Statistics and Data Analysis

Unit 02 – Lecture 02: Dispersion and Covariance

Tofik Ali

School of Computer Science, UPES Dehradun

February 14, 2026

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

Quick Links

Agenda

Learning Outcomes

- Explain why dispersion is needed (beyond the mean/median/mode)

Learning Outcomes

- Explain why dispersion is needed (beyond the mean/median/mode)
- Compute and interpret range and IQR

Learning Outcomes

- Explain why dispersion is needed (beyond the mean/median/mode)
- Compute and interpret range and IQR
- Compute sample variance and standard deviation

Learning Outcomes

- Explain why dispersion is needed (beyond the mean/median/mode)
- Compute and interpret range and IQR
- Compute sample variance and standard deviation
- Compute coefficient of variation (CV) and simple z-scores

Learning Outcomes

- Explain why dispersion is needed (beyond the mean/median/mode)
- Compute and interpret range and IQR
- Compute sample variance and standard deviation
- Compute coefficient of variation (CV) and simple z-scores
- Use the IQR rule to flag potential outliers

Learning Outcomes

- Explain why dispersion is needed (beyond the mean/median/mode)
- Compute and interpret range and IQR
- Compute sample variance and standard deviation
- Compute coefficient of variation (CV) and simple z-scores
- Use the IQR rule to flag potential outliers
- Define covariance and interpret its sign and units

Warm-up: Same Mean, Different Spread

Two datasets can have the same mean but different variability:

Dataset A

10 15 20

Dataset B

14 15 16

Checkpoint: Which dataset is more variable? Why?

What is Dispersion?

Dispersion describes how spread out the data is around the center.

- **Range:** max – min

What is Dispersion?

Dispersion describes how spread out the data is around the center.

- **Range:** max – min
- **IQR:** spread of the middle 50% (Q3 – Q1)

What is Dispersion?

Dispersion describes how spread out the data is around the center.

- **Range:** max – min
- **IQR:** spread of the middle 50% (Q3 – Q1)
- **Variance/SD:** average squared deviation / typical deviation

Range and Interquartile Range (IQR)

- **Range** $= \max(x) - \min(x)$ (very sensitive to outliers)

Range and Interquartile Range (IQR)

- **Range** $= \max(x) - \min(x)$ (very sensitive to outliers)
- **Quartiles** split sorted data into quarters

Range and Interquartile Range (IQR)

- **Range** $= \max(x) - \min(x)$ (very sensitive to outliers)
- **Quartiles** split sorted data into quarters
- **IQR** = Q3 – Q1 (more robust than range)

Exercise 1: Range and IQR

Dataset (Scores):

$$11 \quad 13 \quad 15 \quad 15 \quad 17 \quad 19$$

Task: Compute Range, Q1, Q3, and IQR.

Solution 1

Sorted data: 11, 13, 15, 15, 17, 19

- Range $= 19 - 11 = 8$
- Lower half: 11, 13, 15 $\Rightarrow$ Q1 $= 13$
- Upper half: 15, 17, 19 $\Rightarrow$ Q3 $= 17$
- IQR $= 17 - 13 = 4$

Variance and Standard Deviation

- Variance measures average squared deviation from the mean

Variance and Standard Deviation

- Variance measures average squared deviation from the mean
- Standard deviation is the square root of variance

Variance and Standard Deviation

- Variance measures average squared deviation from the mean
- Standard deviation is the square root of variance
- **Units:** variance has squared units; SD has original units

Sample Variance (Why $n-1$?)

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

- Using $n-1$ helps correct bias when estimating population variance

Sample Variance (Why $n-1$?)

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

- Using $n-1$ helps correct bias when estimating population variance
- We lose one "degree of freedom" because $\bar{x}$ is estimated from the data

Exercise 2: Sample Variance and SD

Use the same dataset: 11, 13, 15, 15, 17, 19
Mean: $\bar{x} = 15$

- Compute s^2 and s
- Hint: $\sum(x_i - \bar{x})^2 = 40$

Solution 2

- $n = 6$
- $s^2 = \frac{40}{6-1} = \frac{40}{5} = 8$
- $s = \sqrt{8} \approx 2.83$

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

Coefficient of Variation (CV)

CV compares spread *relative to the mean*:

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

- Unitless (percentage) $\Rightarrow$ useful to compare variability across different scales

Coefficient of Variation (CV)

CV compares spread *relative to the mean*:

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

- Unitless (percentage) $\Rightarrow$ useful to compare variability across different scales
- Works best when the mean is positive and not near zero

Exercise 3: Coefficient of Variation

Use the same dataset: 11, 13, 15, 15, 17, 19
From Exercise 2: $\bar{x} = 15$, $s \approx 2.83$

Task: Compute CV (in %).

Solution 3

$$\mathrm{CV} = \frac{2.83}{15} \times 100\% \approx 18.9\%$$

Interpretation: The typical spread is about 19% of the mean.

Standardization (z-score)

A z-score tells how many standard deviations a value is from the mean:

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

- $z > 0$: value is above the mean; $z < 0$: below the mean

Standardization (z-score)

A z-score tells how many standard deviations a value is from the mean:

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

- $z > 0$: value is above the mean; $z < 0$: below the mean
- $|z|$ near 2 or 3 often indicates an unusually large/small value

Exercise 4: z-score

Using $\bar{x} = 15$ and $s \approx 2.83$, compute the z-score of $x = 19$.

Task: Compute z and interpret it in one sentence.

Solution 4

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

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

Outlier Detection (IQR Rule)

A common rule to flag potential outliers uses **fences**:

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

Values outside the fences are possible outliers.

Exercise 5: IQR Outlier Check

Monthly income (INR thousands):

20 22 23 24 25 26 27 28 60

Task: Compute Q_1, Q_3, IQR and decide if 60 is an outlier.

Solution 5

Median $= 25$ (since $n = 9$).

- Lower half: 20, 22, 23, 24 $\Rightarrow Q_1 = (22 + 23)/2 = 22.5$
- Upper half: 26, 27, 28, 60 $\Rightarrow Q_3 = (27 + 28)/2 = 27.5$
- IQR $= 27.5 - 22.5 = 5$
- Fences: $22.5 - 7.5 = 15$ and $27.5 + 7.5 = 35$

Conclusion: $60 > 35 \Rightarrow 60$ is an outlier.

Think-Pair-Share (2 minutes)

Prompt: Suppose you have a dataset with a few extreme outliers. Which spread measure would you report first: **IQR** or **SD**? Why?

What is Covariance?

Covariance measures how two variables vary together.

- Positive covariance: both tend to increase together

What is Covariance?

Covariance measures how two variables vary together.

- Positive covariance: both tend to increase together
- Negative covariance: one increases while the other decreases

What is Covariance?

Covariance measures how two variables vary together.

- Positive covariance: both tend to increase together
- Negative covariance: one increases while the other decreases
- Near zero: no *linear* co-variation (could still be non-linear)

Sample Covariance

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:** (units of x) $\times$ (units of y)

Sample Covariance

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:** (units of x) $\times$ (units of y)
- Covariance depends on scale (change units $\Rightarrow$ covariance changes)

Exercise 6: Covariance (Positive)

Dataset (Hours studied vs Score):

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

Task: Compute sample covariance s_{xy}. Interpret the sign.
(Means: $\bar{x} = 3$, $\bar{y} = 60$)

Solution 6

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

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 42$$

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

Interpretation: Positive covariance $\Rightarrow$ as hours increase, scores tend to increase.

Scale Dependence (Important)

- If we multiply y by 10 (change units), covariance multiplies by 10

Scale Dependence (Important)

- If we multiply y by 10 (change units), covariance multiplies by 10
- So covariance is hard to compare across different unit scales

Scale Dependence (Important)

- If we multiply y by 10 (change units), covariance multiplies by 10
- So covariance is hard to compare across different unit scales
- Next lecture: **correlation** standardizes covariance to $[-1, 1]$

Exercise 7: Covariance (Negative)

Dataset (Price vs Demand):

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

Task: Compute s_{xy} and interpret the sign.
(Means: $\bar{x} = 3$, $\bar{y} = 60$)

Solution 7

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

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = -100$$

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

Interpretation: Negative covariance $\Rightarrow$ higher price tends to reduce demand.

Exercise 8: Unit Change and Covariance

From Exercise 6, covariance (hours, score) is 10.5.
Suppose we measure time in minutes: $x' = 60x$.

Task: What is covariance of (x', y)? (No re-calculation needed.)

Solution 8

Property: $\mathrm{cov}(aX, Y) = a\,\mathrm{cov}(X, Y)$.

$$\mathrm{cov}(60X, Y) = 60 \times 10.5 = 630$$

Interpretation: Units changed $\Rightarrow$ covariance changed (scale-dependent).

Exercise 9: Covariance 0 but Strong Relationship

Consider:

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

Task: Compute sample covariance. Are x and y independent?

Solution 9

$\bar{x} = 0$, $\bar{y} = 2$.
Products $(x - \bar{x})(y - \bar{y})$: $-4, 1, 0, -1, 4 \Rightarrow \text{sum} = 0$.

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

Key point: Covariance $0 \neq$ independence (here y is determined by x).

Mini Demo (Python)

Run:

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

What it does:

- Computes range, IQR, variance, SD for `data/scores_small.csv`
- Flags outliers for `data/incomes_outlier.csv` using the IQR rule
- Computes covariance for two paired datasets
- (Optional) Saves plots to `images/` if matplotlib is installed

Summary

- Mean alone is not enough; dispersion describes spread

Exit question: For a dataset with strong outliers, which spread measure would you report first and why?

Summary

- Mean alone is not enough; dispersion describes spread
- IQR is robust; variance/SD quantify typical deviation

Exit question: For a dataset with strong outliers, which spread measure would you report first and why?

Summary

- Mean alone is not enough; dispersion describes spread
- IQR is robust; variance/SD quantify typical deviation
- Covariance captures joint variation (sign matters; scale matters)

Exit question: For a dataset with strong outliers, which spread measure would you report first and why?