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

Tofik Ali

February 9, 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.