STATISTICS FOR DATA SCIENCE

Concepts • Formulas • Excel • Examples

Why Do We Use Statistics in Data Science?

Turn raw data into meaningful information and insights

Summarize data: What is typical? How much variation?

Understand relationships between variables

Make decisions under uncertainty using samples

Foundation for ML models: feature understanding, assumptions, evaluation

Types of Data – Quick Recap

- Qualitative (Categorical)
  - **Nominal:** categories without order (e.g., gender, color)
  - **Ordinal:** ordered categories (e.g., Low/Medium/High)
- Quantitative (Numerical)
  - **Discrete:** counts (e.g., number of items sold)
  - **Continuous:** measured (e.g., height, weight, temperature)

	Categorical	**Quantitative**
Definition	*Take on names or labels*	*Take on numeric values*
Examples	Marital Status	Height
	Smoking Status	Population Size
	Eye Color	Square Footage
	Level of Education	Class Size

Levels of Measurement (Scale Types)

Nominal: labels only, no order (e.g., Male/Female, City)

Ordinal: order exists, gaps unknown (e.g., 1st, 2nd, 3rd rank)

Interval: numeric, equal intervals, no true zero (e.g., °C temperature)

Ratio: numeric, equal intervals, true zero (e.g., height, weight, sales)

Levels of Measurement

Nominal	Ordinal	Interval	Ratio
"Eye color"	"Level of satisfaction"	"Temperature"	"Height"
Named	Named	Named	Named
	Natural order	Natural order	Natural order
		Equal interval between variables	Equal interval between variables
			Has a "true zero" value, thus ratio between values can be calculated

[Levels of Measurement: Nominal, Ordinal, Interval and Ratio]

Descriptive vs Inferential Statistics

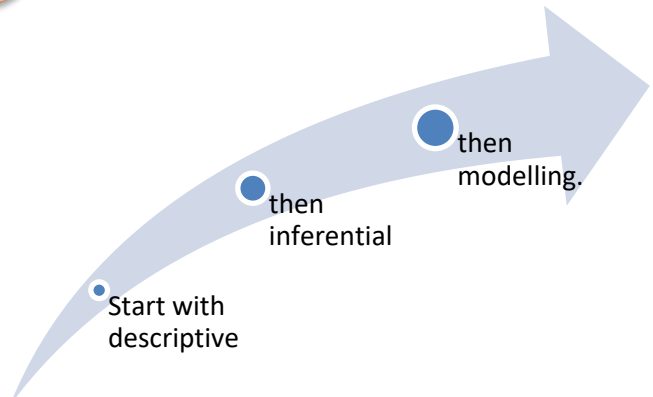
Descriptive Statistics

- Summarize and describe the dataset you have
- Examples: mean, median, standard deviation, histogram, boxplot

Inferential Statistics

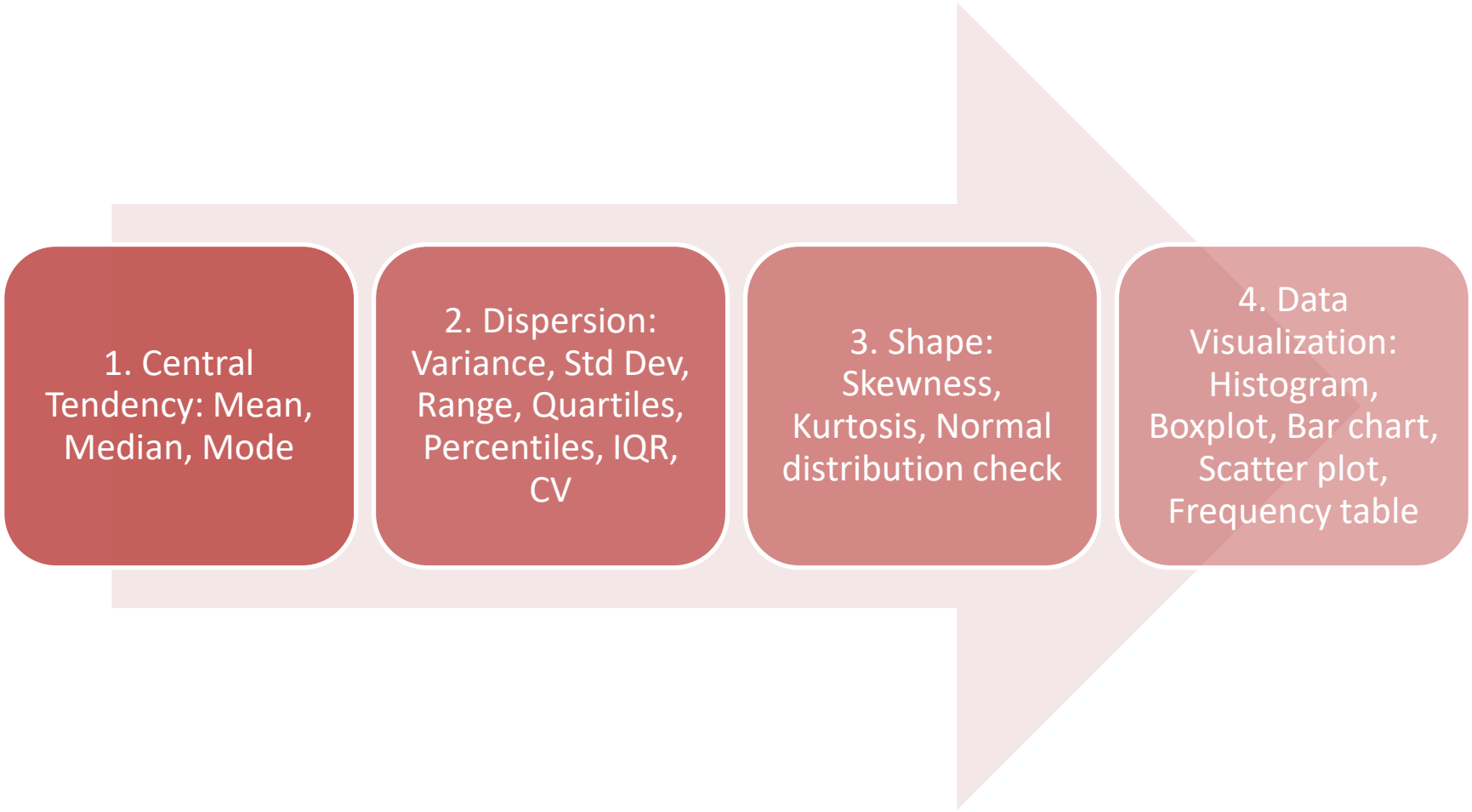
- Use sample data to draw conclusions about a larger population
- Examples: confidence intervals, hypothesis tests, t-test, chi-square

In data science



DESCRIPTIVE STATISTICS

What Is Included?



Central Tendency – Mean

- Definition: Arithmetic average;

  sum of values ÷ number of values.
- Formula (sample): $\bar{x} = (\Sigma xi) / n$
- Example: Values = 10, 12, 15, 13, 20 → $\bar{x}$ = (10+12+15+13+20)/5 = 14
- In Excel: =AVERAGE(A2:A6)
- Use when: data is roughly symmetric and no extreme outliers.

📌 1. Mean (Average)

👉 Formula

$$\text{Mean} = \frac{\sum X}{N}$$

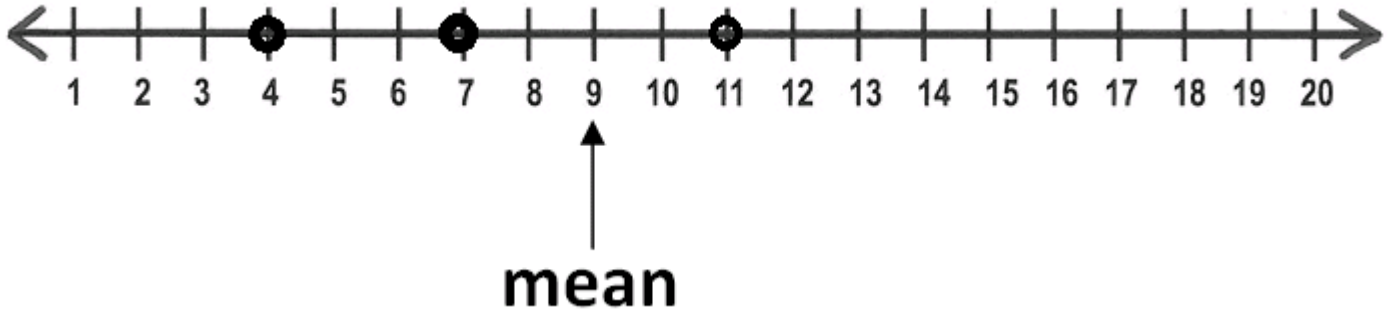
Where:

- $\sum X$ = sum of all observations
- N = number of observations

Example:

Data = 10, 20, 30

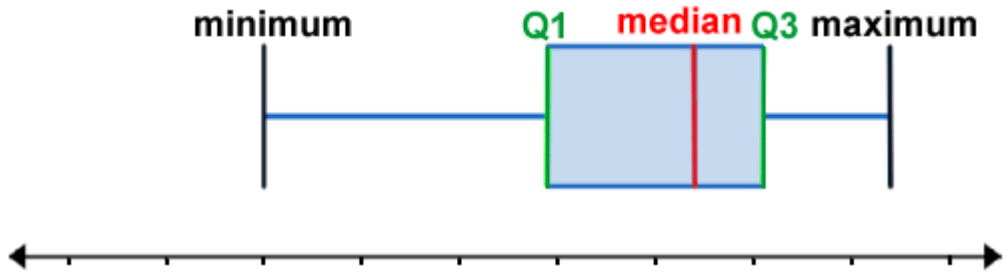
$$\text{Mean} = (10 + 20 + 30)/3 = 60/3 = 20$$



Central Tendency

Median

- Definition: Middle value when data is sorted.
- If n is odd: median is middle value; if even: average of two middle values.
- Example: 5, 7, 9, 12, 100 → Median = 9 (more robust than mean with outlier 100)
- In Excel: =MEDIAN(A2:A6)
- Use when: data is skewed or has outliers.



2. Median

The value which lies in the center when data is arranged in ascending order.

Case 1: Odd number of observations

$$\text{Median} = X_{\frac{n+1}{2}}$$

Example:

Data = 2,4,6,8,10

n = 5

Median = 6

Case 2: Even number of observations

Median = Average of the middle 2 values

$$\text{Median} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Example:

Data = 2,4,6,8

n=4

Median = (4+6)/2 = 5

Central Tendency

Mode

- Definition: Most frequently occurring value in the dataset.
- Example: 1, 1, 2, 3, 3, 3, 4 → Mode = 3
- Useful for: categorical data (e.g., most common category).
- In Excel: =MODE.SNGL(A2:A8)
- Can have: no mode, one mode, or multiple modes.

3. Mode

Value which occurs **most frequently** in a dataset

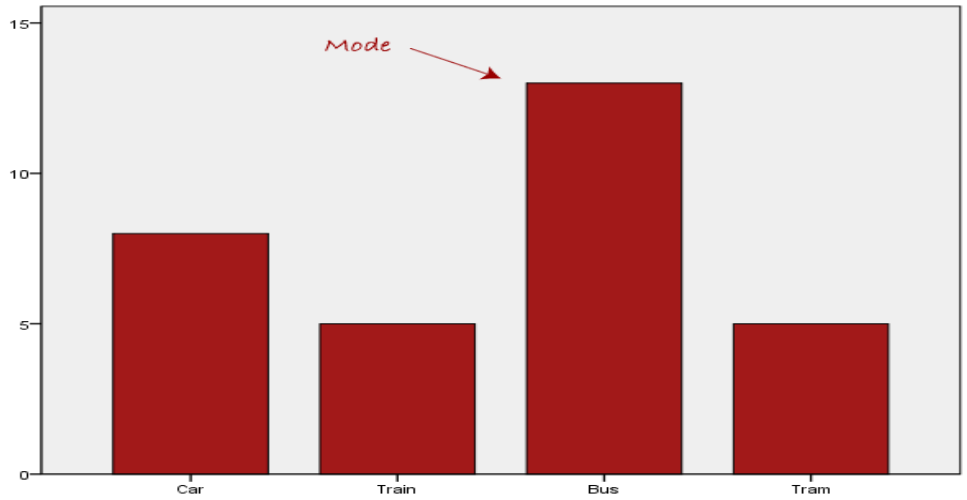
Formula (for simple dataset)

$$\text{Mode} = \text{most repeated value}$$

Example:

Data = 3,4,4,6,7

Mode = 4



continuous Data – Mean

Class Interval	Frequency
0–10	4
10–20	6
20–30	10
30–40	5
40–50	5

⭐ 1️⃣ **Mean (Grouped Data)**

Formula

$$x = \frac{\sum f\, m}{N}$$

F – Means Frequency
M – mid value of class
N – sum of frequency

Step-1: Find Mid Values

Class	F	Mid value (m)
0–10	4	5
10–20	6	15
20–30	10	25
30–40	5	35
40–50	5	45

Step-2: Find f×m

Class	F	m	f×m
0–10	4	5	20
10–20	6	15	90
20–30	10	25	250
30–40	5	35	175
40–50	5	45	225
	N = ΣF = 30		Σfm = 760

- Σfm = 20 + 90 + 250 + 175 + 225
- Σfm = **760**

⭐ 1 Mean

Formula

$$x = \frac{\sum f\, m}{N}$$

F – Means Frequency
M – mid value of class
N – sum of frequency

Mean

$$\bar{x} = \frac{760}{30} = 25.33$$

✓Mean = 25.33

Central Tendency

Median

⭐ 2️⃣ **Median (Grouped Data)**

Formula

$$Median = l + \frac{\frac{N}{2} - c.f}{f} \times h$$

Where
l = lower limit of median class
cf = cumulative frequency before median class
f = frequency of median class
h = class width

Class Interval	Frequency
0–10	4
10–20	6
20–30	10
30–40	5
40–50	5

Step-1: N/2

- N = 30
- N/2 = 15

Step-2: Find cumulative frequency

Class	f	c.f
0–10	4	4
10–20	6	10
20–30	10	20
30–40	5	25
40–50	5	30

Median class = 20–30
(because c.f just crosses 15)

Assign values

l = 20

cf = 10

f = 10

h = 10

Apply formula

$$Median = 20 + \frac{15 - 10}{10} \times 10$$

$$= 20 + \frac{5}{10} \times 10$$

$$= 20 + 5 = 25$$

✓**Median = 25**

⭐ **2 Median (Grouped Data)**

Formula

$$Median = l + \frac{\frac{N}{2} - c.f}{f} \times h$$

Where

l = lower limit of median class

cf = cumulative frequency before median class

f = frequency of median class

h = class width

Central Tendency

Mode

Mode (Grouped Data)
Formula

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Symbol	Meaning	How to find it
l	Lower class boundary of the modal class	Look at the class with highest frequency; take its lower limit
f_1	Frequency of the modal class	Highest frequency in the table
f_0	Frequency of the class before modal class	Frequency just above the modal class
f_2	Frequency of the class after modal class	Frequency just below the modal class
h	Class width	Upper limit – lower limit of any class (all must be equal)

Problem

Class Interval	Frequency
0–10	4
10–20	6 f0
20 L –30	10 f1
30–40	5 f2
40–50	5

Solution

Step 1

Class	F
0–10	4
10–20	6
20–30	10 ← highest (modal class)
30–40	5

Note:

- If modal class = last class ⇒ **f2 = 0**
- If modal class = first class ⇒ **f0 = 0**

Step 2

Modal class = class having maximum frequency → 20–30
So
l = 20
f1 = 10
f0 = 6
f2 = 5
h = 10

Step 3

Apply formula

$$Mode = 20 + \frac{10 - 6}{2(10) - 6 - 5} \times 10$$

$$= 20 + \frac{4}{20 - 11} \times 10$$

$$= 20 + \frac{4}{9} \times 10$$

$$= 20 + 4.44 = 24.44$$

✓**Mode = 24.44**

Question 1

Class Interval	Frequency
0–10	5
10–20	8
20–30	12
30–40	7
40–50	8

Question 2

Class Interval	Frequency
5–15	6
15–25	9
25–35	14
35–45	10
45–55	6
55–65	5

Question 3

Class Interval	Frequency
0–20	3
20–40	11
40–60	18
60–80	9
80–100	4

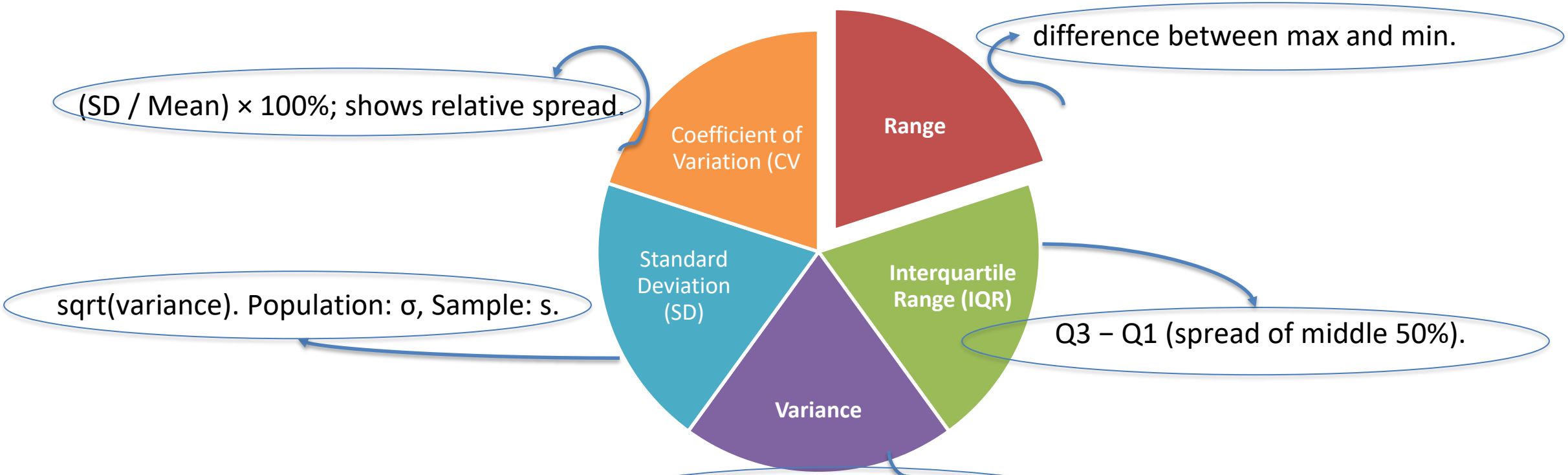
Find

- ✓ Mean
- ✓ Median
- ✓ Mode

Dispersion

What is *Dispersion*

Dispersion = how spread out the data are



(SD / Mean) × 100%; shows relative spread.

difference between max and min.

sqrt(variance). Population: σ, Sample: s.

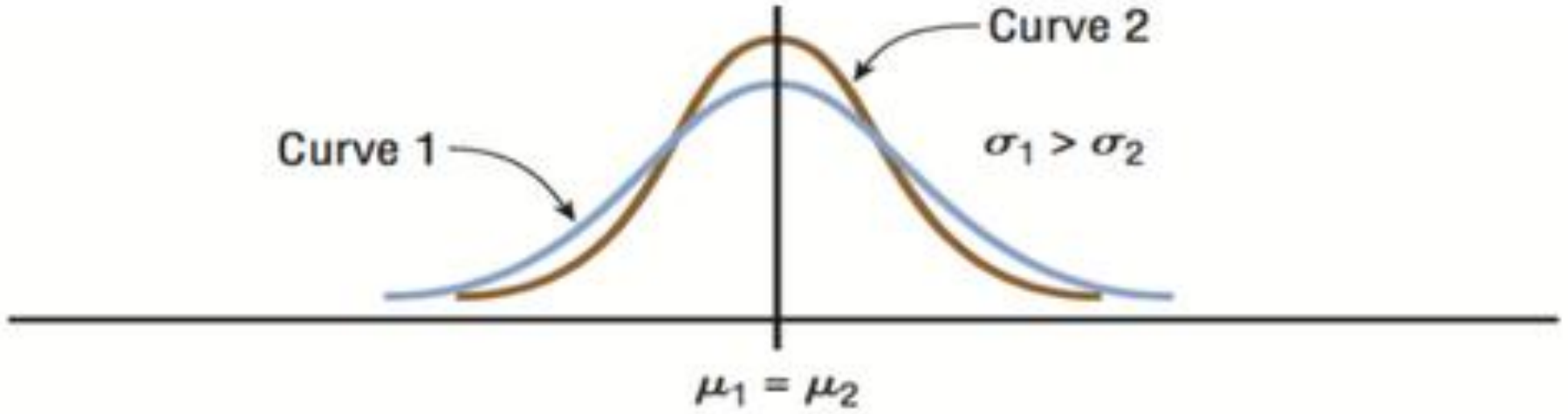
Q3 − Q1 (spread of middle 50%).

average of squared deviations from mean.

- **Population variance** uses N in denominator (σ^2).
- **Sample variance** uses (n−1) in denominator (s^2)

Measures of Dispersion – Why We Need Them

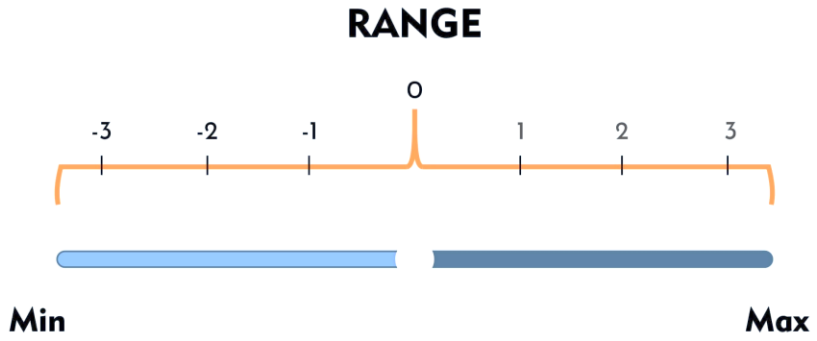
- Two datasets can have same mean but different variability.
- Dispersion tells us how spread out the data is.
- Key measures: Range, Variance, Standard Deviation, Quartiles, IQR, Percentiles, CV.



(a) Same means but different standard deviations

Dispersion – Range

- Definition: Difference between maximum and minimum value.
- Formula: Range = Max – Min
- Example: Values = 10, 12, 15, 13, 20 → Range = 20 – 10 = 10
- In Excel: =MAX(A2:A6)-MIN(A2:A6)
- Simple but sensitive to outliers.



Dispersion – Range

- **Formula:** Range = max(X) − min(X)
- **Excel:** =MAX(range) - MIN(range)

Problem: Discrite

5, 8, 2, 4, 8, 6, 8, 5, 4

Solution:

Step 1 — arrange (sorted)
Sorted: 2, 4, 4, 5, 5, 6, 8, 8, 8
N = 9
Range
Range = max - min = 8 - 2 = 6

Mean
Sum = 5+8+2+4+8+6+8+5+4 = 50
Mean = $\bar{x} = 50/9 = 5.\bar{5} \approx$ **.5 5556** $\approx$ 6

Range for Continuous (Grouped) Data

Range=Highest Upper Limit−Lowest Lower Limit
Because in grouped data we don’t have exact min and max values—so we use class boundaries.

Class Interval
0–10
10–20
20–30
30–40
40–50

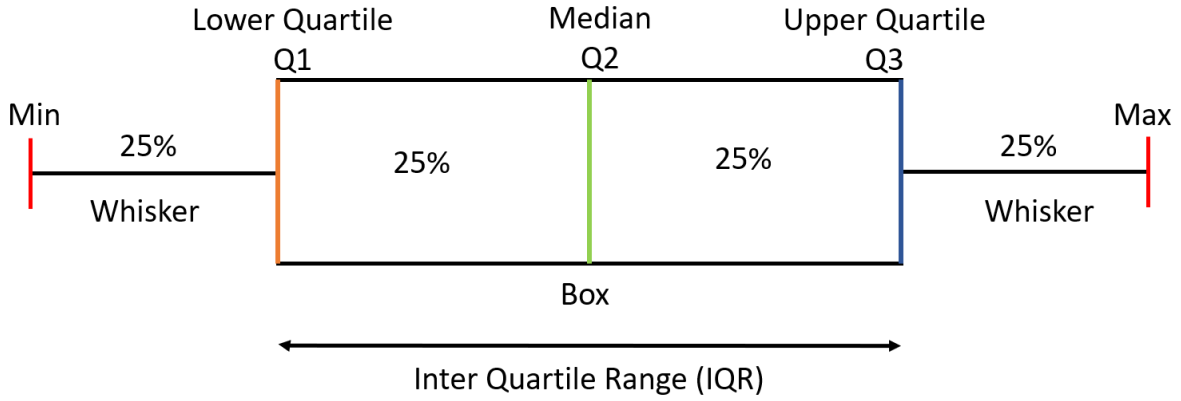
Class
5–15
15–25
25–35

Range = 50 − 0 = **50**

Range = 35 − 5 = **30**

Quartiles, Percentiles & IQR

- Quartiles split data into 4 parts:
  - • Q1: 25th percentile
  - • Q2: 50th percentile (Median)
  - • Q3: 75th percentile
- Percentiles: e.g., 90th percentile (P90) – value below which 90% of data lies.
- Interquartile Range (IQR) = Q3 - Q1; measures middle 50% spread.
- In Excel:
  - • Q1: =QUARTILE.INC(A2:A11,1)
  - • Q2: =QUARTILE.INC(A2:A11,2)
  - • Q3: =QUARTILE.INC(A2:A11,3)
  - • Percentile: =PERCENTILE.INC(A2:A11,0.9)



IQR = Inter Quartile Range

What is IQR?
IQR = Inter Quartile Range
It measures how spread out the **middle 50% of values** are.
So instead of looking at the entire data, IQR only looks at the **core/central region** of the data (without extremes).

Why "Inter Quartile"?
The data is divided into **4 quartiles**
- **Q1 → 25%**
- **Q2 → 50% (median)**
- **Q3 → 75%**
- Q4 → 100%

IQR uses **Q3 and Q1 only**

$$IQR = Q3 - Q1$$

Meaning:
spread of middle portion = values between 25% and 75%

🔷 Why AI and Machine Learning need IQR?

Because

- models work best when data is clean
- outliers mislead learning
- outliers produce wrong predictions
- noise increases errors

So before training:

we remove outliers using IQR ✓

🔷 IQR is used everywhere in AI:

✓ anomaly detection
✓ fraud detection
✓ feature engineering
✓ distribution analysis
✓ data preprocessing
✓ model performance improvement

✓ Summary in 5 lines

- IQR measures central spread
- uses only middle 50%
- avoids extreme values
- ideal for skewed data
- best method to detect outliers

Example A — Odd N

Data:

5, 8, 2, 4, 8, 6, 8, 5, 4 Sorted: 2, 4, 4, 5, 5, 6, 8, 8, 8

IQR (Q1, Q3) — method:

- Median (Q2) position = (n+1)/2 = (9+1)/2 = 5 → median is 5th value = **5**
- Lower half (values before median): 2, 4, 4, 5 → Q1 = median of these 4 values = average of 2nd & 3rd = (4 + 4)/2 = **4**
- Upper half (after median): 6, 8, 8, 8 → Q3 = median of these 4 = average of 2nd & 3rd = (8 + 8)/2 = **8**
- IQR = Q3 − Q1 = **8 – 4 = 4**

Example B — Even N

Data:

12, 15, 11, 14, 10, 13, 16, 18

Step 1 — sort

Sorted: 10, 11, 12, 13, 14, 15, 16, 18

N = 8

IQR (Q1 and Q3) — using the common method where lower half = first n/2 values and upper half = last n/2 values:

- Lower half = 10, 11, 12, 13 → Q1 = median of these = (11 + 12)/2 = **11.5**
- Upper half = 14, 15, 16, 18 → Q3 = (15 + 16)/2 = **15.5**
- IQR = 15.5 − 11.5 = **4.0**

Q1, Q3 and IQR step-by-step using the grouped formula

To find Q1 we use

$$Q_1 = L + \frac{\frac{N}{4} - F}{f} \times h$$

To find Q3

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times h$$

Symbol	Meaning
L	lower boundary of quartile class
N	total frequency
F	cumulative frequency before quartile class
f	frequency of quartile class
h	class width

DATA

Class	f
0–10	6
10–20	8
20–30	14
30–40	12
40–50	10
Total	**50**

So here

$$N = 50$$

Step–1 : Find cumulative frequency

Class	f	c.f
0–10	6	6
10–20	8	14
20–30	14	28
30–40	12	40
40–50	10	50

Step – 2 :

Find Q1

$$Q1 = \frac{N}{4} = \frac{50}{4} = 12.5$$

So Q1 lies in cumulative frequency **>12.5** = **10–20 class**

Symbol	Value
L	10
F	6
f	8
h	10

Symbol	Value
L	10
F	6
f	8
h	10

Formula

$$\boldsymbol{Q_1 = L + \frac{\frac{N}{4} - F}{f} \times h}$$

Put values

$$Q_1 = 10 + \frac{12.5 - 6}{8} \times 10$$

$$Q_1 = 10 + \frac{6.5}{8} \times 10$$

$$Q_1 = 10 + 0.8125 \times 10$$

$$Q_1 = 10 + 8.125 = \boldsymbol{81.521}$$

Step – 3 :

Find Q3

Symbol	Value
L	30
F	28
f	12
h	10

⭐ **Find Q3**

$$Q3 = \frac{3N}{4} = \frac{150}{4} = 37.5$$

Look in cumulative frequency >37.5 → **30–40 class**

To find Q3

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times h$$

Formula

$$Q_3 = 30 + \frac{37.5 - 28}{12} \times 10$$

$$Q_3 = 30 + \frac{9.5}{12} \times 10$$

$$Q_3 = 30 + 0.7916 \times 10$$

$$Q_3 = 30 + 7.916 = \boldsymbol{73.619}$$

⭐ **Find IQR**

$$IQR = Q3 - Q1$$
$$IQR = 37.916 - 18.125 = \boldsymbol{19.791}$$

Middle 50% values of data lie within a spread of **19.79 units**

Why we use IQR?

🟢 **used to measure spread**

It shows **how spread the central 50% of values are**

🟢 **not affected by extreme values**

Better than range because range includes extreme values.

🟢 **used in skewed data**

When data is skewed, IQR gives better idea than mean and SD.

What does IQR tell us?

✓ how tightly packed the central data is
✓ measure of variability
✓ less affected by extreme values
✓ more reliable than range

Why is IQR important?

Because **real world data has outliers**

Examples:

•salary data
•hospital expenses
•house price
•height/weight
•customer purchase amount
•traffic speed
•machine readings etc.

Range, variance, standard deviation—all get influenced by outliers.

But **IQR avoids extreme values**

so it gives a **robust statistical spread**

Why IQR is used in Data Science?

✓ to detect outliers
✓ to handle noise
✓ used while cleaning data
✓ used in boxplots
✓ used in anomaly detection
✓ used before applying ML algorithms

Most common usage:

Outlier detection rule

$$Lower = Q1 - 1.5 \times IQR$$
$$Upper = Q3 + 1.5 \times IQR$$

This helps us automatically find extreme data points.

PERCENTILES

⭐ **What are Percentiles?**

Percentiles divide a dataset into **100 equal parts**.

So each percentile represents **1%** of the ordered data.

Examples:

- **P50** = the value below which 50% of the data lies
- **P90** = the value below which 90% of the data lies
- **P10** = means 10% are below this point

Discrete data (showing position, interpolation)

Discrete data — position method with interpolation:

$$\text{position} = \frac{P}{100}(n+1)$$

Grouped (continuous) data (use grouped formula)

Reminder grouped formula:

$$P_k = L + \frac{\frac{kN}{100} - F}{f} \times h$$

Find $kN/100$(the position in cumulative frequency), identify the class where cumulative frequency ≥ that value, then apply formula.

Symbol	Meaning
L	lower limit of percentile class
K	percentile (like 25 for P25)
N	total freq
F	cumulative freq before that class
f	frequency of that class
h	class width.

Discrete data (showing position, interpolation)

Q1
Find **P25 and P75** for the following data:
Values:
4, 6, 7, 9, 10, 12, 15, 18

Q1 — Data
`4, 6, 7, 9, 10, 12, 15, 18` (n = 8)
P25
•position = $0.25 \times (n+1) = 0.25 \times 9 = 2.25$ →between 2nd and 3rd values.
•2nd = 6, 3rd = 7. Interpolate: $6 + 0.25(7-6) = 6 + 0.25 = \mathbf{6.25}$.
P75
•position = $0.75 \times 9 = 6.75$→ between 6th and 7th values.
•6th = 12, 7th = 15. Interpolate: $12 + 0.75(15-12) = 12 + 0.75 \times 3 = 12 + 2.25 = \mathbf{14.25}$.

Continuous Grouped Data Percentile Questions

Class Interval	f
0–10	4
10–20	6
20–30	10
30–40	15
40–50	5

Find **P30** and **P70**

Reminder grouped formula:

$$P_k = L + \frac{\frac{kN}{100} - F}{f} \times h$$

Find $kN/100$(the position in cumulative frequency), identify the class where cumulative frequency ≥ that value, then apply formula.

Class	f	c.f
0–10	4	4
10–20	6	10
20–30	10	20
30–40	15	35
40–50	5	40

Total $N = 4 + 6 + 10 + 15 + 5 = 40$.

P30

•$kN/100 = 0.30 \times 40 = 12$. Cumulative freqs: 4, 10, **20**, 35, 40 → first ≥12 is class **20–30**.

•For 20–30: $L = 20,\ F = 10,\ f = 10,\ h = 10$.

•$P_{30} = 20 + \frac{12-10}{10} \times 10 = 20 + \frac{2}{10} \times 10 = 20 + 2 = \mathbf{22.0}$.

P70

•$kN/100 = 0.70 \times 40 = 28$. Cumulative freqs: 4,10,20,**35** → first ≥28 is class **30–40**.

•For 30–40: $L = 30,\ F = 20,\ f = 15,\ h = 10$.

•$P_{70} = 30 + \frac{28-20}{15} \times 10 = 30 + \frac{8}{15} \times 10 = 30 + 5.3333\ldots = \mathbf{35.3333}$.

)You can round as needed; I show 4-decimal or 3-decimal where helpful.)

Q1
Find **P40** and **P90**
Values:
2, 2, 5, 8, 10, 12, 15, 25, 26, 28

Q2
Find **P30 and P60**
Values:
5, 7, 7, 8, 9, 11, 12, 14

Q3.
Find **P10, P50, P75**

Class Interval	f
0–5	5
5–10	8
10–15	12
15–20	20
20–25	10

Q4
Find **P25 & P90**

Class Interval	f
0–10	3
10–20	5
20–30	7
30–40	10
40–50	15

ANSWERS:
- Q1: P40 = **8.8**, P90 = **27.8**
- Q2: P30 = **7.0**, P60 = **9.8**

- Q3: P10 = **5.3125**, P50 = **15.625**, P75 = **19.0625**
- Q4: P25 = **22.8571**, P90 = **47.3333**

Variance & Standard Deviation

- Variance measures average squared distance from mean.
- Sample variance: $s^2 = \Sigma(xi - \bar{x})^2 / (n-1)$
- Standard deviation: $s = \sqrt{s^2}$ (same units as data).
- Example idea: marks in a test – higher s means marks are more spread out.
- Excel:
  - • Sample variance: =VAR.S(A2:A11)
  - • Sample std dev: =STDEV.S(A2:A11)
- Use to compare spread between two datasets with similar scale.
- [IMAGE PLACEHOLDER: bell curve with ±1σ, ±2σ, ±3σ marked]

Variance

What is Variance?
Variance tells you **how spread out** the data is from the **mean**.
Simply:
How much values are deviating away from the average.
If values are near mean → variance small
If values far apart → variance large

Mathematical meaning
Variance = average of squared deviations from the mean

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{N}$$

Why squared?
If we sum just (x - mean), positives and negatives cancel.
So we square them → make all values positive.

Standard Deviation (SD)

Standard deviation = **square root** of variance

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

So SD gives deviation **in original units** (variance square units → not useful directly)

Example:

- variance = 25
- SD = √25 = 5

That's why SD is used more than variance.

Imagine heights:

- Everyone exactly 170 cm → variance = 0
- Some 150, some 190 → variance big

So it measures **spread (variation).**

Why these used in Data Science / AI?

Used for:

✓ feature scaling
✓ evaluating spread
✓ standard scores
✓ normal distribution
✓ anomaly detection
✓ control limits
✓ forecasting
✓ machine learning models

Examples:

- StandardScaler uses mean & SD
- Z-score uses SD
- Gaussian distribution uses SD
- Outlier detection depends on SD

Variance and SD basically describe **uncertainty/spread/noise** in data.

POPULATION VS SAMPLE

✓ **Population**

Population = **ALL** the data you want to study

Example:

- All students in India
- All employees in a company
- All items produced in a month

If we have **every value**, we can calculate **PERFECT variance** → Population variance

✓ **Sample**

Sample = **only some part** of the data

Example:

- 100 students from a school (not all Indian students)
- 20 customers from Amazon
- 30 patients from a hospital

We **don't know the entire population** so we **estimate** the variance

1 Discrete data variance formula

Population variance

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{N}$$

Sample variance

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

Mean:

$$\bar{x} = \frac{\sum x}{n}$$

- Population variance = actual spread
- Sample variance = estimated spread

2 Continuous (Grouped) Variance formula

Use midpoints **m**

Population grouped variance

$$\sigma^2 = \frac{\sum f\, m^2}{\sum f} - \left(\frac{\sum f\, m}{\sum f}\right)^2$$

Mean:

$$\bar{x} = \frac{\sum f\, m}{\sum f}$$

Standard Deviation (SD)

What is SD?

Standard deviation measures **how spread out the values are from the mean**.

👉 Low SD → values are close to mean

👉 High SD → values are widely spread from mean

It tells you the **dispersion**, but in the **same units** as the original data (variance is in squared units).

Why SD instead of Variance?

Variance is:

$$\sigma^2$$

but units are squared (rupees², cm², marks² — doesn't make sense!).

SD solves this:

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

So SD is easy to interpret.

Example:

•variance = 25

•SD = √25 = 5

So average deviation is "5 units", NOT "25 units²"

⭐ **SD Formula (Population)**

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

⭐ **SD Formula (Sample)**

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Discrete

Discrete data variance formula

Population variance

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{N}$$

Sample variance

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

Mean:

$$\bar{x} = \frac{\sum x}{n}$$

Example 1 — Discrete (odd N = 9)

Data: 5, 8, 2, 4, 8, 6, 8, 5, 4

Step 1 — basic counts & mean

N = 9

Σx = 5+8+2+4+8+6+8+5+4 = **50**

Mean $\bar{x} = 50/9 = 5.5555555556$

Step 2 — deviations table

x	x − mean	(x − mean)²
5	−0.5555556	0.30864198
8	2.4444444	5.97530864
2	−3.5555556	12.64197531
4	−1.5555556	2.41975309
8	2.4444444	5.97530864
6	0.4444444	0.19753086
8	2.4444444	5.97530864
5	−0.5555556	0.30864198
4	−1.5555556	2.41975309

Σ (x − mean)² = **36.22222222**

sum of squares (SS)

$$SS = \sum (x - \bar{x})^2$$

Step 3 — variances & SD

•Population variance:

$$\sigma^2 = SS/N = 36.22222222/9 = 4.024691358$$

•Population SD:

$$\sigma = \sqrt{4.024691358} = 2.006163343$$

•Sample variance:

$$s^2 = SS/(n-1) = 36.22222222/8 = 4.527777778$$

•Sample SD:

$$s = \sqrt{4.527777778} = 2.127857556$$

Example 2 — Discrete (even N = 8)

Data: 12, 15, 11, 14, 10, 13, 16, 18

Step 1 — basic counts & mean

N = 8

Σx = 109

Mean $\bar{x} = 109/8 = 13.625$

Step 2 — deviations table

x	x − mean	(x − mean)²
12	−1.625	2.640625
15	1.375	1.890625
11	−2.625	6.890625
14	0.375	0.140625
10	−3.625	13.140625
13	−0.625	0.390625
16	2.375	5.640625
18	4.375	19.140625

Σ (x − mean)² = **49.875**

Step 3 — variances & SD

•Population variance:

$$\sigma^2 = 49.875/8 = 6.234375$$

•Population SD:

$$\sigma = \sqrt{6.234375} = 2.496873044$$

•Sample variance:

$$s^2 = 49.875/7 = 7.125$$

•Sample SD:

$$s = \sqrt{7.125} = 2.669269563$$

Grouped (continuous)

Example 3 — Grouped (continuous) — (h = 10)

Class	f
0–10	4
10–20	6
20–30	10
30–40	5
40–50	5

Step 1 — midpoints m

m = 5, 15, 25, 35, 45

Continuous (Grouped) Variance formula

Use midpoints **m**

Population grouped variance

$$\sigma^2 = \frac{\sum f\, m^2}{\sum f} - \left(\frac{\sum f\, m}{\sum f}\right)^2$$

Mean:

$$\bar{x} = \frac{\sum f\, m}{\sum f}$$

Step 2 — compute f×m and f×m²

Class	f	m	f×m	m²	f×m²
0–10	4	5	20	25	100
10–20	6	15	90	225	1,350
20–30	10	25	250	625	6,250
30–40	5	35	175	1,225	6,125
40–50	5	45	225	2,025	10,125

Σf = N = 30 | Σfm = 760 | Σfm² = 23,950

Step 3 — mean

$$\bar{x} = \frac{\sum f\, m}{N} = 760/30 = 25.33333333$$

Step 4 — grouped population variance & SD

$$\sigma^2 = \frac{\sum f\, m^2}{N} - \left(\frac{\sum f\, m}{N}\right)^2 = \frac{23950}{30} - \left(25.33333333\right)^2$$

Compute:

- 23950/30 = 798.3333333
- $(25.33333333)^2 = 641.7777778$

$$\sigma^2 = 798.3333333 - 641.7777778$$
$$= 156.5555556$$
$$\sigma = \sqrt{156.5555556} = 12.51221625$$

Example 4 — Grouped (continuous) — (h = 5)

Class	f
0–5	5
5–10	8
10–15	12
15–20	20
20–25	10

Step 1 — midpoints m

m = 2.5, 7.5, 12.5, 17.5, 22.5

Step 2 — compute f×m and f×m²

Class	f	m	f×m	m²	f×m²
0–5	5	2.5	12.5	6.25	31.25
5–10	8	7.5	60	56.25	450
10–15	12	12.5	150	156.25	1,875
15–20	20	17.5	350	306.25	6,125
20–25	10	22.5	225	506.25	5,062.5

$\Sigma f = N = 55$

$\Sigma fm = 797.5$

$\Sigma fm^2 = 13{,}543.75$

Step 3 — mean

$$\bar{x} = 797.5/55 = 14.5$$

Step 4 — grouped variance & SD

$$\sigma^2 = \frac{13543.75}{55} - 14.5^2 = 246.25 - 210.25 = 36.0$$

$$\sigma = \sqrt{36.0} = 6.0$$

Continuous (Grouped) Variance formula

Use midpoints **m**

Population grouped variance

$$\sigma^2 = \frac{\sum f\, m^2}{\sum f} - \left(\frac{\sum f\, m}{\sum f}\right)^2$$

Mean:

$$\bar{x} = \frac{\sum f\, m}{\sum f}$$

A. DISCRETE DATA (2 Questions)

Q1 – Test Scores (Scenario)
A teacher records marks of 6 students:
50, 60, 70, 80, 60, 50

Q2 – Machine Production (Scenario)
Daily production units of a machine:
12, 15, 14, 18, 21

B. CONTINUOUS (GROUPED DATA)

Q3 – Worker Monthly Salary (Scenario)

Salary Range	Frequency
10–20	4
20–30	7
30–40	6
40–50	3

Q4 – Student Height Distribution

Height (cm)	Frequency
140–150	3
150–160	8
160–170	12
170–180	5

ANSWERS

Q1.

Population variance
= 133.33
Population SD
= 11.55

Q2.

Sample Variance
= 12.5
Sample SD
= 3.53

Q3.

Variance
= **95.35**
SD
= **9.76**

Q4.

Variance
= **109.69**
SD
= **10.47**

Coefficient of Variation (CV)

- Definition: Relative measure of spread = Standard Deviation / Mean.
- Formula: $CV = s / \bar{x}$
- Usage: Compare variability between datasets with different means/units.
- Example: Compare sales variability of two products with different average sales.
- In Excel: =STDEV.S(A2:A11)/AVERAGE(A2:A11)

Coefficient of Variation (CV)

⭐ Coefficient of Variation (CV)

One of the most important metrics in statistics, data science, ML, and AI when comparing variability between datasets.

✅ 1. What is Coefficient of Variation?

The **Coefficient of Variation (CV)** measures how large the standard deviation is compared to the mean.
It tells:
👉 "How much variability exists *relative* to the average value?"
So, it is a **relative measure of dispersion**, not an absolute one.

Why CV is Used? (Purpose)

✓ Compare variability between two datasets with different units or different means.
Example:
•Dataset A mean = 10, SD = 2
•Dataset B mean = 100, SD = 5
Even though SD of B is higher, variability **relative to mean** may be lower.
CV helps compare this fairly.

Formula for Coefficient of Variation

Population CV

$$CV = \frac{\sigma}{\mu} \times 100$$

Sample CV

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

Where:

- σ = population standard deviation
- s = sample standard deviation
- μ = mean
- $\bar{x}$ =sample mean

Example 1: Daily Sales (Discrete Data)
A shop recorded the number of items sold in 7 days:
12, 15, 18, 20, 17, 16, 14
👉 Find:
•Mean
•Standard Deviation
•Coefficient of Variation (CV)

Data:
12, 15, 18, 20, 17, 16, 14
n = 7

Step 1 — Mean

$$\bar{x} = \frac{12 + 15 + 18 + 20 + 17 + 16 + 14}{7}$$

$$\bar{x} = \frac{112}{7} = 16$$

Step 2 — Standard Deviation (Population SD)

Use:

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N}}$$

x	x − mean	(x − mean)²
12	-4	16
15	-1	1
18	2	4
20	4	16
17	1	1
16	0	0
14	-2	4

$$\sum(x-\bar{x})2=42$$

$$\sigma = \sqrt{\frac{42}{7}} = \sqrt{6} = 2.45$$

Step 3 — Coefficient of Variation

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$CV = \frac{2.45}{16} \times 100 = 15.31\%$$

- **Final Answer (Q1)**

- **Mean = 16**
- **SD = 2.45**
- **CV = 15.31%**

CONTINUOUS (GROUPED) DATA

Q3 – Employee Salary Distribution

Class	f
10–20	5
20–30	8
30–40	12
40–50	10
50–60	5

Step 1 — Midpoints (m)
10–20 → 15
20–30 → 25
30–40 → 35
40–50 → 45
50–60 → 55

Step 2 — Mean Formula

$$\bar{x} = \frac{\sum f\, m}{\sum f}$$

Compute fm:

m	f	fm
15	5	75
25	8	200
35	12	420
45	10	450
55	5	275

∑f=40,∑fm=1420

$$\bar{x} = \frac{1420}{40} = 35.5$$

Step 3 — SD

Formula:

$$\sigma = \sqrt{\frac{\sum f\left(m - \bar{x}\right)^2}{\sum f}}$$

m	f	m−mean	(m−mean)²	f × (m−mean)²
15	5	-20.5	420.25	2101.25
25	8	-10.5	110.25	882
35	12	-0.5	0.25	3
45	10	9.5	90.25	902.5
55	5	19.5	380.25	1901.25

$\sum f(m - \bar{x})^2 = 5790$

$$\sigma = \sqrt{\frac{5790}{40}} = \sqrt{144.75} = 12.03$$

Step 4 — CV

$$CV = \frac{12.03}{35.5} \times 100 = 33.90\%$$

- **Final Answer (Q3)**

•Mean = 35.5
•SD = 12.03
•CV = 33.90%

Measure	What It Means	Formula (Basic)	Units	What It Tells You	When to Use	Example Interpretation
Variance (σ^2 or s^2)	Average of squared deviation from mean	$\sigma^2 = \frac{\sum(x-\bar{x})^2}{N}$ (pop) $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ (sample)	**Squared units**	How spread out the data points are	When measuring overall dispersion; in statistical models	Variance = 100 → On average, values deviate "100 units²" from mean
Standard Deviation (σ or s)	Square root of variance	σ=sqrt(σ2)	**Same units as data**	How far each value is from the mean on average	To understand real-world variability easily	SD = 10 → Values vary ~10 units around mean
Coefficient of Variation (CV)	Relative variation (SD compared to mean)	(SD / Mean)×100	**Percentage**	Variation in proportion to the mean	To compare datasets with different units or scales	CV = 25% → Data has moderate variability

Q2 – Machine Output (Discrete Data)

A machine produced the following units over 6 hours:

25, 30, 28, 32, 29, 26

👉 Find:

- Mean
- Standard Deviation
- Coefficient of Variation (CV)

Q4 – Student Height Distribution

Height (cm)	Frequency
140–150	4
150–160	10
160–170	15
170–180	8
180–190	3

👉 Find:

- Mean (Grouped)
- Standard Deviation (Grouped)
- CV

Shape

Skewness, Kurtosis, Normal distribution check

Shape of a Distribution

The **shape** of a data distribution tells us **how the data values are spread and clustered**.
The main shape measures are:

1. **Skewness** – symmetry
2. **Kurtosis** – peakedness / tail behavior
3. **Normal Distribution check** – whether data follows a bell shape

Shape – Skewness

- Skewness measures asymmetry of the distribution.
- Right-skewed (positive): long tail on right (e.g., income data).
- Left-skewed (negative): long tail on left.
- Skewness ≈ 0: roughly symmetric.
- In Excel: =SKEW(A2:A11)
- [IMAGE PLACEHOLDER: diagrams of left/right skewed distributions]

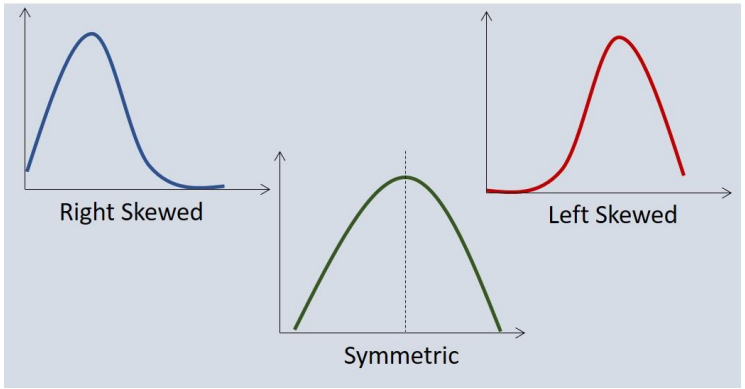
Skewness (Measure of Symmetry)

What is Skewness?
Skewness shows **whether data is symmetric or tilted (skewed)** to one side.

Types of Skewness

Type	Description	Example
Zero skewness	Perfectly symmetric	Height of people
Positive skewness	Tail towards right	Income, salary
Negative skewness	Tail towards left	Marks in easy exam

Visual Understanding
•**Positive Skewed (Right-skewed)**
Mean > Median > Mode
•**Negative Skewed (Left-skewed)**
Mean < Median < Mode



Formula for Skewness (Karl Pearson)

$$\text{Skewness} = \frac{Mean - Mode}{Standard\ Deviation}$$

OR

$$\text{Skewness} = \frac{3(Mean - Median)}{Standard\ Deviation}$$

Example Sum – Skewness

Question:

Mean = 60

Median = 55

Standard Deviation = 10

Answer:

$$\text{Skewness} = \frac{3(60 - 55)}{10} = \frac{15}{10} = 1.5$$

Interpretation:

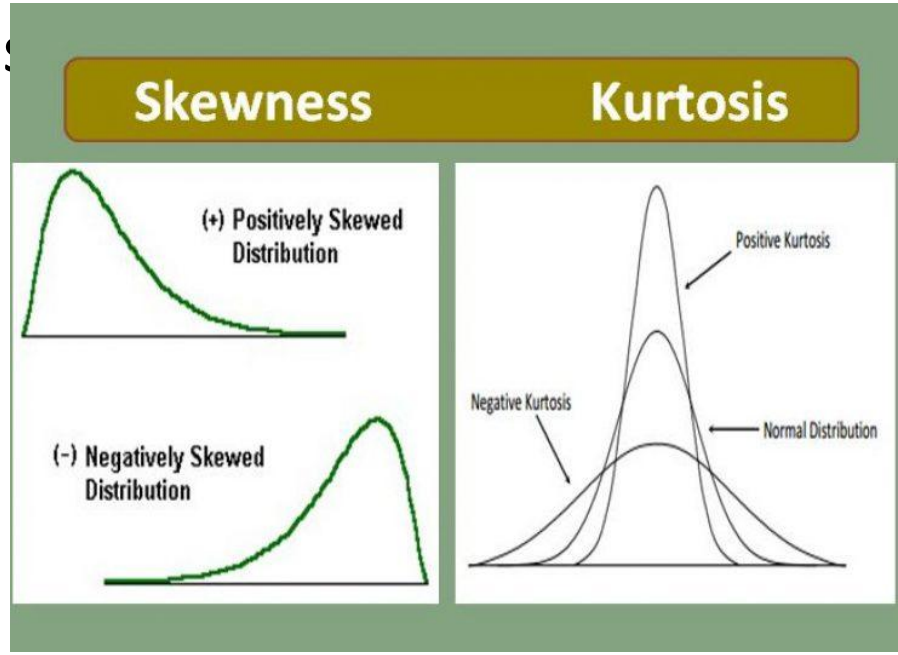
✓ Positively skewed distribution

Interpretation Table

Skewness Value	Meaning
0	Symmetrical
> 0	Positively skewed
< 0	Negatively skewed

Shape – Kurtosis & Normal Distribution Check

- Kurtosis measures 'tailedness' – how heavy the tails are.
- High kurtosis: more extreme values (heavy tails).
- Low kurtosis: fewer extreme values (light tails).
- In Excel: =KURT(A2:A11)
- Normal distribution check:
  - • Histogram looks bell-shaped
  - • Mean ≈ median
  - • Many values within ±1, ±2 standard deviations
- [IMAGE PLACEHOLDER: normal distribution curve]



Kurtosis (Measure of Peakedness)

What is Kurtosis?
Kurtosis measures **how peaked or flat a distribution is compared to a normal distribution**.

Types of Kurtosis

Type	Shape	Meaning
Mesokurtic	Normal	Average peak
Leptokurtic	High peak	More outliers
Platykurtic	Flat	Fewer extreme values

Kurtosis Values

Kurtosis Value	Distribution
= 3	Mesokurtic
> 3	Leptokurtic
< 3	Platykurtic

Note: Some books use **Excess Kurtosis**
Excess Kurtosis = Kurtosis − 3

Example Sum – Kurtosis
Question:
Calculated Kurtosis = 4.8
Answer:
4.8 > 3 → **Leptokurtic**
Interpretation:
✓ Distribution is highly peaked
✓ More extreme values (outliers)

Another Example

Question:

Excess Kurtosis = −1.2

Answer:

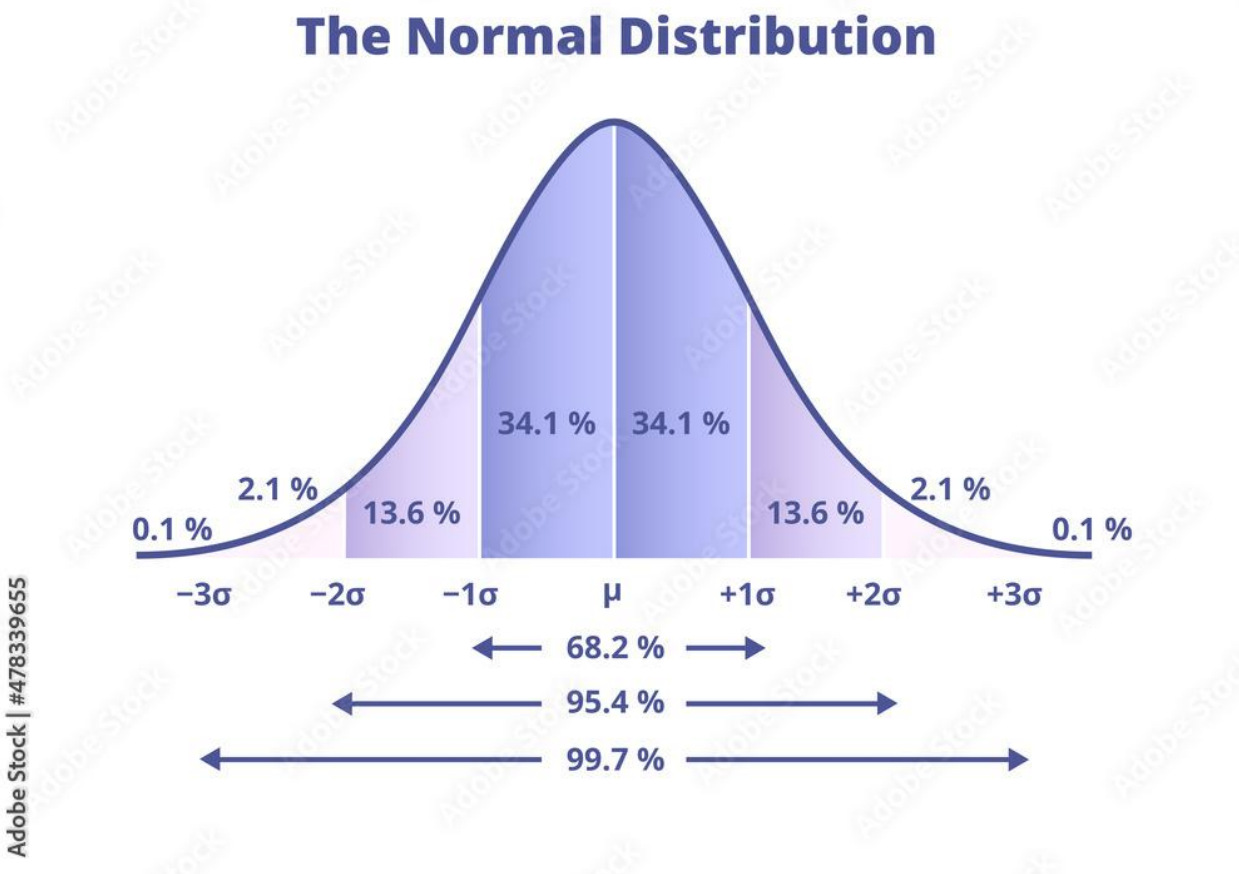
Negative value → **Platykurtic distribution**

Normal Distribution Check

What is Normal Distribution?
A **Normal Distribution** is a **bell-shaped, symmetric distribution** where:
- Mean = Median = Mode
- Skewness = 0
- Kurtosis = 3

Properties of Normal Distribution

Rule	Explanation
68% rule	Data within ±1 SD
95% rule	Data within ±2 SD
99.7% rule	Data within ±3 SD



Example Question – Normal Check

Question:

Mean = 50

Median = 50

Mode = 50

Skewness = 0

Kurtosis = 3

Answer:

✓ Data follows **Normal Distribution**

Practical Normality Check (for students)

Data is approximately normal if:

- Skewness is between **–1 and +1**
- Kurtosis is between **2 and 4**

Univariate vs Multivariate Analysis

What is Univariate Analysis?
Analysis of **one variable only**.

Example	Tool
Student marks	Mean, Median
Age	Histogram
Salary	Skewness

What is Multivariate Analysis?
Analysis of **two or more variables together**.

Example	Tool
Height vs Weight	Correlation
Sales vs Profit	Regression
Marks vs Attendance	Scatter plot

Comparison Table

Feature	Univariate	Multivariate
No. of variables	One	Two or more
Purpose	Describe	Relationship
Examples	Mean, SD	Correlation
Charts	Histogram	Scatter plot

Example Question

Question:

Studying **only student marks** → which analysis?

Answer:

✓ Univariate Analysis

Question:

Studying **marks and attendance together**?

Answer:

✓ Multivariate Analysis

Descriptive Data Visualization in Excel

- Histogram: shows distribution of numeric data.
- Boxplot: shows median, quartiles, outliers.
- Bar chart: compares categories (categorical data).
- Scatter plot: shows relationship between two numeric variables.
- Frequency table: counts in each category or bin.
- Excel:
  - • Histogram/Boxplot: Insert → Statistical Chart
  - • Bar/Column chart: Insert → Column/Bar
  - • Scatter: Insert → Scatter

Inferential Statistics – Overview (Context Only)

- Goal: Use sample to infer about population.
- Key ideas:
  - • Population vs sample
  - • Sampling error & Central Limit Theorem
  - • Confidence Intervals (estimate ranges)
  - • Hypothesis Testing (t-test, chi-square, ANOVA)
- Use after descriptive stats when you want to generalize or test decisions.

Excel – Descriptive Statistics ToolPak

- ToolPak can give a full summary in one step:
- Steps:
  - 1. Enable Analysis ToolPak (File → Options → Add-ins).
  - 2. Data → Data Analysis → Descriptive Statistics.
  - 3. Select input range, check 'Summary statistics'.
- Output: mean, median, mode, std dev, variance, range, skewness, kurtosis, etc.
- [IMAGE PLACEHOLDER: Excel Descriptive Statistics dialog & output table]

Summary – What Students Should Learn Today

- Why statistics is essential in data science.
- Difference between descriptive and inferential statistics.
- Descriptive Statistics:
  - • Central Tendency: Mean, Median, Mode
  - • Dispersion: Variance, Std Dev, Range, Quartiles, Percentiles, IQR, CV
  - • Shape: Skewness, Kurtosis, Normal check
  - • Visualization: Histogram, Boxplot, Bar, Scatter, Frequency tables
- Excel functions and charts to compute and visualize all of the above.