Chapter 2

Describing Distributions with Numbers

Intended Learning Outcomes

- Compute and interpret measures of central tendency: mean, median, and mode
- Explain the difference between resistant and non-resistant measures
- Compute and interpret measures of variability: IQR, variance, and standard deviation
- Construct and interpret standard and modified boxplots
- Identify potential outliers using the $1.5 \times$ IQR rule
- Choose appropriate summary statistics based on distribution shape

Why Numerical Summaries Matter

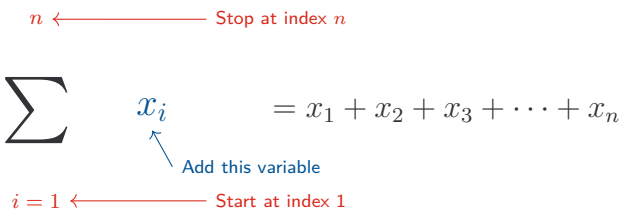
Numerical Summary

A **numerical summary** reduces an entire distribution to a few key numbers that capture its essential features: where the data is **centred** and how much it **varies**.

Two Fundamental Questions:

1. Where is the distribution **centred**? (Central tendency)
2. How **spread out** is the distribution? (Variability)

Indices and Summation



Key Point: The symbol Σ (capital Greek sigma) is an instruction to **add up** everything that follows it.

Measuring Centre: The Mean

Mean (Average)

The **mean** of n observations $x_1, x_2, \ldots, x_n$ is the _____ of all values divided by the _____:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

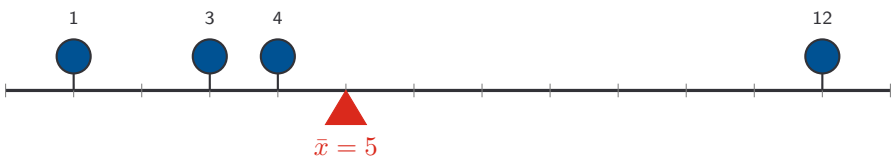


Figure: The mean as a balance point. Data: {1, 3, 4, 12}

Example 2.1: Calculating a Sum

Find $\sum_{i=1}^{3} x_i$ for the `countries_visited` variable.

ID	x
1	2
2	5
3	1

Example 2.2: Calculating a Sum and Mean

Five students reported the number of books they read last year.

(a) Find $\sum_{i=1}^{5} x_i$

(b) Calculate the mean $\bar{x}$

Student	x
1	3
2	7
3	2
4	10
5	5

Robustness of a Measure

Resistant Measure

A statistical measure is **resistant** (or robust) if it is not sensitive to the influence of a few _____ (outliers). A resistant measure focuses on the _____ of the data rather than the tails.

Why does this matter?

- Real data sometimes contains unusual values (typos, measurement errors, genuinely extreme cases)
- Choosing the wrong summary can give a misleading picture

Example: The Mean is Non-Resistant

Ages of five book club members:

```
22, 25, 28, 30, 32
```

(a) Calculate the mean age.

(b) A 95-year-old joins. Calculate the new mean.

(c) Does the new mean represent "typical"?

Measuring Centre: The Mode

Mode

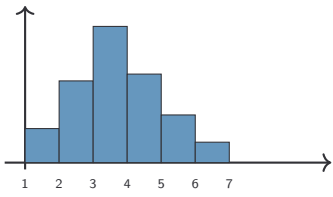
The **mode** is the value that appears _____ in a dataset. It is the only measure of centre that we **can** use for _____ data.

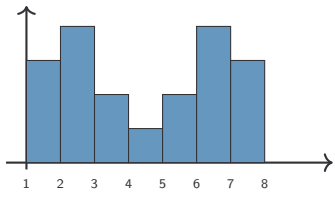
Types of Modality:

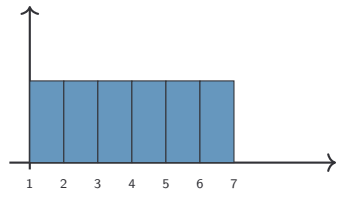
- Unimodal: One peak
- Bimodal: Two peaks
- Multimodal: More than two peaks
- No mode: No value repeats

Visualizing Modality

In each histogram, identify the modes:







Example 2.3: Finding the Mode

For each dataset, identify the mode(s):

Dataset	Mode(s)
$\{4, 7, 7, 7, 9, 12\}$	
$\{2, 4, 6, 8, 10\}$	
$\{1, 3, 3, 4, 5, 6, 6, 9\}$	
T-shirt sizes: {S, M, L, M, L, XL, L}	

Measuring Centre: The Median

Median

The **median** is the _____ when data are arranged in order. It divides the distribution so that _____ of observations fall below and _____ fall above.

Finding the Median

1. Sort the data from smallest to largest.
2. If n is odd: The median is the middle value (position $\frac{n+1}{2}$).
3. If n is even: The median is the average of the two middle values.

Example 2.4: Finding the Median (Odd n)

Find the median of:

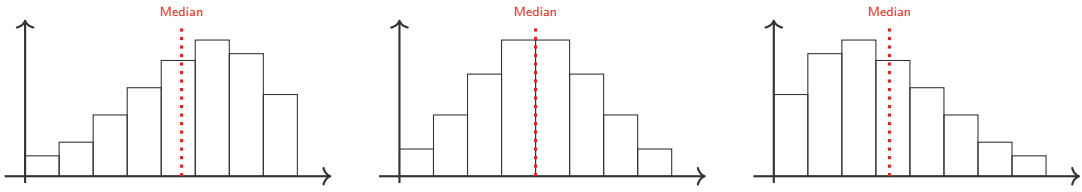
```
5, 21, 4, 8, 22, 11, 5
```

Example 2.5: Finding the Median (Even n)

Find the median of:

```
5, 6, 7, 7.5, 7.5, 8, 8, 8
```

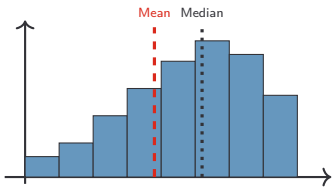
Visualizing the Median on Histograms



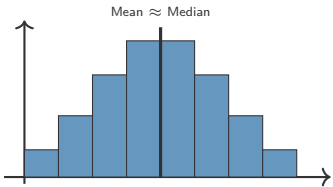
Left-Skewed **Symmetric** **Right-Skewed**

Key Point: The median divides the area under the histogram in half: 50% of data are to the left, 50% to the right.

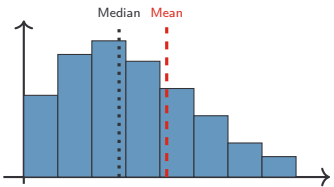
Mean vs. Median and Skewness



Left-Skewed
Mean $<$ Median



Symmetric



Right-Skewed
Mean $>$ Median

From Centre to Spread: The Missing Piece

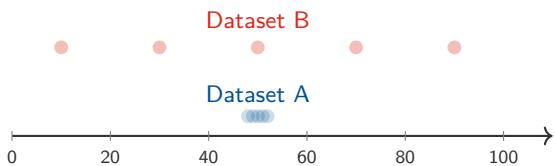
Motivating Question: Consider two datasets with the same mean of 50:

Dataset A:

`48, 49, 50, 51, 52`

Dataset B:

`10, 30, 50, 70, 90`



🔑 **Key Point:** The mean alone cannot distinguish these very different distributions. We need measures of **spread** to complete the picture.

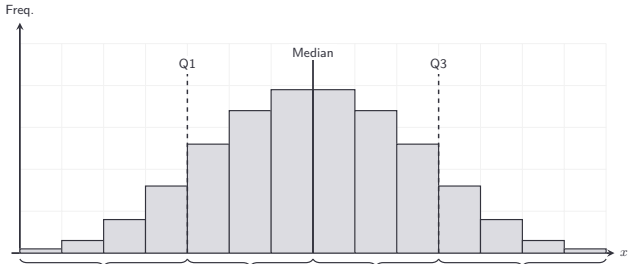
Quartiles

Quartiles

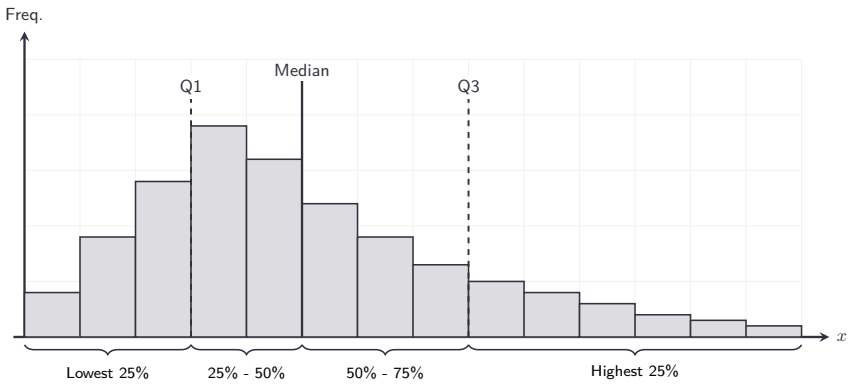
Quartiles divide a sorted dataset into four equal parts.

- **First Quartile (Q_1)**: 25% of data fall below this value
- **Second Quartile (Q_2)**: The median (50% below, 50% above)
- **Third Quartile (Q_3)**: 75% of data fall below this value

Quartiles help describe the spread and center of a distribution.



Quartiles



Quartiles and the Interquartile Range

⚙ Calculating Quartiles

1. **Sort** the data from smallest to largest.
2. Find the **Median (Q_2)**.
3. **If n is odd:** Exclude the median value.
   - Lower Half: All values strictly below the median.
   - Upper Half: All values strictly above the median.
4. **If n is even:** Split the data into two equal halves.
5. **Q_1:** The median of the Lower Half.
6. **Q_3:** The median of the Upper Half.

Example 2.6: Computing Quartiles (Odd n)

Find Q_1, Q_2 (median), and Q_3 for:

```
3, 5, 8, 12, 15, 28, 35
```

Example 2.7: Computing Quartiles (Even n)

Find Q_1, Q_2 (median), and Q_3 for:

```
2, 4, 7, 9, 10, 12, 15, 18
```

The Five-Number Summary

Five-Number Summary

The **five-number summary** of a variable consists of the following five values, arranged in order from smallest to largest:

$$\text{Min} \qquad Q_1 \qquad \text{Median} \qquad Q_3 \qquad \text{Max}$$

Example: Coffee expenses ($)

```
3, 5, 8, 12, 15, 28, 35
```

Find the five-number summary.

Example 2.8: The Interquartile Range

Interquartile Range (IQR)

The **interquartile range (IQR)** is the range of the middle [ ] of the data:

$$\text{IQR} = [\quad]$$

Example: Coffee expenses ($)

3, 5, 8, 12, 15, 28, 35

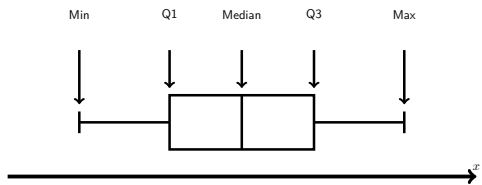
Compute the IQR.

Boxplot

Boxplot

A **boxplot** is a graphical summary of a dataset based on the five-number summary:

- The **box** spans from the first quartile (Q_1) to the third quartile (Q_3).
- A line inside the box marks the **median**.
- **Whiskers** extend from the box to the minimum and maximum values.



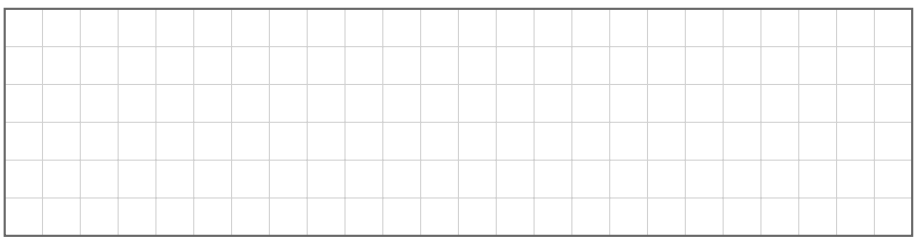
Example 2.9: Drawing a Boxplot

Data:

```
5, 10, 14, 16, 18, 20, 25, 30, 40,
45, 60, 83
```

Five-Number Summary:

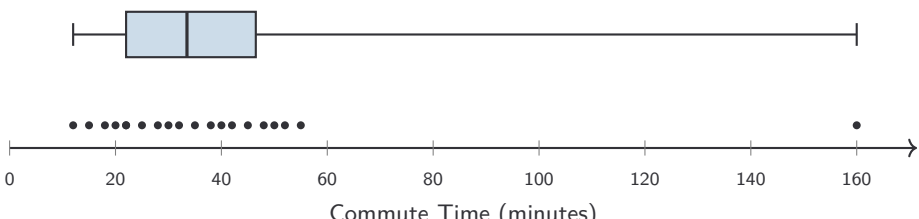
- Min =
- Q_1 =
- Median =
- Q_3 =
- Max =



Example 2.10: The Challenge with Boxplots

Data (minutes):

```
12, 15, 18, 20, 22, 22, 25, 28, 30, 32, 35, 38, 40, 42, 45, 48, 50, 52, 55,
160
```



Identifying Outliers: The $1.5 \times$ IQR Rule

Potential Outliers

A value is a **potential outlier** if it falls more than _____ beyond the quartiles:

- **Lower Fence:** $Q_1 - 1.5 \times \text{IQR}$
- **Upper Fence:** $Q_3 + 1.5 \times \text{IQR}$

Any value _____ the lower fence or _____ the upper fence is flagged as an outlier.

This choice '1.5' IQR is arbitrary and was popularized by John Tukey in the 1970s. Other methods for identifying outliers exist.

Modified Boxplot

Modified Boxplot

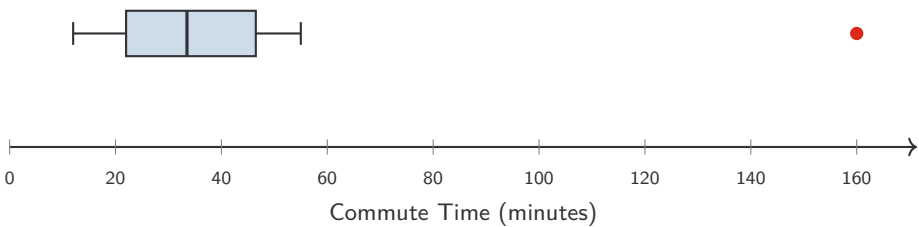
A **modified boxplot**

- displays potential outliers as [ ] and
- its whiskers extend only to the most extreme values *within* the fences.

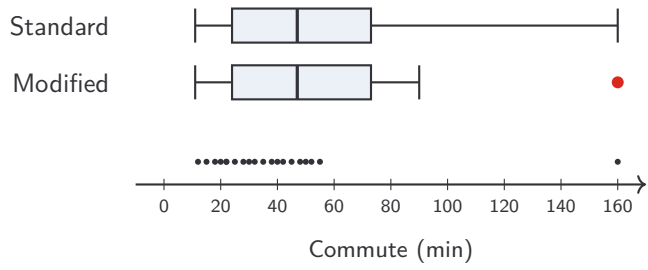
Example 2.10: Modified Boxplot (Daily Commute Times)

Data (minutes):

```
12, 15, 18, 20, 22, 22, 25, 28, 30, 32, 35, 38, 40, 42, 45, 48, 50, 52, 55, 160
```



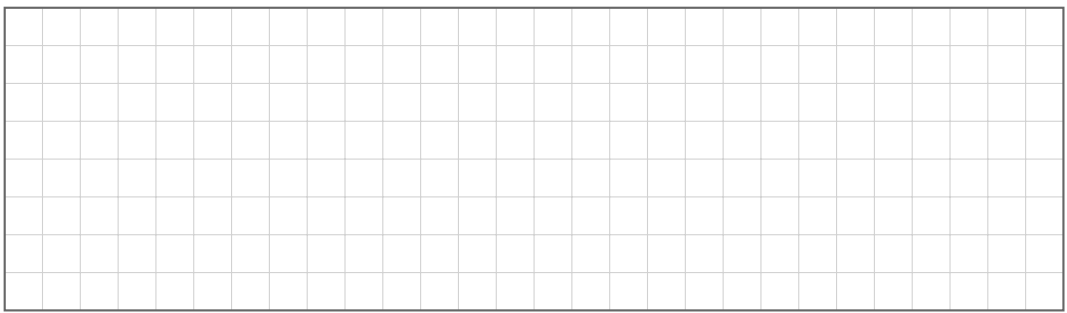
Standard vs. Modified Boxplot



Example 2.11: Modified Boxplot with Two Outliers

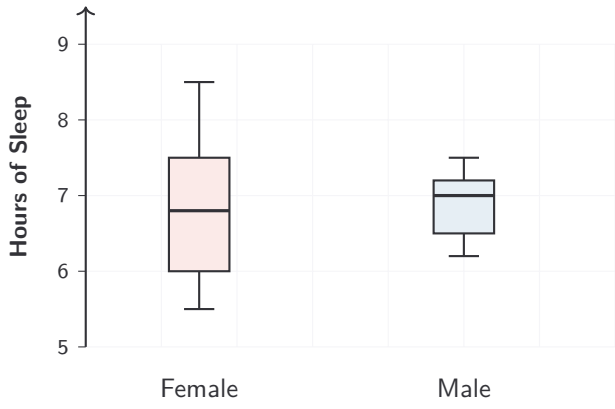
Draw a modified boxplot for the following data:

```
5, 45, 48, 50, 52, 55, 58, 60, 62, 110
```



Example 2.12: Comparing Distributions with Boxplots

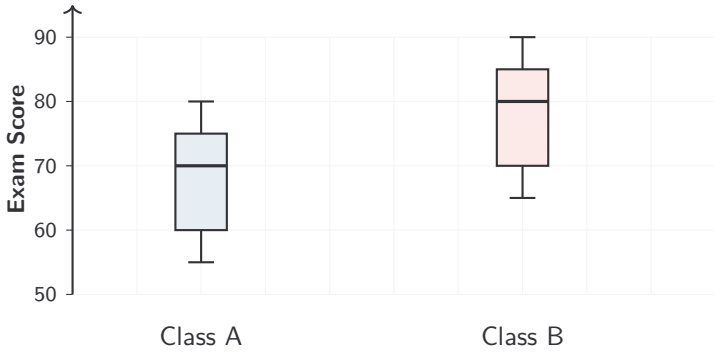
Who sleeps more?



Example 2.13: Comparing Exam Scores: Two Classes

Same spread, different centres

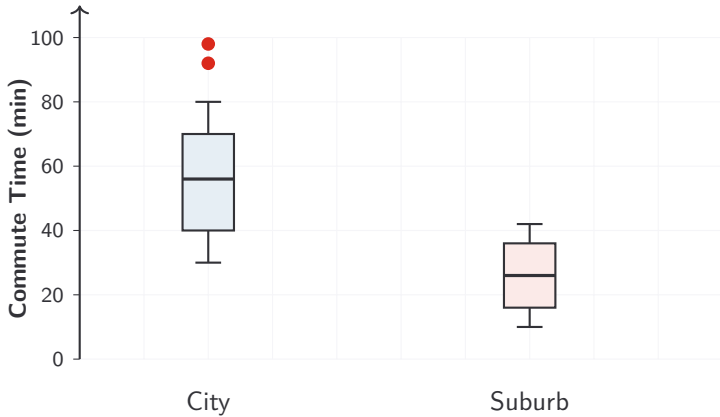
Context: Final exam scores (out of 100) for two statistics classes.



Example 2.14: Comparing Commute Times: City vs. Suburb

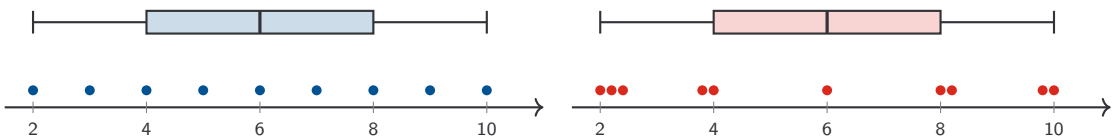
Different centres, different spreads, with outliers

Context: Daily commute times (minutes) for workers in two locations.



Limitation: Same Boxplot, Different Distributions

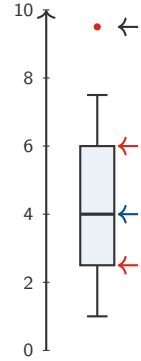
Boxplots hide details about distribution shape



Key Point: Two very different distributions can produce identical boxplots

Reading a Boxplot: Centre, Spread, and Outliers

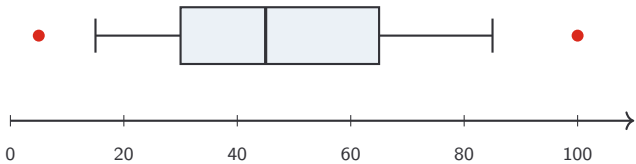
A boxplot reveals all three key features at once



Key Point: A single boxplot shows where the data is centred, how spread out it is, and identifies potential outliers.

Example 2.15: Practice Reading a Boxplot

Daily study time (minutes) for 50 students



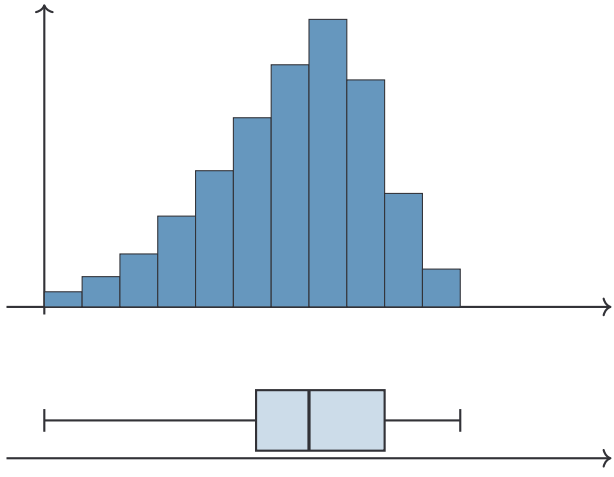
(a) What is the median study time?

(b) What is the IQR?

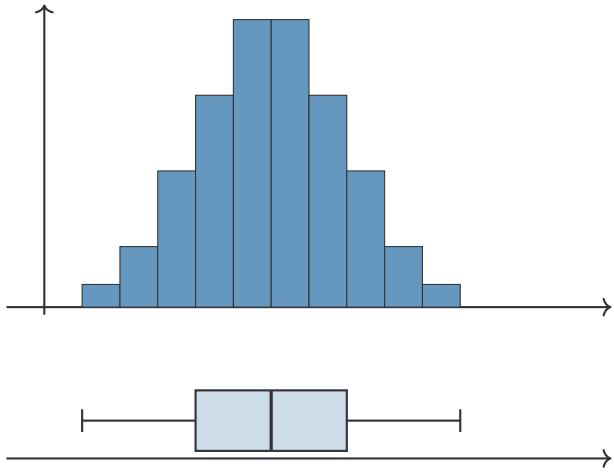
(c) Identify the outliers.

(d) Is the distribution symmetric?

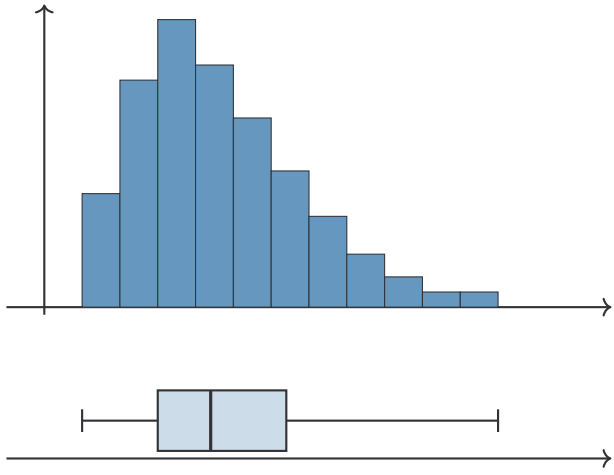
Distribution Shape: Left Skewed



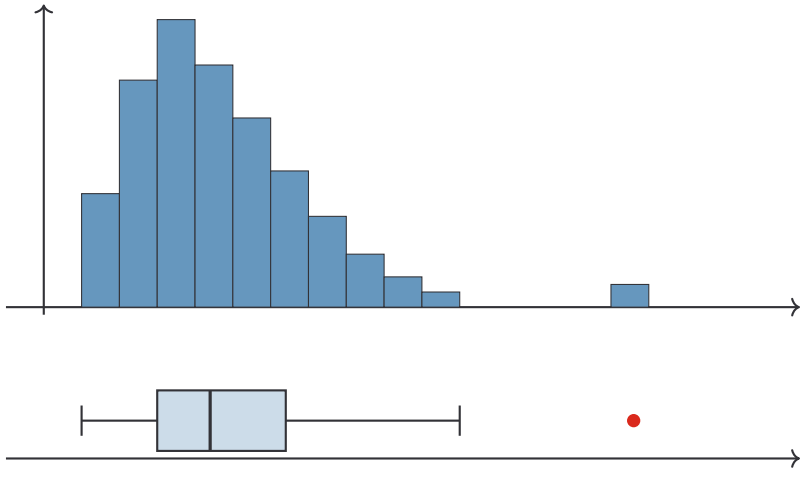
Distribution Shape: Symmetric



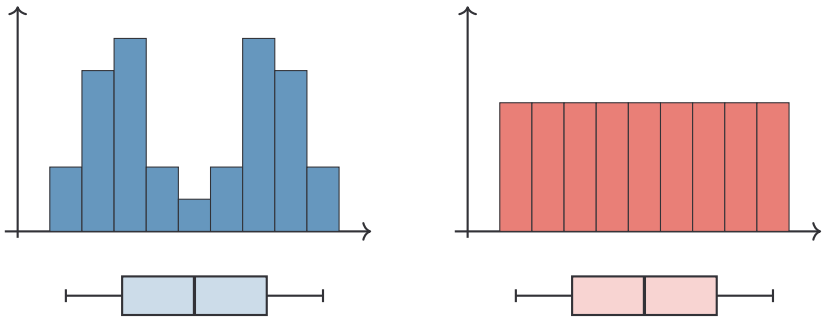
Distribution Shape: Right Skewed



Distribution Shape: Right Skewed with Outlier



Limitation: Boxplots



Beyond Quartiles: A More Precise Measure of Spread

The IQR tells us about the middle 50% of data.

Limitation of IQR:

- Ignores 50% of observations

We could, alternatively

- Measure how far each observation is from the centre
- Average these distances
- This results in the **standard deviation**

🔑 **Key Point:** The standard deviation uses *all* data points to quantify spread, making it a more comprehensive (but less resistant) measure.

Measuring Variability: Deviation

Deviation

A **deviation** is the difference between an observation and the

$$\text{deviation} = x_i - \bar{x}$$

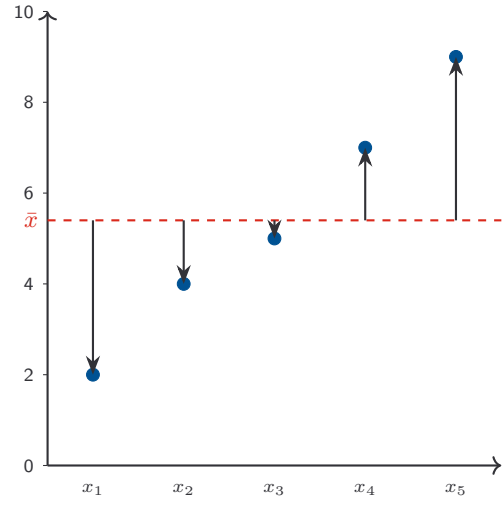
Example 2.16: Visualizing Deviations from the Mean

Data:

```
2, 4, 5, 7, 9
```

Mean: $\bar{x} = 5.4$

x_i	Deviation
2	
4	
5	
7	
9	



Measuring Variability: Variance

Variance

The **variance** measures the average squared deviation from the mean

We divide by $n - 1$ (not n) to get an unbiased estimate of the population variance.

Example 2.17: Computing Variance

Data:

```
2, 4, 5, 7, 9
```

Mean: $\bar{x} = 5.4$

Find the variance.

Example 2.18: Variance Practice

Find the variance for the following dataset

```
3, 6, 7, 8, 13
```

Measuring Variability: Standard Deviation

Standard Deviation

The **standard deviation** is the _____ of the variance:

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

It has the same _____ as the original data, making it more interpretable than variance.

Example 2.19: Computing Standard Deviation

Find the standard deviation for the following data

```
2, 4, 5, 7, 9
```

Example 2.20: Same Mean, Different Spread

Two sections of a statistics course have quiz scores:

Section A:

73, 74, 75, 76, 77

Mean = $\square$

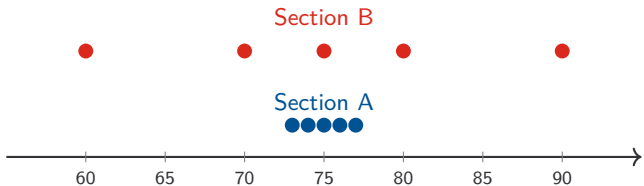
$s \approx \square$

Section B:

60, 70, 75, 80, 90

Mean = $\square$

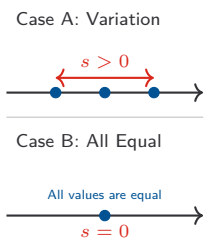
$s \approx \square$

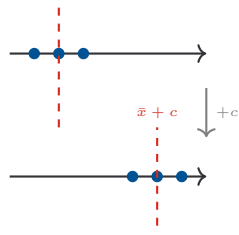


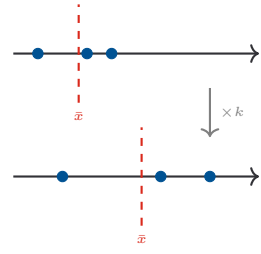
Properties of Standard Deviation

Key Point: The standard deviation satisfies the following properties:

- $s \geq 0$; $s = 0$ only if all values are equal
- Adding/subtracting a constant does not change s
- Multiplying all values by k changes s to $|k|s$







Example 2.21: Adding a Constant to All Values

Original scores:

```
65, 70, 75, 80, 85
```

Mean $= 75$
$s \approx 7.91$

After adding 10 points:

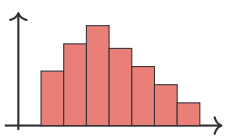
```
75, 80, 85, 90, 95
```

Mean $= 85$
$s \approx$

Choosing Measures of Centre and Variability

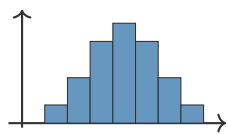
Median & IQR

Use when distribution is **skewed** or when **outliers** are present.



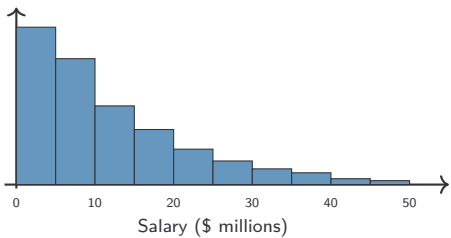
Mean & Standard Deviation

Use when distribution is **symmetric** without outliers.



Example 2.22: NBA Player Salaries (2024–25)

The histogram below shows the distribution of NBA player salaries for the 2024–25 season ($n = 450$ players).



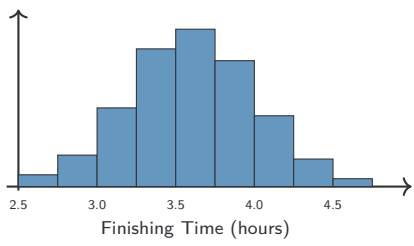
Summary Statistics:
Mean: \$10.2M
Median: \$4.5M
SD: \$11.8M
IQR: \$12.1M

(a) Is this distribution symmetric or skewed?

(b) Which measures should we report?
Centre:
Spread:

Example 2.23: Boston Marathon Finishing Times (2024)

Context: Finishing times (hours) for 500 randomly sampled runners from the 2024 Boston Marathon.



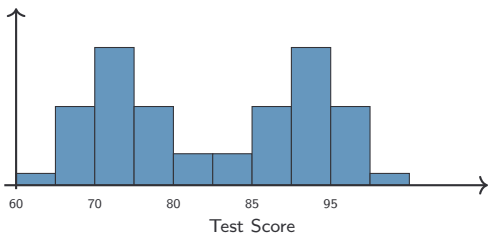
Summary Statistics:
Mean: 3.72 hrs
Median: 3.68 hrs
SD: 0.48 hrs
IQR: 0.65 hrs

(a) Is this distribution symmetric or skewed?

(b) Which measures should we report?
Centre:
Spread:

Example 2.24: Bimodal Symmetric Distribution

A combined sample of test scores (out of 100) from two sections of a course is shown below:



Summary Statistics:
Mean: 80.0
Median: 80.0
Mode: 72 and 90 (bimodal)
SD: 9.4

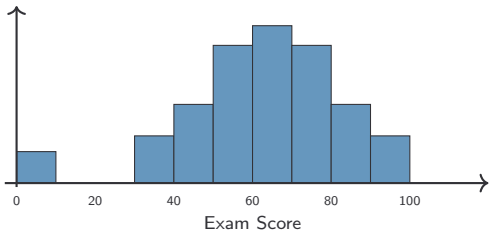
(a) Describe the shape.

(b) What might explain this pattern?

(c) Which statistics are appropriate?

Example 2.25: Symmetric Distribution with Outliers

A class of students took an exam, and their scores (out of 100) are shown below:



Summary Statistics:
Mean: 81.0
Median: 86.0
SD: 19.8
IQR: 16.0

(a) Describe the shape of the distribution.

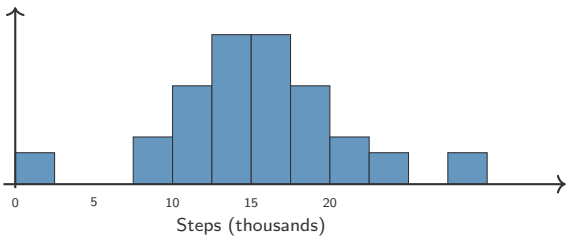
(b) Identify the outlier.

(c) Which statistics should we report?
Centre:
Spread:

Example 2.26: Symmetric Distribution with Outliers on Both Sides

Daily step counts with extreme values

A fitness tracker recorded daily step counts for a group of individuals over a month.



Summary Statistics:
Mean: 10,913
Median: 10,800
SD: 3,445
IQR: 2,600

(a) Describe the shape of the distribution.

(b) Identify the outliers.

(c) Which statistics should we report?
Centre:
Spread:

Example 2.27: Toronto Home Prices (December 2025)

Prices ($ thousands) for 15 homes sold in a Toronto neighbourhood.

685, 720, 745, 780, 795, 810, 825, 850, 875, 890, 920, 985, 1050, 1180, 2450

(a) Calculate the five-number summary.

(b) Check for outliers using the $1.5\times$IQR rule.

(c) Which statistics best describe "typical" prices?

Example 2.28: AP Statistics Exam Scores (2024)

A sample of 12 students' AP Statistics exam scores (out of 100):

```
72, 75, 78, 79, 81, 82, 84, 85, 87, 88, 90, 91
```

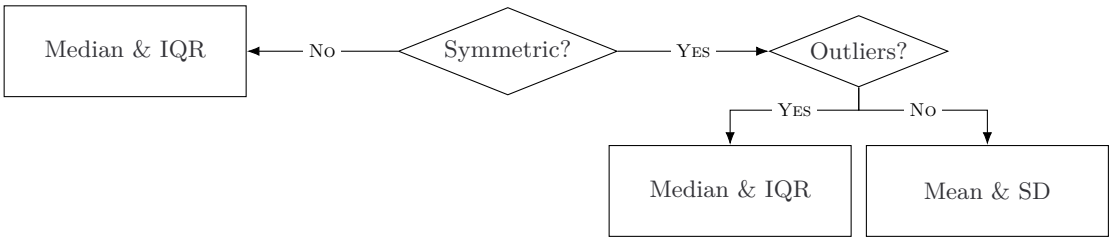
(a) Calculate the mean and median.

(b) What does the relationship between mean and median suggest about the shape?

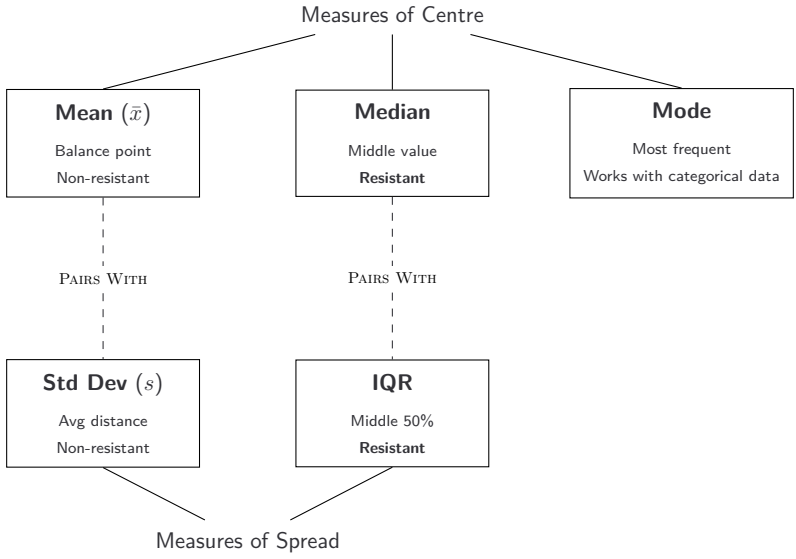
(c) Calculate the standard deviation.

(d) Which summary statistics are most appropriate?

Decision Guide: Choosing Summary Statistics



Summary: Properties of Statistical Measures



Chapter Summary

Measures of Centre

- **Mean ($\bar{x}$)** — Balance point; not resistant
- **Median** — Middle value; resistant
- **Mode** — Most frequent; works for categorical

Measures of Spread

- **IQR** — $Q_3 - Q_1$; resistant
- **Variance (s^2)** — Avg squared deviation
- **Std Dev (s)** — $\sqrt{s^2}$; same units as data

Graphical Summaries

- **Five-Number Summary:** Min, Q_1, Median, Q_3, Max
- **Standard Boxplot:** Whiskers to min/max
- **Modified Boxplot:** Outliers shown separately
- $1.5\times$ **IQR Rule:** Identifies outliers

Choosing Statistics

- Skewed/Outliers $\Rightarrow$ Median & IQR
- Otherwise $\Rightarrow$ Mean & Std Dev

Exercise: Mean vs. Median

Data: $\{15, 22, 29, 31, 35, 42, 88\}$

(a) Calculate the mean.

(b) Calculate the median.

(c) Which measure better represents the "typical" value? Why?

Exercise: Outlier Detection

Data:

```
12, 14, 15, 16, 17, 18, 19, 20, 45
```

(a) Find Q_1, Q_3, and IQR.

(b) Calculate the fences.

(c) Are there any outliers? If so, identify them.

Exercise: Interpreting a Boxplot

A company reports employee salaries with this five-number summary:

Min = \$35,000 Q_1 = \$48,000 Median = \$55,000 Q_3 = \$72,000 Max = \$250,000

(a) What is the IQR?

(b) Is the distribution symmetric, left-skewed, or right-skewed? Explain.

(c) Should the company report the mean or median salary? Why?