-skewed (negative skew).

Exercise 6: Compute moment skewness (income)

Suppose for Dataset A (income) we have:

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

Then:

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

Interpretation: strongly right-skewed distribution.

5. Kurtosis

5.1 What kurtosis means

Kurtosis is commonly used to describe **tail heaviness** (how often extreme values occur). It is often reported as **excess kurtosis**:

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

The normal distribution has excess kurtosis 0.

5.2 Moment kurtosis (one common definition)

Moment kurtosis:

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

Excess kurtosis:

$$g_2 - 3$$

Interpretation:

- Excess > 0: heavier tails (more extreme values than normal).
- Excess < 0: lighter tails (fewer extremes than normal).

6. Exercises (Kurtosis)

Exercise 7: Excess kurtosis for 1,2,3,4,5

Dataset: 1, 2, 3, 4, 5. Mean is 3. Deviations: $[-2, -1, 0, 1, 2]$.

Step 1: Compute m_2.

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

Step 2: Compute m_4.

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

Step 3: Compute kurtosis and excess.

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

Interpretation: negative excess kurtosis (lighter tails than normal).

Exercise 8: Excess kurtosis for the income example

Suppose for the income dataset:

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

Then:

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

Interpretation: large positive excess kurtosis indicates heavy tails / extreme values (outliers).

7. Mini Demo (Python)

Run this from the lecture folder:

`python demo/correlation_skew_kurt_demo.py`

The script:

- computes Pearson correlation for:
  - hours vs score (positive)
  - price vs demand (negative)
  - x vs x^2 (non-linear example)
- prints correlation values among features in `data/student_metrics.csv`
- computes moment skewness and excess kurtosis for example univariate datasets
- optionally saves plots into `images/` if `matplotlib` is installed

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley, 7th ed., 2020.
- Gupta, S. C., & Kapoor, V. K. *Fundamentals of Applied Statistics*, Sultan Chand & Sons, 4th rev. ed., 2007.
- McKinney, W. *Python for Data Analysis*, O'Reilly, 2022.

Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

Title Slide

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

Quick Links

Correlation Skewness Kurtosis Demo Summary

Agenda

- Overview
- Correlation
- Skewness
- Kurtosis
- Demo
- Summary

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
- 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
- $r < 0$: 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
- Unitless (no units)
- Sensitive to outliers

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)

$\lvert r \rvert$ 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)
- 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)
- **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
- 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 kth 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 mean $>$ median or mean $<$ median.

Solution 5

- Dataset A: one large value (60) creates a right tail $\Rightarrow$ right-skewed, mean $>$ median
- Dataset B: one small value (50) creates a left tail $\Rightarrow$ left-skewed, mean $<$ 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 $g_1 = \dfrac{m_3}{m_2^{3/2}}$ 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
- 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)
- 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 $g_2 = \dfrac{m_4}{m_2^2}$ 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
- 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
- $r = 0$ 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.