Statistics and Data Analysis
Unit 02 – Lecture 02 Notes
Dispersion and Covariance

Tofik Ali

February 4, 2026

What You Will Learn (Beginner-Friendly)

In this lecture we answer two practical questions:

1. If two datasets have the same "average," how do we tell which one is more spread out?
2. If we have two variables (like hours studied and score), how do we measure whether they move together?

By the end, you should be able to:

- Explain why measures of **dispersion** (spread) are needed in addition to mean/median/mode.
- Compute **range** and **IQR** (interquartile range).
- Compute **sample variance** and **sample standard deviation**.
- Compute **coefficient of variation (CV)** and a simple **z-score**.
- Use the **IQR rule** to flag potential outliers.
- Define **covariance**, compute it for paired data, and interpret its sign.

1. Warm-up: Same Mean, Different Spread

Consider:

- Dataset A: 10, 15, 20
- Dataset B: 14, 15, 16

Both have mean 15, but Dataset A is much more spread out. This shows why a single "center" value (mean/median/mode) is not enough. We need a second type of summary: **dispersion**.

2. Dispersion (Spread) Measures

Dispersion describes how far data values are from the center.

2.1 Range

Range is the simplest spread measure:

$$\text{Range} = \max(x) - \min(x)$$

It is easy to compute, but it uses only two values (the minimum and maximum), so one extreme outlier can make the range very large.

2.2 Quartiles and IQR

To define IQR, we first sort the data. Quartiles are values that split the sorted data into four parts:

- Q_1: first quartile (25% point)
- Q_2: second quartile (the median, 50% point)
- Q_3: third quartile (75% point)

IQR (Interquartile Range) measures spread of the middle 50%:

$$\text{IQR} = Q_3 - Q_1$$

IQR is more robust than range because it ignores extreme tails.

How we compute Q_1 and Q_3 in this course (median-of-halves method). If n is even, split the sorted data into two halves of size $n/2$. Compute the median of the lower half as Q_1 and the median of the upper half as Q_3. If n is odd, exclude the middle value before splitting.

2.3 Variance and Standard Deviation

Range and IQR are intuitive, but many statistical methods use variance and standard deviation.

Mean (reminder). For values $x_1, \ldots, x_n$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample variance.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Key idea: we measure each deviation from the mean, square it (to remove negative signs), and average the squared deviations.

Why divide by $n-1$? Because we are estimating the population variance from a sample. After using the data to compute $\bar{x}$, the deviations are not all free to vary independently. The $n-1$ correction reduces bias in the variance estimate (this is a standard result in statistics).

Standard deviation.

$$s = \sqrt{s^2}$$

Standard deviation has the same unit as the original data (unlike variance, which has squared units).

2.4 Coefficient of Variation (CV)

Sometimes we want to compare "how variable" two datasets are, but their units or scales are different. For example, Dataset A might be salaries in rupees and Dataset B might be prices in rupees, but the means are very different. In such cases, the standard deviation alone can be misleading because it is measured in the original units.

Coefficient of variation (CV) measures spread *relative to the mean*:

$$\text{CV} = \frac{s}{\bar{x}} \times 100\%$$

It is a percentage (unitless), so it is useful for comparison across datasets. A larger CV means more variability relative to the mean.

Important note: CV is most meaningful when the mean is positive and not close to zero. If $\bar{x}$ is near 0, the CV can become extremely large and hard to interpret.

2.5 Standardization (z-score)

A **z-score** tells how far a value is from the mean, measured in standard deviations:

$$z = \frac{x - \bar{x}}{s}$$

How to interpret:

- $z = 0$ means the value equals the mean.
- $z = 1$ means "1 standard deviation above the mean."
- $z = -2$ means "2 standard deviations below the mean."

Z-scores are useful because they put values on a common scale, which helps comparison.

2.6 Outlier detection using the IQR rule (fences)

In practice, we often want a quick way to *flag* potential outliers. A common rule-of-thumb uses **IQR fences**:

$$\text{Lower fence} = Q_1 - 1.5 \times \text{IQR}, \quad \text{Upper fence} = Q_3 + 1.5 \times \text{IQR}$$

Any value smaller than the lower fence or larger than the upper fence is marked as a possible outlier. This is not a "proof" that a point is wrong; it simply signals that we should investigate.

3. Exercises (Dispersion)

Exercise 1: Range and IQR

Dataset (Scores): 11, 13, 15, 15, 17, 19

Step 1: Sort. It is already sorted: 11, 13, 15, 15, 17, 19.

Step 2: Range.

$$\text{Range} = 19 - 11 = 8$$

Step 3: Quartiles and IQR. Since $n = 6$ (even), split into halves:

- Lower half: 11, 13, 15 $\Rightarrow Q_1 = 13$
- Upper half: 15, 17, 19 $\Rightarrow Q_3 = 17$

So:

$$\text{IQR} = 17 - 13 = 4$$

Exercise 2: Sample variance and standard deviation

Use the same dataset and $\bar{x} = 15$.

Step 1: Compute deviations and squares.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
11	-4	16
13	-2	4
15	0	0
15	0	0
17	2	4
19	4	16

Sum of squares:

$$\sum (x_i - \bar{x})^2 = 16 + 4 + 0 + 0 + 4 + 16 = 40$$

Step 2: Sample variance. Here $n = 6$:

$$s^2 = \frac{40}{6 - 1} = \frac{40}{5} = 8$$

Step 3: Sample standard deviation.

$$s = \sqrt{8} \approx 2.83$$

Interpretation: A typical score is about 2.8 points away from the mean of 15.

Exercise 3: Coefficient of variation (CV)

Using $\bar{x} = 15$ and $s \approx 2.83$ for the scores dataset:

$$\text{CV} = \frac{s}{\bar{x}} \times 100\% = \frac{2.83}{15} \times 100\% \approx 18.9\%$$

Interpretation: The standard deviation is about 19% of the mean. If another dataset had CV of 40%, it would be *more variable relative to its mean.*

Exercise 4: z-score

Compute the z-score for $x = 19$ using $\bar{x} = 15$ and $s \approx 2.83$:

$$z = \frac{19 - 15}{2.83} \approx 1.41$$

Interpretation: 19 is about 1.4 standard deviations above the mean.

Exercise 5: IQR outlier check (income example)

Monthly income (INR thousands): 20, 22, 23, 24, 25, 26, 27, 28, 60

Step 1: Sort. Already sorted. Here $n = 9$ so the median is the 5th value: 25.

Step 2: Find Q_1 and Q_3 (median-of-halves). Exclude the median (25) and split:

- Lower half: 20, 22, 23, 24 $\Rightarrow Q_1 = (22 + 23)/2 = 22.5$
- Upper half: 26, 27, 28, 60 $\Rightarrow Q_3 = (27 + 28)/2 = 27.5$

So IQR $= 27.5 - 22.5 = 5$.

Step 3: Compute fences.

$$\text{Lower fence} = 22.5 - 1.5(5) = 15, \quad \text{Upper fence} = 27.5 + 1.5(5) = 35$$

Since $60 > 35$, the value 60 is an outlier by the IQR rule.

4. Covariance

Now we move from one-variable spread to two-variable joint behavior.

4.1 What covariance measures

Covariance measures whether two variables tend to move together:

- Positive covariance: when x is above its mean, y tends to be above its mean.
- Negative covariance: when x is above its mean, y tends to be below its mean.
- Near zero covariance: no *linear* co-variation (but a non-linear pattern can still exist).

4.2 Sample covariance formula

For paired data (x_i, y_i):

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

Units and scale dependence. If x is in hours and y is in marks, then covariance unit is "hours $\times$ marks". If you change units (e.g., marks to percentage), covariance changes. This is why covariance is hard to compare across different scales. Next lecture, **correlation** normalizes covariance to $[-1, 1]$.

4.3 A useful scaling property

Covariance has a simple scaling behavior:

$$\text{cov}(aX, Y) = a \, \text{cov}(X, Y)$$

If you multiply a variable by a constant (changing units), the covariance multiplies by the same constant. This is one reason we often prefer correlation for comparing strength across different unit scales.

5. Exercises (Covariance)

Exercise 6: Hours studied vs score (positive covariance)

Data:

- x (hours): 1, 2, 3, 4, 5 $\Rightarrow \bar{x} = 3$
- y (score): 52, 55, 60, 65, 68 $\Rightarrow \bar{y} = 60$

Step 1: Compute deviations.

$$x - \bar{x} = [-2, -1, 0, 1, 2]$$
$$y - \bar{y} = [-8, -5, 0, 5, 8]$$

Step 2: Multiply deviations and sum.

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

Step 3: Divide by $n - 1$. Here $n = 5$:

$$s_{xy} = \frac{42}{5 - 1} = 10.5$$

Interpretation: Positive covariance means higher hours are associated with higher scores.

Exercise 7: Price vs demand (negative covariance)

Data:

- x (price): 1, 2, 3, 4, 5 $\Rightarrow \bar{x} = 3$
- y (demand): 80, 70, 60, 50, 40 $\Rightarrow \bar{y} = 60$

Deviations:

$$x - \bar{x} = [-2, -1, 0, 1, 2]$$
$$y - \bar{y} = [20, 10, 0, -10, -20]$$

Products sum:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = -40 - 10 + 0 - 10 - 40 = -100$$

Sample covariance:

$$s_{xy} = \frac{-100}{5 - 1} = -25$$

Interpretation: Negative covariance means higher price is associated with lower demand.

Exercise 8: Unit change and covariance (no full re-calculation)

From Exercise 6 we found:

$$\text{cov}(\text{hours}, \text{score}) = 10.5$$

If we change the unit of time from hours to minutes, then $x' = 60x$. Using the scaling property:

$$\text{cov}(x', y) = \text{cov}(60x, y) = 60\,\text{cov}(x, y) = 60(10.5) = 630$$

Key message: covariance changes when units change.

Exercise 9: Covariance 0 but strong relationship

Let:

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

Compute means:

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

Now compute the sum of products:

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

So the sample covariance is:

$$s_{xy} = \frac{0}{5 - 1} = 0$$

Even though covariance is 0, y is completely determined by x (because $y = x^2$). So the variables are **not independent**. This is why "covariance 0" only means "no linear association."

6. Mini Demo (Python)

Run this from the lecture folder:

```
python demo/dispersion_covariance_demo.py
```

The script:

- loads `data/scores_small.csv` and computes range, IQR, variance, SD
- loads `data/incomes_outlier.csv` (if present) and flags outliers using the IQR rule
- loads `data/pairs_hours_score.csv` and `data/pairs_price_demand.csv`
- computes sample covariance for the paired 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.