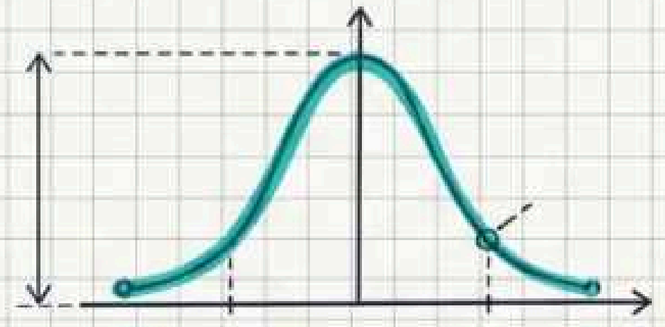
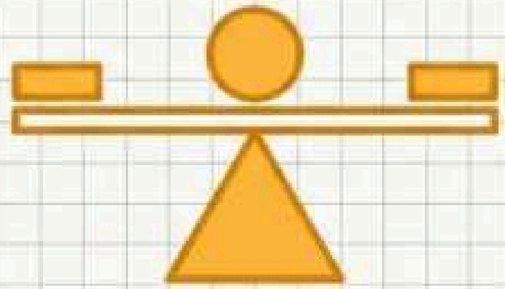
2.1 Overview of One-Variable Distributions

What is a distribution? It tells us what values a variable takes and how often.



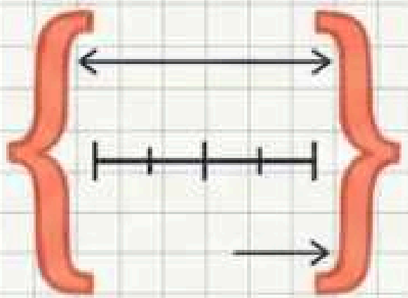
Shape

Draw it.



Center

Mean & Median

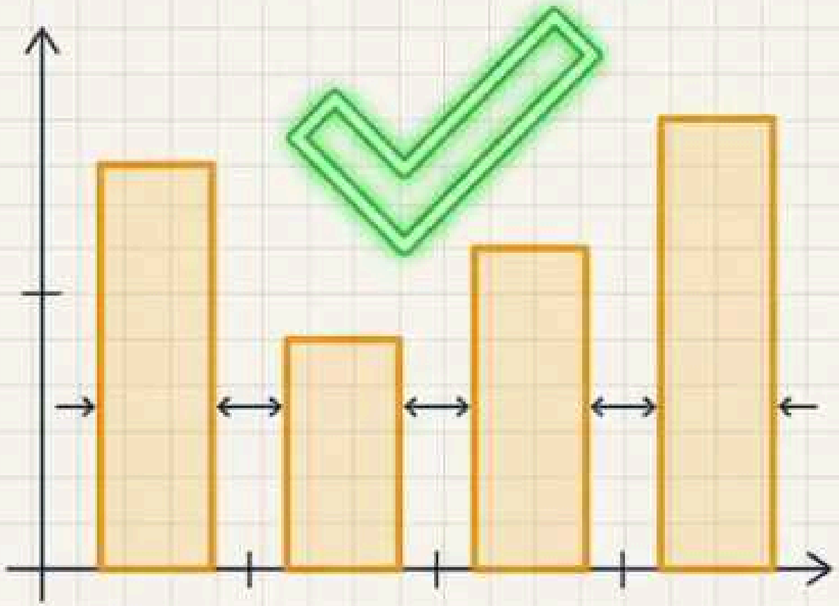


Spread

Range, IQR, & Standard Deviation

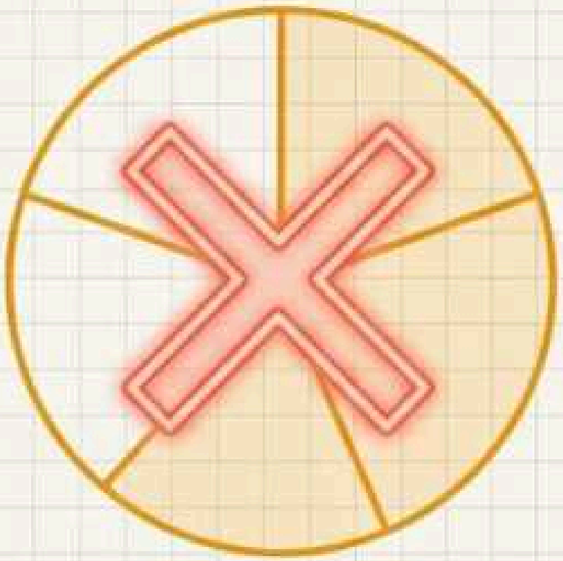
2.2 Categorical Variables: How We Compare

BAR CHART: Length Comparison



Accurate Visual Perception: Clear differences in lengths.

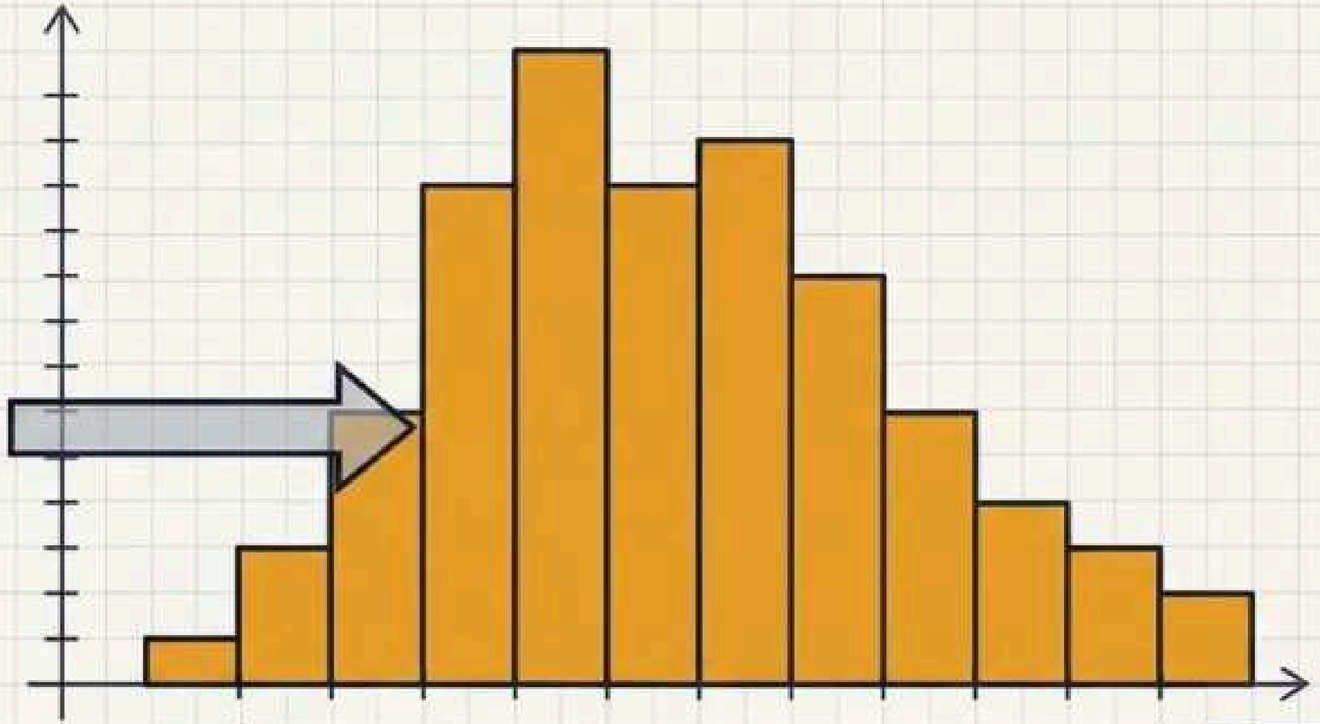
PIE CHART: Angle Comparison



Inaccurate Visual Perception: Difficult angle judgements.

Human eyes compare lengths (bars) much more accurately than angles (slices).
Stick to bar charts for clarity.

2.3 Quantitative Variables & Histograms

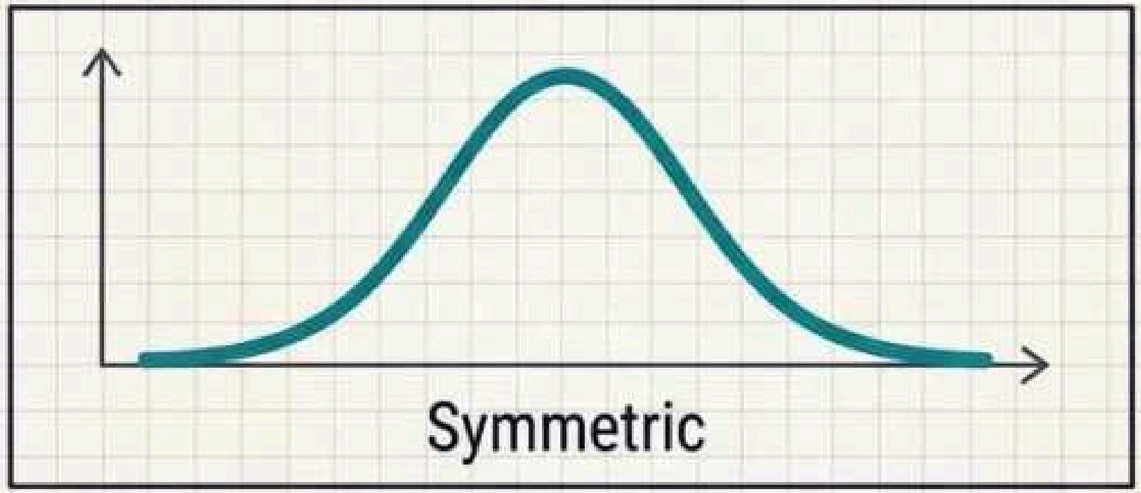


Bars touch because the scale is continuous

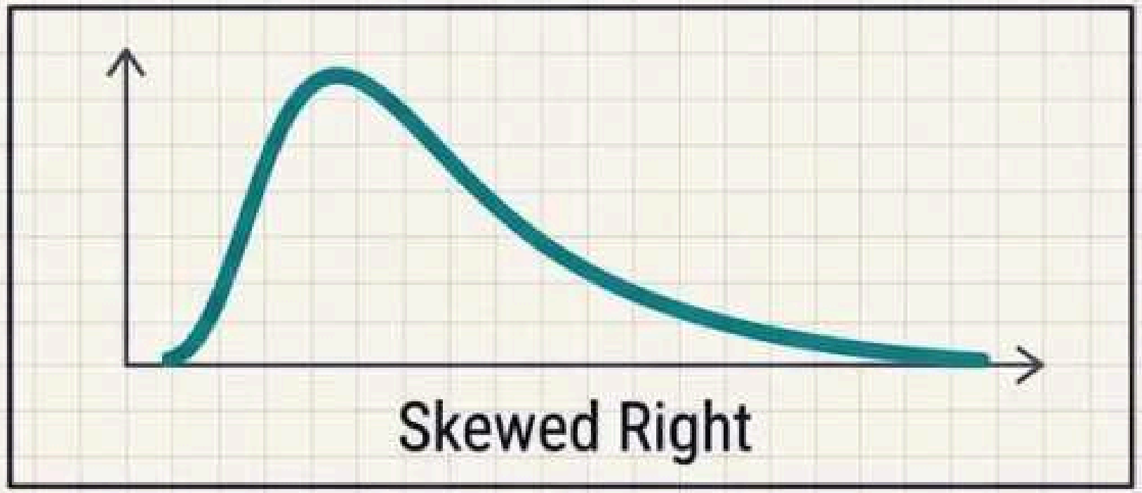
Bins
(Equal intervals on a continuous scale)

PRO-TIP: Bin width changes the look—always try more than one width to see the true shape of the data.

2.4 Describing Shape



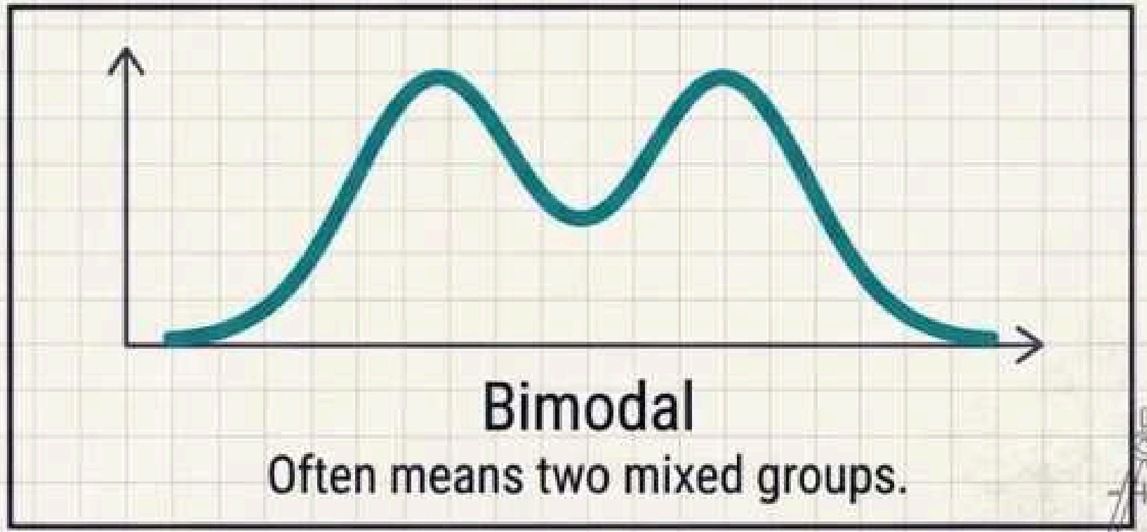
Symmetric



Skewed Right



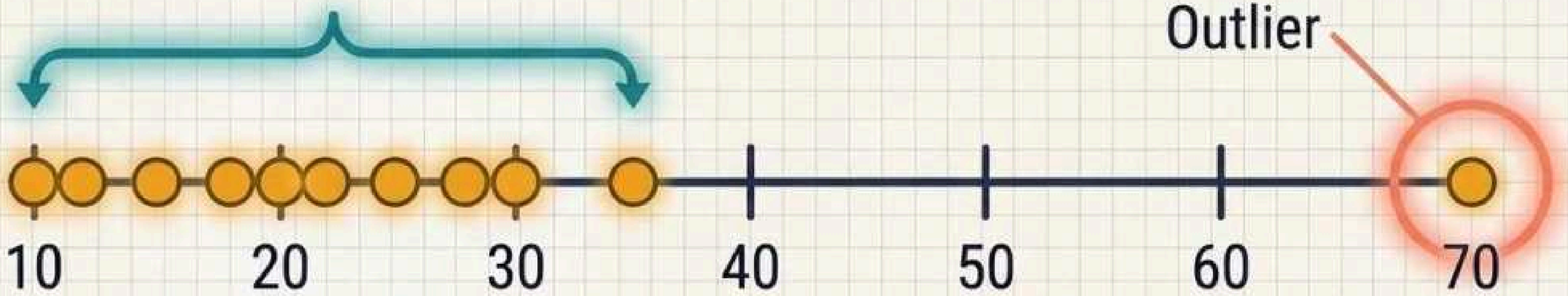
Skewed Left



Bimodal

Often means two mixed groups.

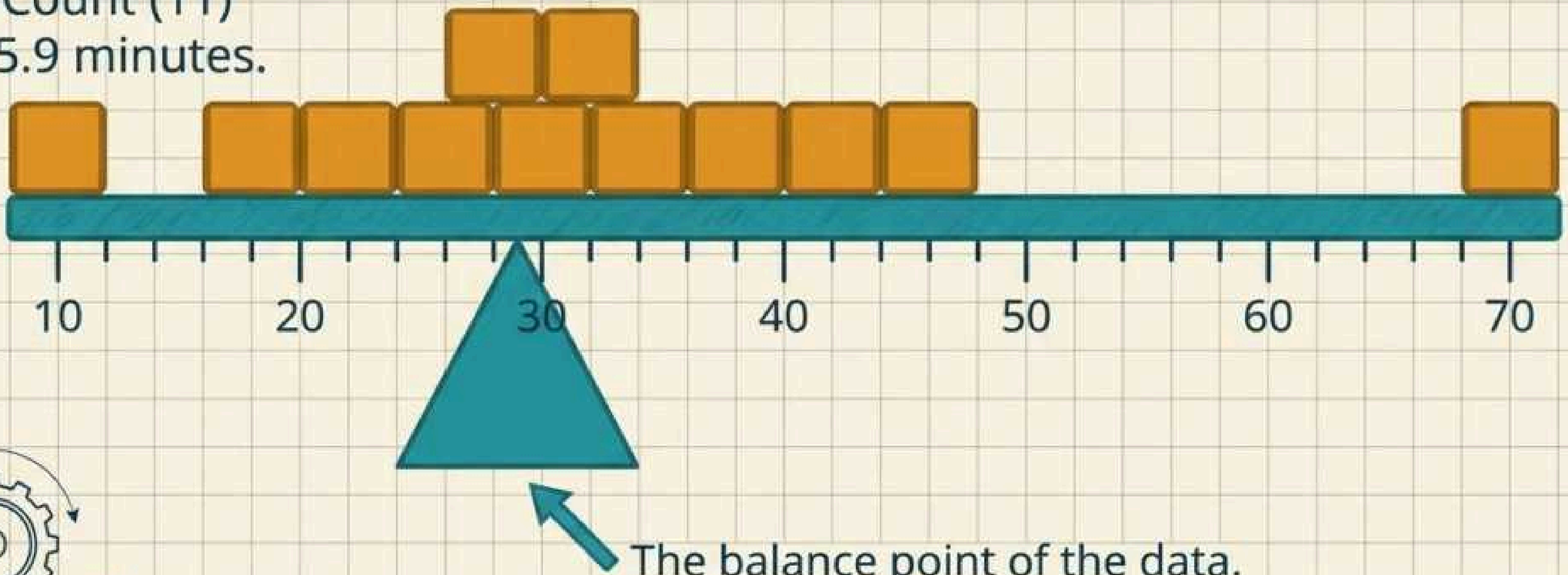
2.5 A Small Data Set: Commute Times



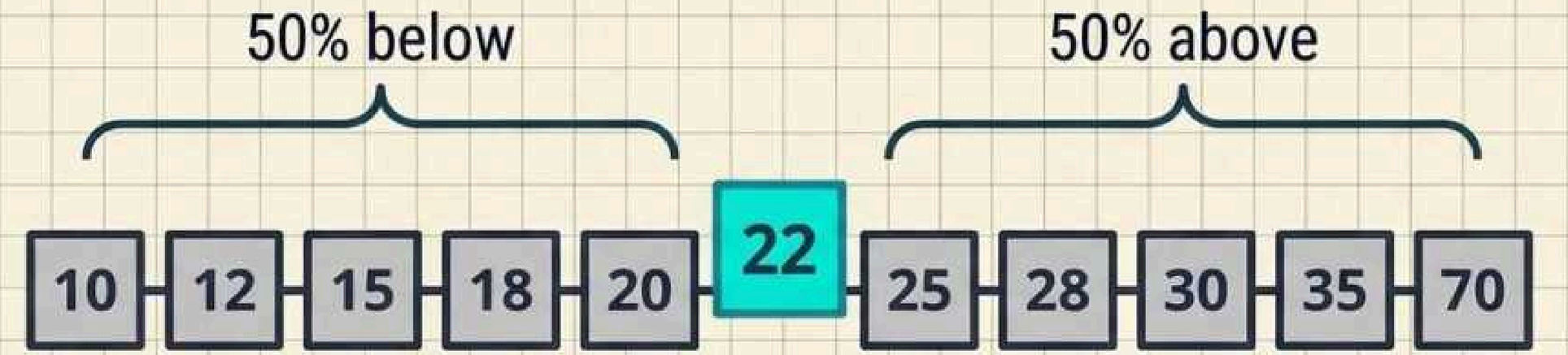
Diagnostic:
Shape is skewed to the right.

2.6 The Mean: Measure of Center

Total (285)
÷ Count (11)
= 25.9 minutes.



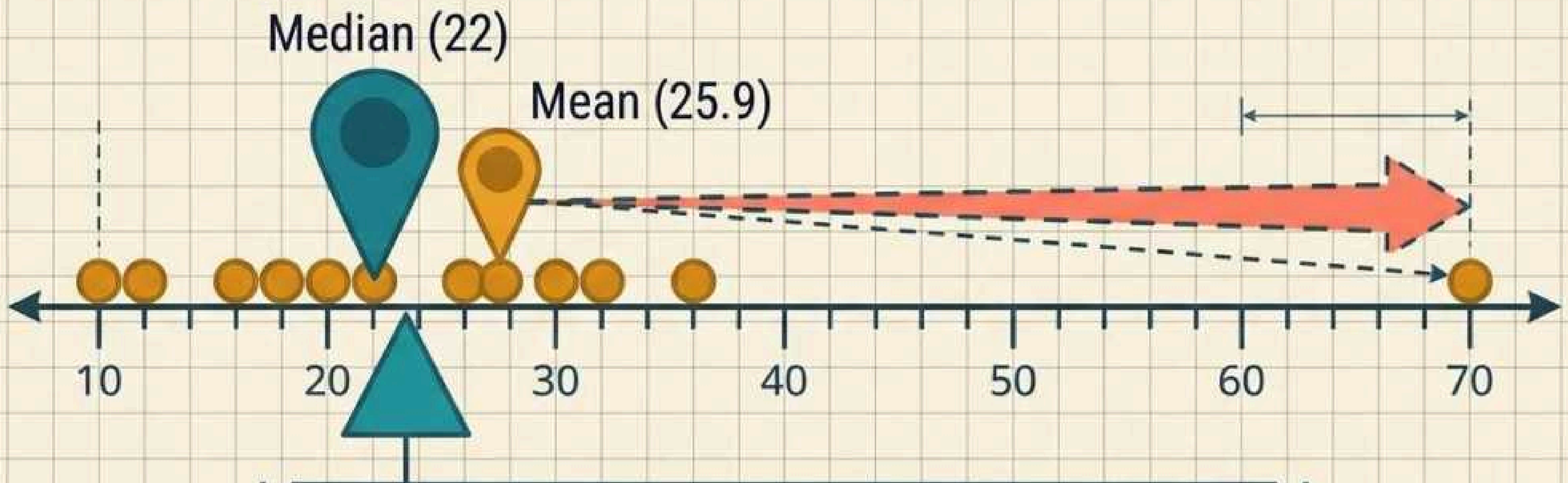
2.7 The Median: The Middle Value



Even counts: If N is even, average the two middle values.

The Mode: The most frequent value (best for categorical data).

2.8 Resistance to Outliers



The **Median** is resistant to outliers.
The **Mean** is pulled toward the long tail.

2.9 Real-Life Impact: House Prices

Context: 9 homes ($210k–$350k) + 1 mansion ($2.1M).

Advert A: Selling Prestige

Average price

$451,000

*(Uses the **Mean**, pulled up by the mansion, to make the neighborhood look wealthy)*

Advert B: Selling Affordability

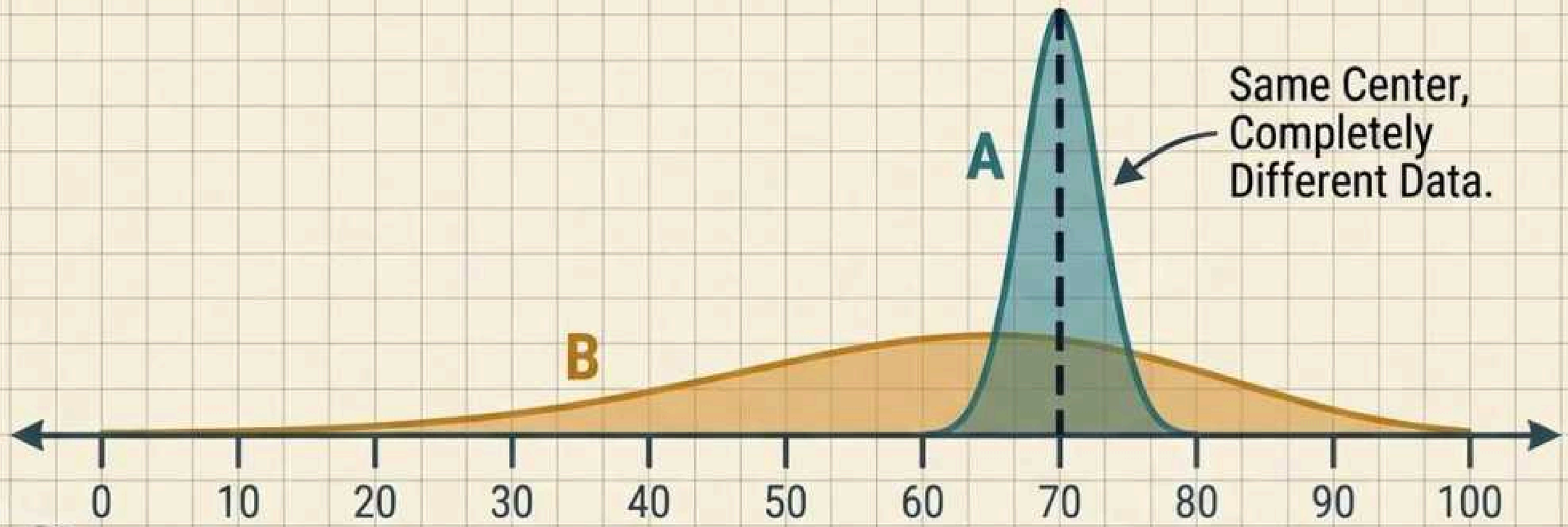
Typical price

$267,500

*(Uses the **Median**, anchored in the middle, to highlight accessibility)*

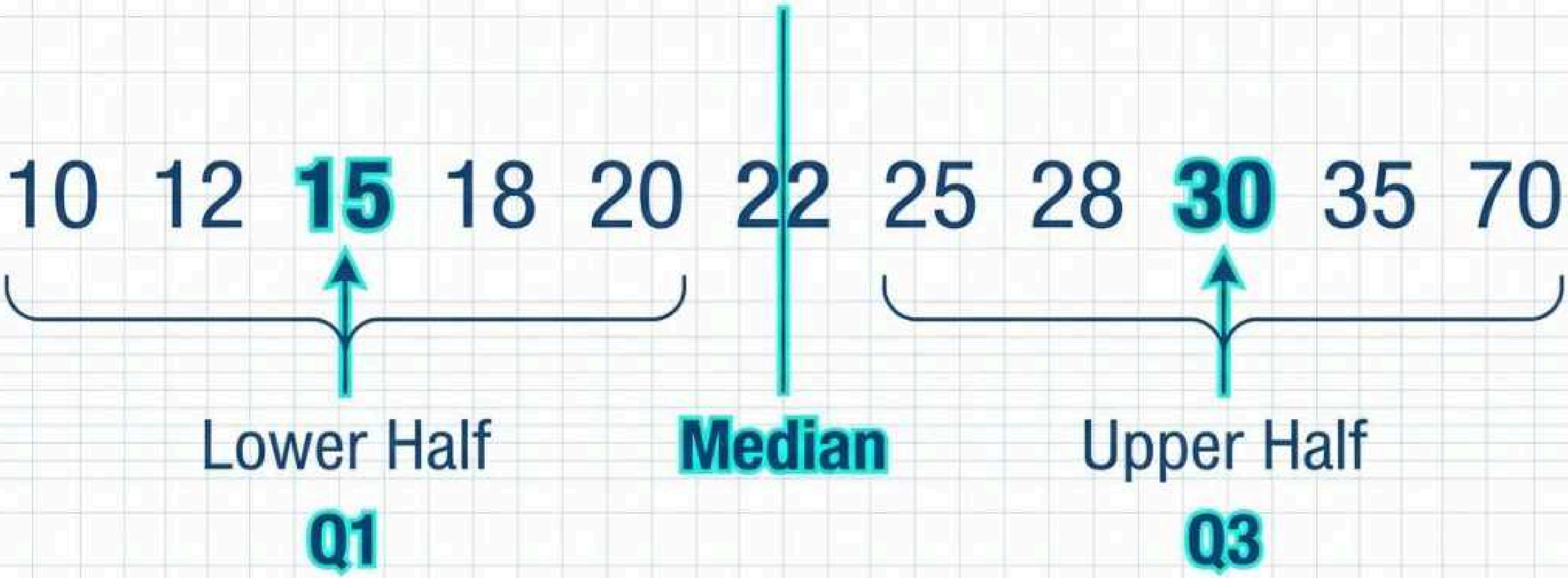
Takeaway: The metric quoted depends entirely on the story the author wants you to believe.

2.10 Why We Need “Spread”



The simplest measure of spread is **Range (Max - Min)**. For commutes: 70 - 10 = 60. **Warning: Fragile**, as it only relies on two extreme values.

2.11 Measuring Spread with Quartiles



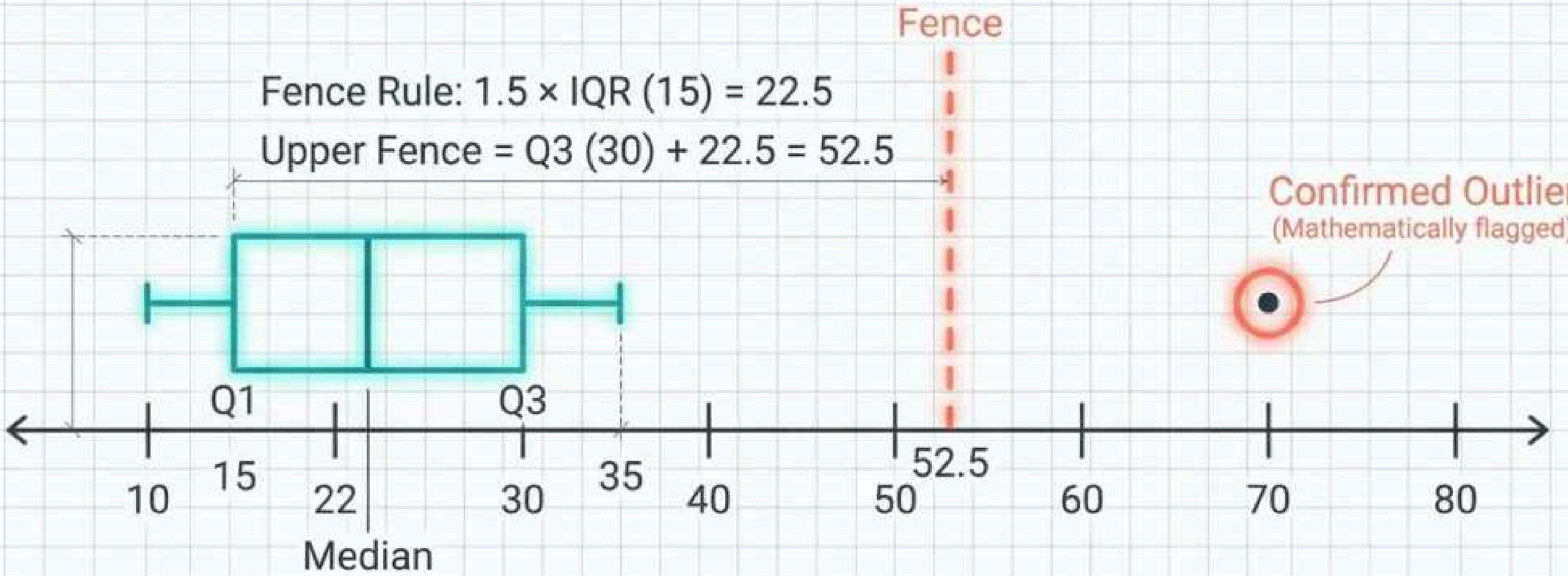
2.12 Interquartile Range (IQR) & 5-Number Summary

IQR = Q3 (30) - Q1 (15) = 15 minutes.

IQR measures the spread of the middle 50% of the data.
Like the median, it is highly resistant to outliers.

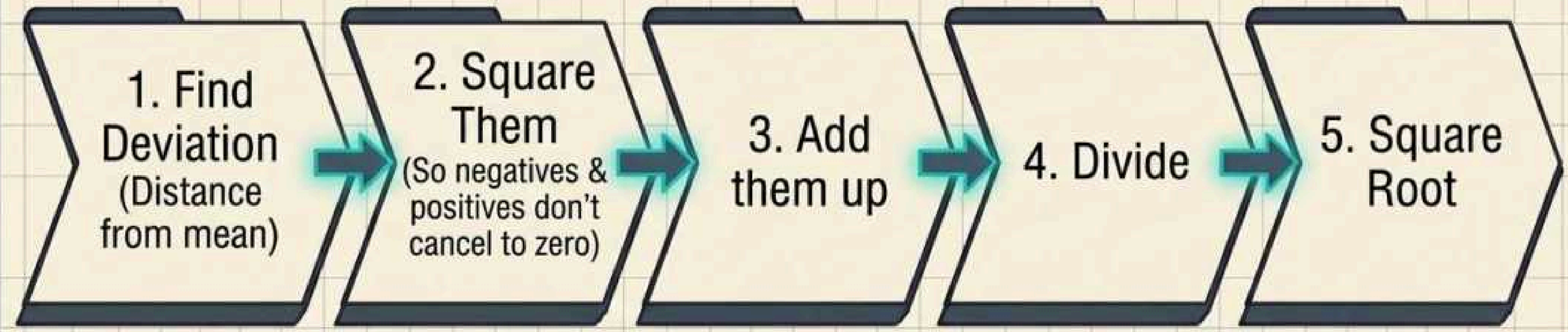
Minimum	Q1	Median	Q3	Maximum
10	15	22	30	70

2.13 Boxplots & The Fence Rule



2.14 Concept: Standard Deviation

Goal: What is the typical distance of observations from the mean?



2.15 Calculating Standard Deviation

Mean of the dataset (2, 4, 4, 4, 5, 5, 7, 9) is 5.

Values	2	4	4	4	5	5	7	9	
Deviations	-3	-1	-1	-1	0	0	2	4	
Squared	9	1	1	1	0	0	4	16	→ Sum = 32

Step 1:	32 ÷ 8 = Variance (4)
Step 2:	Square Root of 4 = Standard Deviation (2)

2.16 Sample vs. Population

Population	Sample
Divide by n (8)	Divide by n-1 (7)
Variance = 4	Variance = 4.57
	Standard Deviation ≈ 2.14

Why n-1? A sample's mean naturally sits closer to its own values than the true population mean would. Dividing by n-1 mathematically corrects this shortfall.

2.17 Which Summaries Should You Use?

Look at the Graph

If roughly symmetric with NO outliers

→ Use **Mean & Standard Deviation**

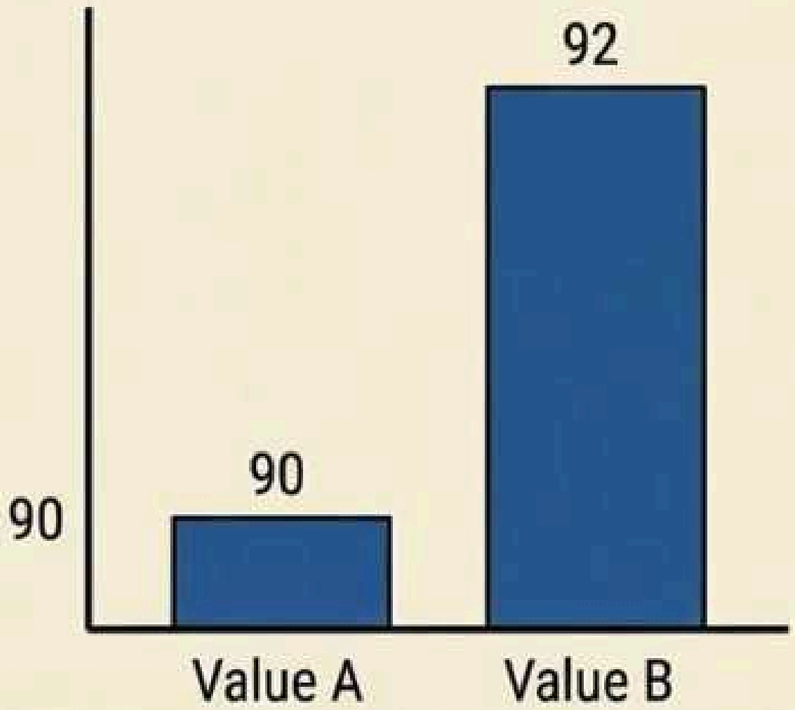
If skewed OR has outliers

→ Use **Median & IQR** (They resist extreme values)

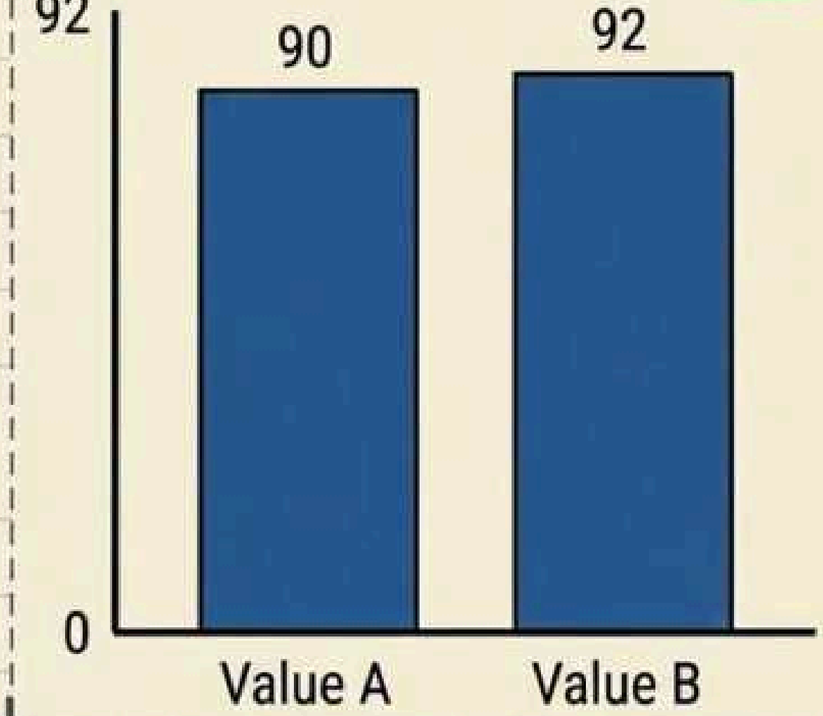
Golden Rule: When in doubt, report both. A number summary without a graph can hide a great deal.

2.18 Misleading Graphs

BAD



GOOD



Warning Checklist:

- ☐ Truncated axes (not starting at zero)
- ☐ 3D effects
- ☐ Uneven scales
- ☐ Cherry-picked timeframes

2.19 Shifting and Scaling Data

Input + C ⇨ Center shifts by **C**, but **Spread** stays exactly the same.

Input × C ⇨ Both **Center** and **Spread** are multiplied by **C**.

Real-World Example:

Converting minutes to hours? You are multiplying by 1/60. Therefore, both the mean and the standard deviation are divided by 60.

2.20 Summary & Next Steps

The Blueprint

Describe any distribution by Shape, Center, Spread, and Outliers.

Spread Measures

Std Dev vs. IQR. Choose resistant measures (Median/IQR) when data is skewed.

Center Measures

Mean (the balance point) vs. Median (the middle value).

Coming Up in Lecture 3:
The Normal Curve

