CHAPTER 2

Describing Distributions with Numbers

✓ Goals

- 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: range, IQR, variance, and standard deviation
- Construct and interpret standard and modified boxplots
- Identify potential outliers using the $1.5 \times \text{IQR}$ rule
- Choose appropriate summary statistics based on distribution shape

2.1. Why Numerical Summaries

In Chapter 1, we learned to *visualise* distributions using histograms, stemplots, and other graphs. These pictures give us an intuitive feel for the shape of our data, but pictures alone are not always enough.

Histograms have a lot of detail and while this is valuable, sometimes we need a quick summary that captures the essence of the distribution in just a few numbers. This is where **numerical summaries** come in.

A **numerical summary** reduces an entire distribution to a few key numbers. These numbers answer two fundamental questions about our data:

1. **Where is the centre?** (Measures of central tendency)
2. **How spread out are the values?** (Measures of variability)

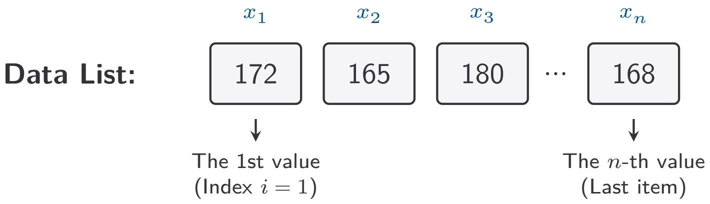
2.2. Measuring Centre: The Mean

We begin with the most familiar measure of centre: the **arithmetic mean**, commonly called the "average." You have likely calculated means since elementary school: add up all the values and divide by how many there are. But before we can write this precisely, we need to establish some notation.

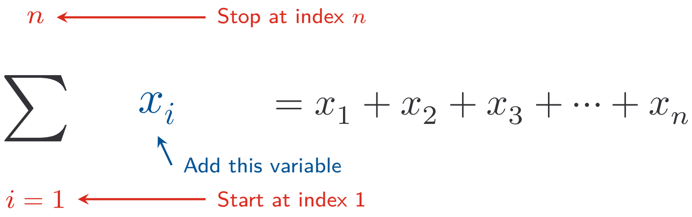
Notation: Indices and Summation

When working with data, we often have many observations of the same variable. Rather than writing "the first person's height, the second person's height, the third person's height…" we use a compact notation:

- Let x represent our variable of interest (e.g., height, income, test score).
- Use subscripts $x_1, x_2, \ldots, x_n$ to identify each individual observation: x_1 is the first value, x_2 is the second value, and so on, up to x_n, the n-th (last) value.
- The letter n always denotes the total number of observations.



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



Example 2.1 *Calculating a Sum*
Suppose that our dataset contains the variable `countries_visited` for three students:

	`countries_visited`
1	2
2	5
3	1

Let x denote the `countries_visited` variable. Find

$$\sum_{i=1}^{3} x_i.$$

Example 2.2 *Calculating a Sum with More Data*
Suppose we have the variable `books_read` for five students:

`books_read`
3
7
2
10
5

Let x denote the `books_read` variable. Find the sum

$$\sum_{i=1}^{5} x_i.$$

In Example 2.2, we do not explicitly list an index for each observation.

This is OK because the order of the data is implied by their position in the list. The first value (3) corresponds to x_1, the second value (7) to x_2, and so on. So long as we are consistent, it does not matter how we assign the values to indices.

The index i is a **dummy variable**: it is just a placeholder that tells us which observation to use. The actual letter used does not matter; we could have written $\sum_{j=1}^{n} x_j$ or $\sum_{k=1}^{n} x_k$ and it would mean the same thing.

Definition 2.3 *Mean (Arithmetic Average)*
The **mean** of n observations $x_1, x_2, \ldots, x_n$ is the sum of all values divided by the

count:

The symbol $\bar{x}$ (read “x-bar”) denotes the sample mean. We will use it throughout the course to represent the average of a variable.

The Mean as a Balance Point

The mean has a beautiful physical interpretation: it is the **balance point** of the distribution.

Intuition: Imagine each data point as a weight placed on a number line (like a seesaw or ruler). The mean $\bar{x}$ is the position where you would place a fulcrum to make the system balance perfectly. Points farther from the centre exert more “leverage”. This is why extreme values have such a large influence on the mean.

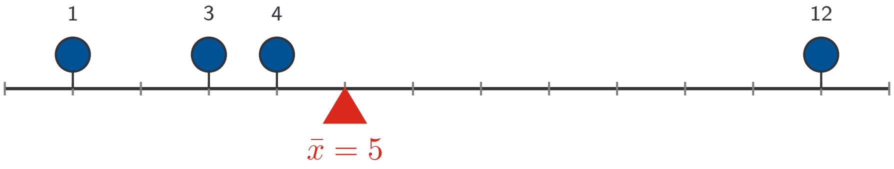


Figure: The mean as a balance point. Data: {1, 3, 4, 12}

Example 2.4 *Calculating the Mean*
Suppose five students report the number of hours they studied for an exam:

$$3 \quad 5 \quad 7 \quad 8 \quad 12$$

. Find the mean number of study hours.

Example 2.5 *The Mean is Non-Resistant*
Consider the ages (in years) of five participants in a local book club:

$$22 \quad 25 \quad 28 \quad 30 \quad 32$$

1. Calculate the mean age of the group.
2. Suppose a 95-year-old member joins the club. Calculate the new mean age.
3. Compare the two means. Does the new mean accurately represent the “typical” age of the book club members? Why or why not?

2.2.1 Robustness of a Measure

Definition 2.6 *Resistant Measure*
A statistical measure is **resistant** (or robust) if it is not sensitive to the influence of a few extreme observations (outliers).

Why does this matter?

- Real data sometimes contains unusual values (typos, measurement errors, or genuinely extreme cases).
- Choosing the wrong summary can give a misleading picture of what is "typical."

Thus, we want to be mindful of the measures we choose based on the characteristics of our data.

2.3. Measuring Centre: The Mode

We have seen that the mean has a significant weakness: it can be pulled dramatically by outliers. This naturally raises the question: is there a measure of centre that is more resistant to extreme values?

The answer is yes: the **mode**. Unlike the mean, which calculates with numerical values, the mode simply asks: "Which value appears most often?" This counting approach makes the mode completely resistant to outliers. If we were to change a datapoint to an extreme value, it would not affect (in general) which value is most frequent.

Definition 2.7 *Mode*
The **mode** is the value that appears **most frequently** in a dataset.

The mode has a unique property: it is the **only measure of centre that works for categorical data**. You cannot calculate the mean of "red, blue, green," but you can identify which colour appears most often.

Here is how we would describe the mode in different types of data:

1. Dataset: $\{2, 3, 3, 5, 7\}$
   Solution: The mode is 3 (appears twice).
2. Dataset: $\{1, 2, 3, 4, 5\}$
   Solution: No mode (no value repeats).
3. Dataset: $\{\text{cat}, \text{dog}, \text{dog}, \text{bird}\}$
   Solution: The mode is "dog" (appears twice).

A distribution can have one mode (unimodal), two modes (bimodal), or more (multimodal). If no value repeats, we say the distribution has no modes.

Example 2.8 *Calculating the Mode*
For each of the following datasets, identify the mode(s):

- $\{4, 7, 7, 7, 9, 12\}$

- $\{2, 4, 6, 8, 10\}$
- $\{1, 3, 3, 4, 5, 6, 6, 9\}$
- For a survey of T-shirt sizes: $\{\text{S}, \text{M}, \text{L}, \text{M}, \text{L}, \text{XL}, \text{L}\}$

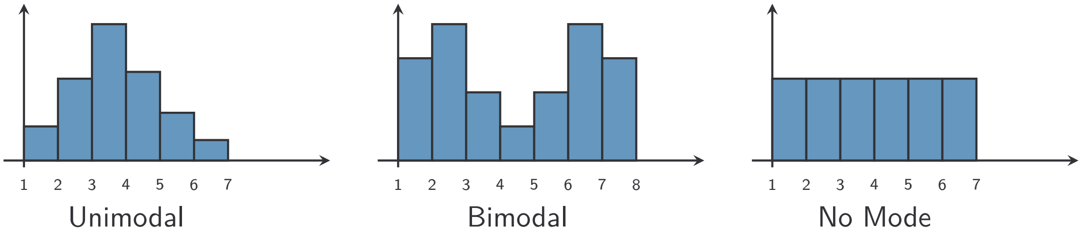


Figure: Three types of modality. Unimodal distributions have one peak; bimodal distributions have two distinct peaks; distributions with no repeating values have no mode.

🔑 **Key point:** The mode is a **resistant** measure.

In the dataset $\{1, 1, 1, 2, 5\}$, the mode is 1. If we change the 5 to 5,000,000, the mode remains 1. The outlier has absolutely no effect on the mode.

The mode answers the question: “What value is most common?” This makes it useful for categorical data but less useful for continuous quantitative data where values rarely repeat exactly. It does not provide us with information about what a ’typical’ value might be for a distribution, but just tells us which value occurs most frequently.

2.4. Measuring Centre: The Median

The mode is completely resistant to outliers, but it has a limitation: it only tells us the most *frequent* value, not the *central* value. For many purposes, we want a measure that locates the “middle” of the distribution while still being resistant to extreme values.

The **median** offers exactly this combination. It finds the central position in the data: the value that splits the distribution in half.

Definition 2.9 *Median*
The **median** is the **middle value** when data are arranged in order. It divides the distribution so that exactly half (50%) of observations fall below it and half fall above.

Unlike the mean, the median cares only about *position*, not *magnitude*. This makes it resistant to outliers.

Finding the Median

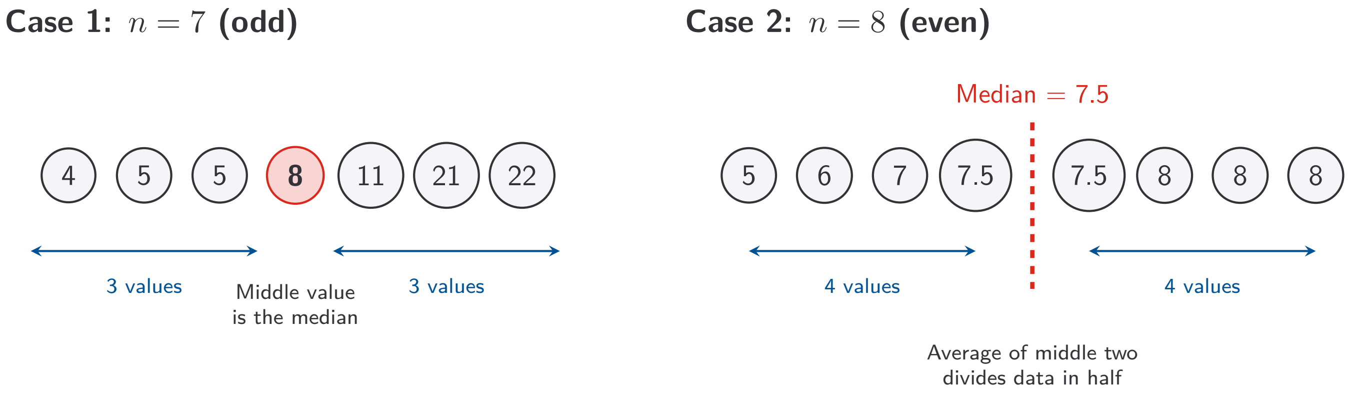
1. **Sort** the data from smallest to largest.
2. **Locate the middle:**
   - If n is odd: The median is the single middle value at position $\frac{n+1}{2}$.
   - If n is even: The median is the average of the two middle values.

Example 2.10 *Median with Odd n*
Find the median of the dataset:

$$4 \quad 5 \quad 5 \quad 8 \quad 11 \quad 21 \quad 22$$

Example 2.11 *Median with Even n*
Find the median of the dataset:

$$5 \quad 6 \quad 7 \quad 7.5 \quad 7.5 \quad 8 \quad 8 \quad 8$$



Intuition: Unlike the mean (which balances weights), the median simply **counts** positions. It finds the midpoint that splits the sorted data into two equal-sized groups, regardless of how extreme the values are.

Exercise 2.12 *Median Calculation*
Find the median of the following datasets:

(a) $\{7, 12, 15, 18, 22\}$

(b) $\{4, 8, 11, 13, 17, 20\}$.

(c) $\{25, 18, 30, 22, 27\}$

(d) $\{5, 7, 7, 8, 10, 12\}$

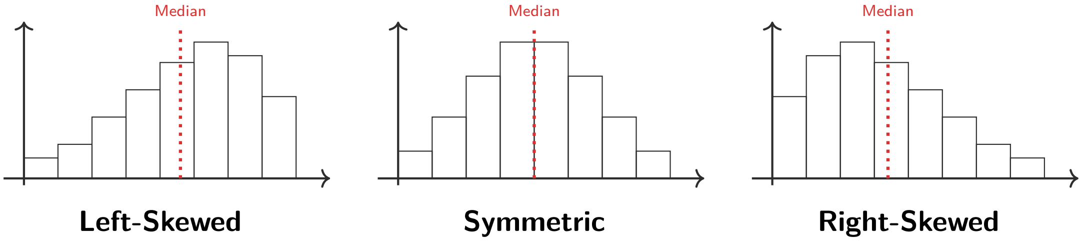


Figure 2.1: The median divides the area under the histogram in half: 50% of the data lie to the left and 50% to the right.

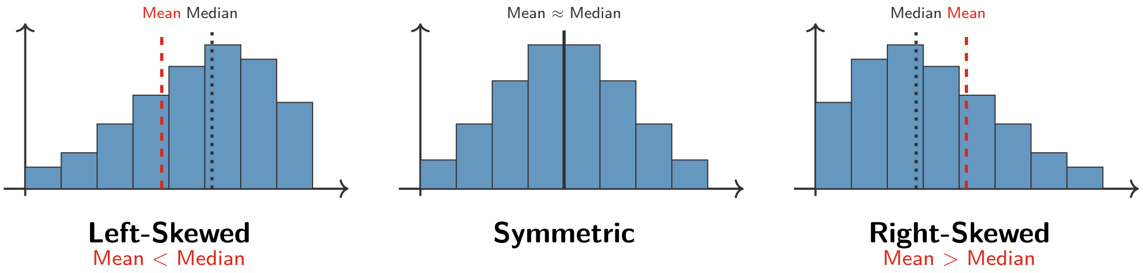
Visualising Shape: Mean vs. Median

For the purposes of our course, most of the distributions we will encounter will satisfy the following rule of thumb.

Mean vs. Median and Skewness

In a skewed distribution, the mean is usually farther out in the long tail than the median. As a practical rule of thumb:

- In a **right-skewed** distribution (long tail to the right): Mean > Median (typically)
- In a **left-skewed** distribution (long tail to the left): Mean < Median (typically)
- In a **symmetric** distribution: Mean ≈ Median

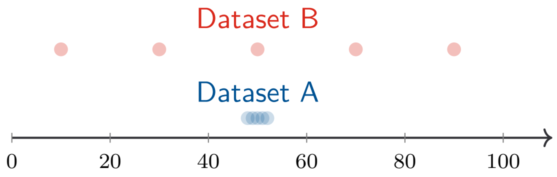


2.5. Measures of Variability

From Centre to Spread: The Missing Piece

We have now explored three measures of centre: the mean, mode, and median. But knowing where a distribution is centred is only half the story. Consider two datasets that both have a mean of 50:

- Dataset A: $\{48, 49, 50, 51, 52\}$
- Dataset B: $\{10, 30, 50, 70, 90\}$



These datasets have identical means (and even medians and modes), yet they describe very different situations. Dataset A has values tightly clustered around 50, while Dataset B has values spread across a wide range. A measure of center on its own would not allow us to distinguish between these distributions.

We now turn to measures that quantify this spread, or **variability**. These measures tell us how far observations typically deviate from the centre.

2.5.1 Quartiles and the Interquartile Range

The first measure of spread we'll examine is based on the idea that the middle 50% of the data often provides a good summary of variability. In other words, if we know how wide the central half of the data is, we can get a sense of how spread out the entire dataset is, as these middle values are not influenced by outliers.

To accomplish this, we need to first define **quartiles**.

Definition 2.13
The **quartiles** divide a sorted dataset into four equal parts:

- The **first quartile** (Q_1) is the value that separates the lowest 25% of the data from the rest.
- The **second quartile** (Q_2) is the median, splitting the data into two equal halves.
- The **third quartile** (Q_3) separates the lowest 75% from the highest 25%.

Quartiles help describe the spread and identify the middle 50% of the data.

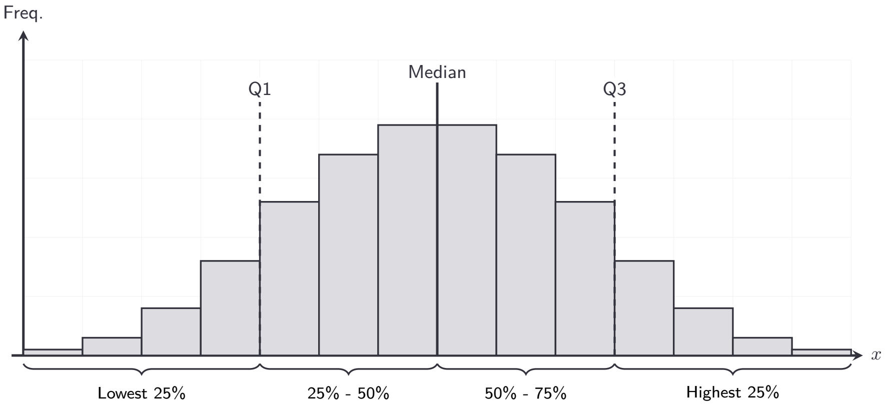


Figure: the length of the right tail is evident in the distance from Q3 to the maximum value relative to distance from Q1 to the minimum value.

Quartiles also help describe skewness: in a skewed distribution, the spacing between quartiles is uneven.

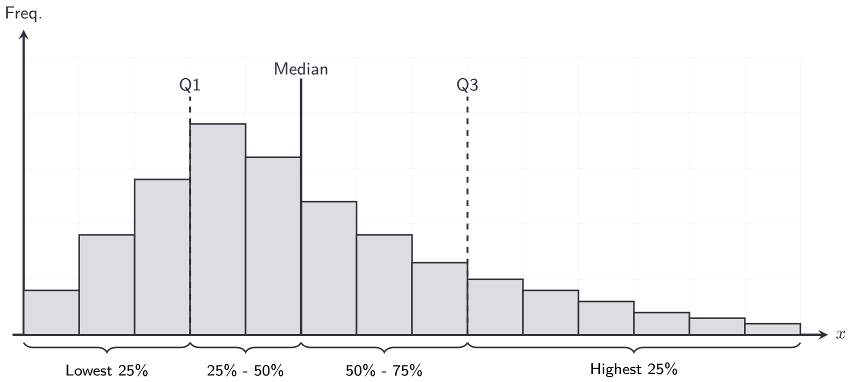


Figure: Uneven spacing between quartiles indicates skewness. Left-skewed distributions have wider spacing on the left; right-skewed distributions have wider spacing on the right.

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.
    - **Lower Half:** The lower $n/2$ values.
    - **Upper Half:** The upper $n/2$ values.
5. Q_1**:** The median of the Lower Half.
6. Q_3**:** The median of the Upper Half.

Example 2.14 *Finding Quartiles: Odd n*
Find the quartiles of the following dataset:

$$4 \quad 5 \quad 5 \quad 8 \quad 11 \quad 21 \quad 22$$

Example 2.15 *Finding Quartiles: Even n*
Find the quartiles of the following dataset:

$$2 \quad 4 \quad 7 \quad 9 \quad 10 \quad 12 \quad 15 \quad 18$$

Definition 2.16 *Interquartile Range (IQR)*
The **interquartile range (IQR)** measures the spread of the middle 50% of the data

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

Example 2.17 *Finding Quartiles and IQR*
Find the quartiles of weekly coffee expenses (in dollars):

$$3 \quad 5 \quad 8 \quad 12 \quad 15 \quad 28 \quad 35$$

Intuition: Think of the IQR as describing where the "typical" values live. By focusing only on the middle half of the data, the IQR completely ignores the minimum and maximum. Thus, the IQR is resistant to outliers.

2.6. The Five-Number Summary and Boxplots

Visualizing Spread: From Numbers to Pictures

So far, we have developed numerical measures to describe spread: quartiles, IQR, and the range. These statistics tell us about variability, but comparing numbers across multiple groups can be cumbersome. The **boxplot** provides a visual way to display all of these measures simultaneously, making patterns and comparisons easier to see at a glance.

The quartiles, together with the minimum and maximum, form the five-number summary. This summary is visualized using a boxplot.

Definition 2.18 *Five-Number Summary*
The **five-number summary** captures the key features of a distribution using exactly five values:

$$\text{Min}, \quad Q_1, \quad \text{Median}, \quad Q_3, \quad \text{Max}$$

These five numbers tell us the range of the data (Min to Max), where the middle 50% lies (Q_1 to Q_3), and where the centre is (Median).

The five number summary is can then be represented visually using the boxplot.

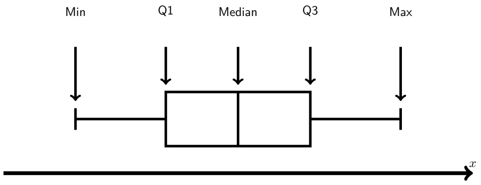


Figure: Anatomy of a boxplot. The box spans from Q_1 to Q_3 (the IQR), with whiskers extending to the minimum and maximum values.

Definition 2.19 *Boxplot*
A **boxplot** (also called a box-and-whisker plot) is a graphical representation of the five-number summary:

- The **box** spans from Q_1 to Q_3, showing the middle 50% of the data.
- A **line inside the box** marks the median.
- **Whiskers** extend from the box to the minimum and maximum values.

Example 2.20 *Creating a Boxplot*
Draw the boxplot for the following dataset:

$$5 \quad 10 \quad 14 \quad 16 \quad 18 \quad 20 \quad 25 \quad 30 \quad 40 \quad 45 \quad 60 \quad 83$$

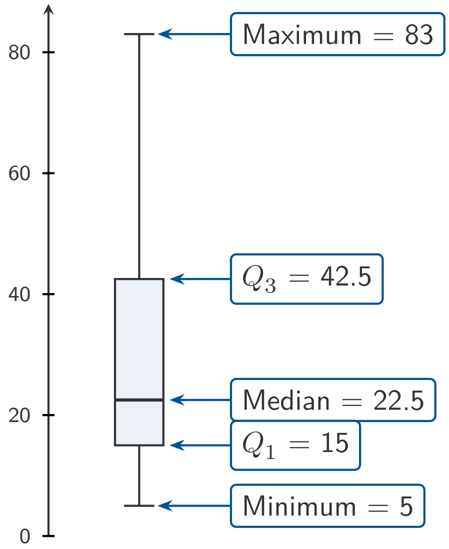
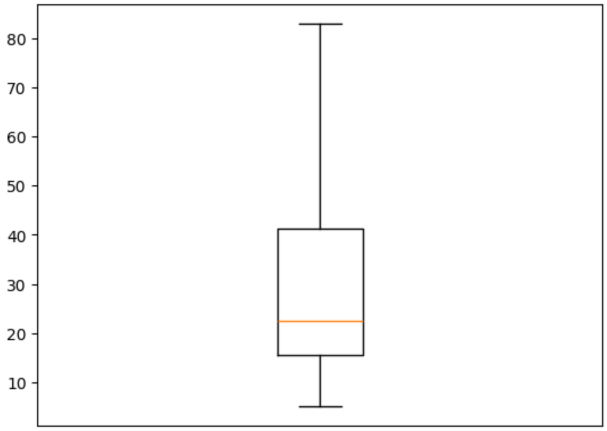


Figure: The boxplot for the dataset above. Note that boxplots can be drawn horizontally or vertically - it does not affect the interpretation.

Intuition: The box shows where most values cluster (the middle 50%), the line inside shows the centre, and the whiskers show how far the extremes reach. A longer whisker on one side suggests skewness in that direction.

Creating a Boxplot in Python

```
<MINTED>
```

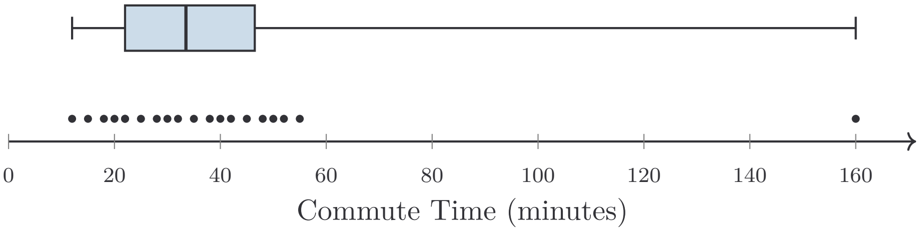


2.6.1 Modified Boxplots and Outliers

Standard boxplots extend their whiskers all the way to the minimum and maximum values. This is simple but has a drawback: if there is an extreme outlier, the whisker stretches far out to reach it, potentially distorting our view of where most of the data lies.

Modified boxplots solve this problem by treating outliers separately. Instead of stretching whiskers to extreme values, we:

1. Define “fences” beyond which values are considered potential outliers.
2. Extend whiskers only to the most extreme values *within* those fences.
3. Plot any outliers as individual points beyond the whiskers.



Definition 2.21 *Potential outliers*
A value is a **potential outlier** if it falls more than 1.5 × IQR beyond the quartiles:

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

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

? *Why* 1.5*?*

The 1.5 × IQR rule is a useful convention popularized by John Tukey in the 1970s. It is not the only way to identify outliers, but it provides a consistent starting point. In different contexts, other thresholds (like 2 × IQR or 3 × IQR) may be more appropriate depending on the data and the analysis goals. However, for our course, we will stick with the 1.5 × IQR rule as the standard for identifying potential outliers *within the context of boxplots*.

Note that we can still identify outliers by inspecting a plot like the stemplot or histogram as we did in chapter 1, just that the modified boxplot provides a systematic numerical method to do so.

Definition 2.22 *Modified Boxplot*
A **modified boxplot** displays outliers as individual points. The whiskers extend only to the most extreme values *within* the fences.

All potential outliers are plotted separately beyond the whiskers (usually as dots or asterisks).

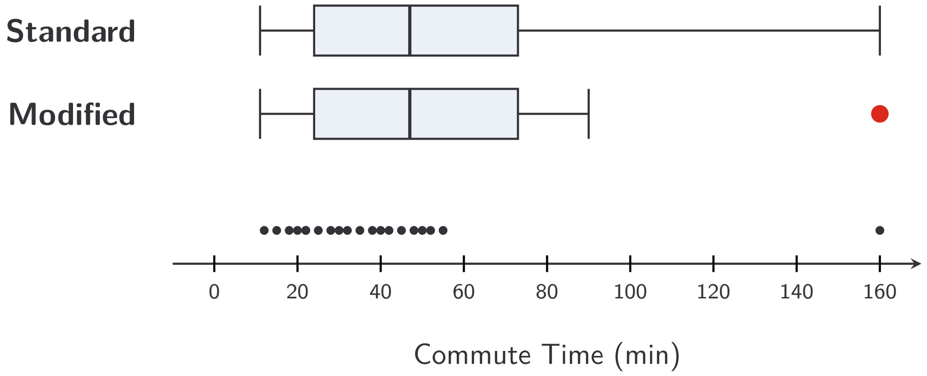
Example 2.23 *Daily Commute Times*
Consider a dataset of daily commute times (in minutes) for 20 employees:

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

Draw the modified boxplot for this data.

Standard vs. Modified Boxplot Comparison

The following diagram illustrates the key difference between a standard boxplot (which extends whiskers to the absolute min/max) and a modified boxplot (which identifies outliers separately).



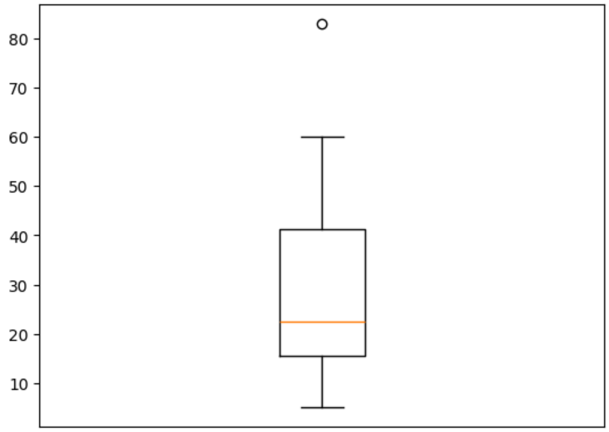
Example 2.24 *Outliers on Both Tails*
Sometimes a distribution has outliers on both the low and high ends. Consider the following dataset:

5 45 48 50 52 55 58 60 62 110

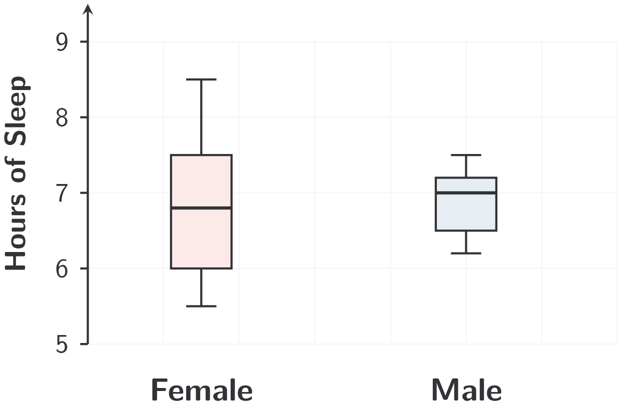
Draw the modified boxplot for this data.

Creating a Boxplot in Python

```
<MINTED>
```



Example 2.25 *Comparison: Who sleeps more?*
Comment on the differences in distribution of sleep hours by gender using the boxplots below.



Answer: The median sleep time (the line inside the box) is similar for both groups, indicating that the typical sleep duration is about the same. However, the boxes differ in height (or width, depending on how we draw it), indicating different variability in sleep habits. The female group has much more variability in sleep duration compared to the men, who have a more consistent sleep pattern.

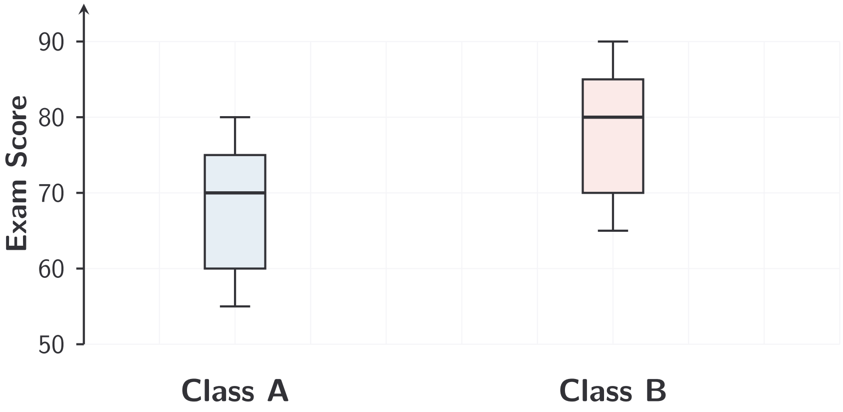
The male distribution appears to be slightly more symmetric, while the female distribution may be skewed having a longer upper whisker (hence indicating some right skewness).

When comparing two distributions using boxplots, consider the following aspects:

- **Centre:** Compare the medians to see which group tends to have higher or lower values.
- **Spread:** Look at the IQRs to assess variability within each group.
- **Outliers:** Note any outliers and consider their potential impact on the analysis.
- **Shape:** Observe the symmetry or skewness indicated by the quartiles (i.e. distance from Q1 to Q2 and Q2 to Q3) and whisker lengths.

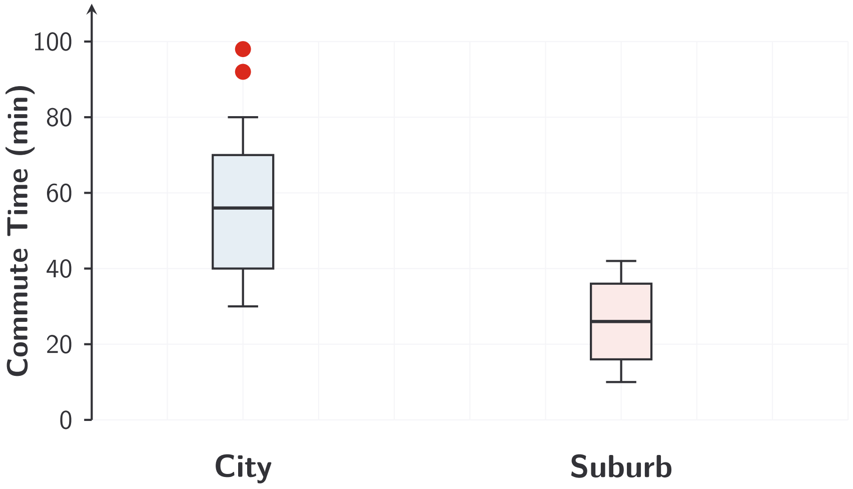
> Skewness might be apparent from the relative lengths of the whiskers and the position of the median within the box but this is less precise than using histograms or density plots.

Example 2.26 *Comparing Exam Scores: Two Classes*
Final exam scores (out of 100) for two statistics classes are shown below.

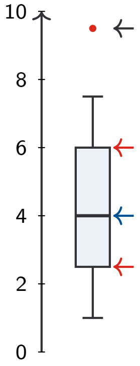


Compare the two distributions.

Example 2.27 *Comparing Commute Times: City vs. Suburb*
Daily commute times (minutes) for workers in two locations are shown below.

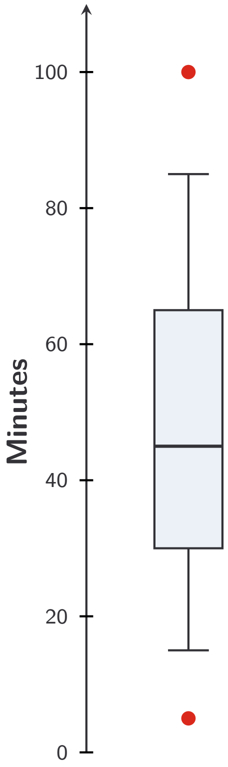


Compare the two distributions.



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

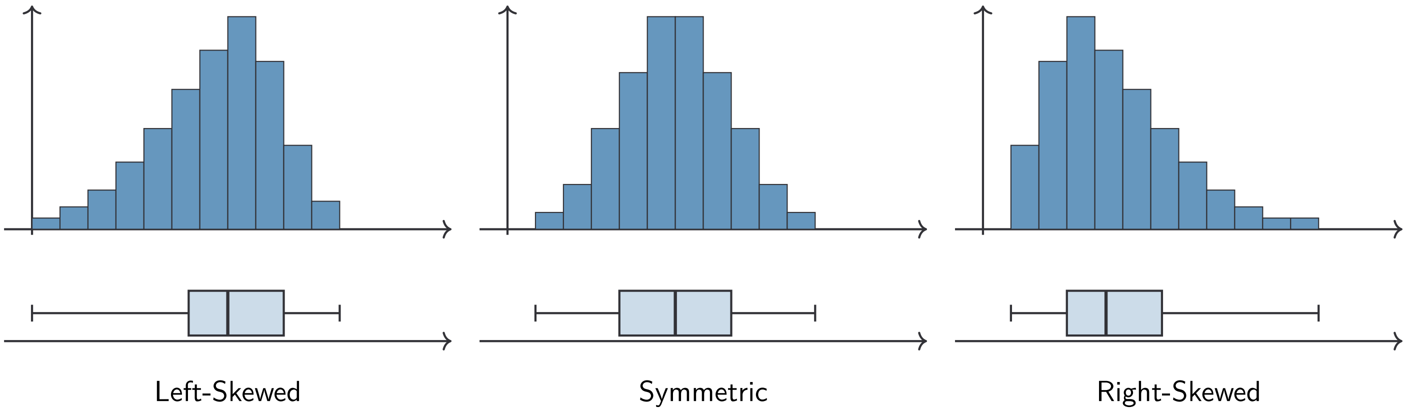
Example 2.28 *Practice Reading a Boxplot*
The boxplot below shows daily study time (minutes) for 50 students.



Identify: (a) the median, (b) the IQR, (c) any outliers, and (d) whether the distribution is approximately symmetric.

Inferring Shape from Boxplots

Boxplots can hint at distribution shape by comparing whisker lengths and the median's position within the box.



Left-Skewed

Symmetric

Right-Skewed

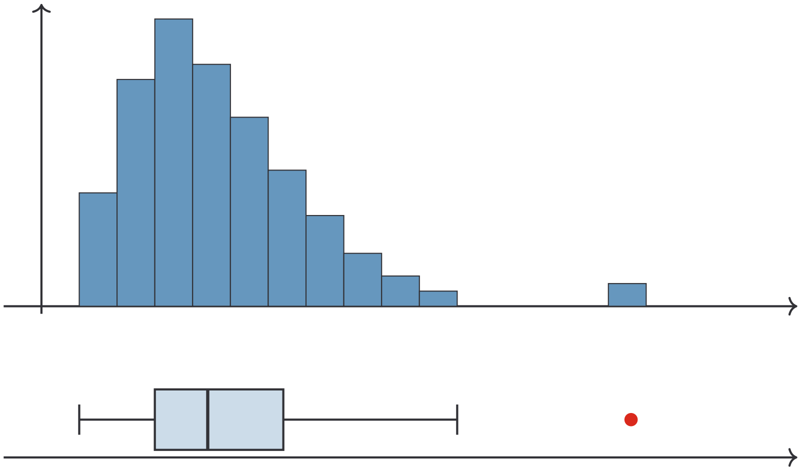


Figure: A right-skewed distribution with a single high outlier.

Limitations of Boxplots

Boxplots summarize centre and spread well, but they can hide important shape details.

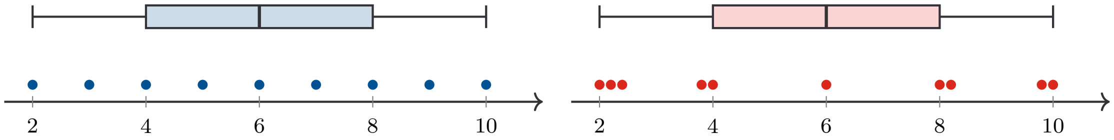
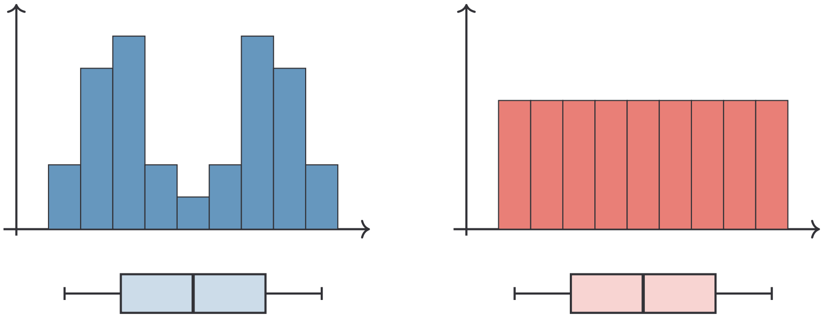


Figure: Two different distributions with identical boxplots.

🔑 **Key point:** Two structurally different distributions can produce identical boxplots.



2.7. Measuring Variability: Variance and Standard Deviation

Beyond Quartiles: Using All the Data

Boxplots and the IQR give us a robust, visual way to compare distributions and detect outliers. However, the IQR only uses two values from the dataset: the quartiles $Q1$ and $Q3$. What if we want a measure that uses *all* the data points?

The variance and standard deviation answer this need. They measure how far observations typically fall from the mean by calculating the average squared distance. This makes them more informative for symmetric distributions, though they share the mean's sensitivity to outliers.

To define variance and standard deviation, we first need to understand *deviations*.

Definition 2.29 *Deviation*
A **deviation** is the signed distance between an observation and the mean:

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

Positive deviations indicate values above the mean; negative deviations indicate values below.

As an example, consider the dataset $\{3, 5, 8\}$. To find the deviation of each value, we first compute the mean, then subtract it from each data point.

Step 1: Calculate the mean

$$\bar{x} = \frac{3+5+8}{3} = \frac{16}{3} \approx 5.33$$

Step 2: Calculate deviations

x_i	$x_i - \bar{x}$
3	$3 - 5.33 = -2.33$
5	$5 - 5.33 = -0.33$
8	$8 - 5.33 = 2.67$

👁 **Intuition:** To measure spread, we want to know: “On average, how far are the data points from the mean?” Deviations help us capture this distance.

But there is a problem: if we simply average the deviations, the positives and negatives cancel out (they always sum to zero). To avoid this, we square each deviation before averaging.

There are mathematical reasons for squaring rather than taking absolute values, which we won't cover here.

Example 2.30 *Computing Deviations*
For the dataset $\{2, 4, 5, 7, 9\}$ with mean $\bar{x} = 5.4$, compute the deviation of each data point from the mean.

Definition 2.31 *Variance*
The **variance** measures the average squared deviation from the mean

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

Example 2.32 *Computing the Variance*
Find the variance for the dataset

$$2 \quad 4 \quad 5 \quad 7 \quad 9$$

Example 2.33 *Variance Practice*
Find the variance s^2 for the dataset:

$$3 \quad 6 \quad 7 \quad 8 \quad 13$$

Definition 2.34 *Standard Deviation*
The **standard deviation** is the square root of the variance

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

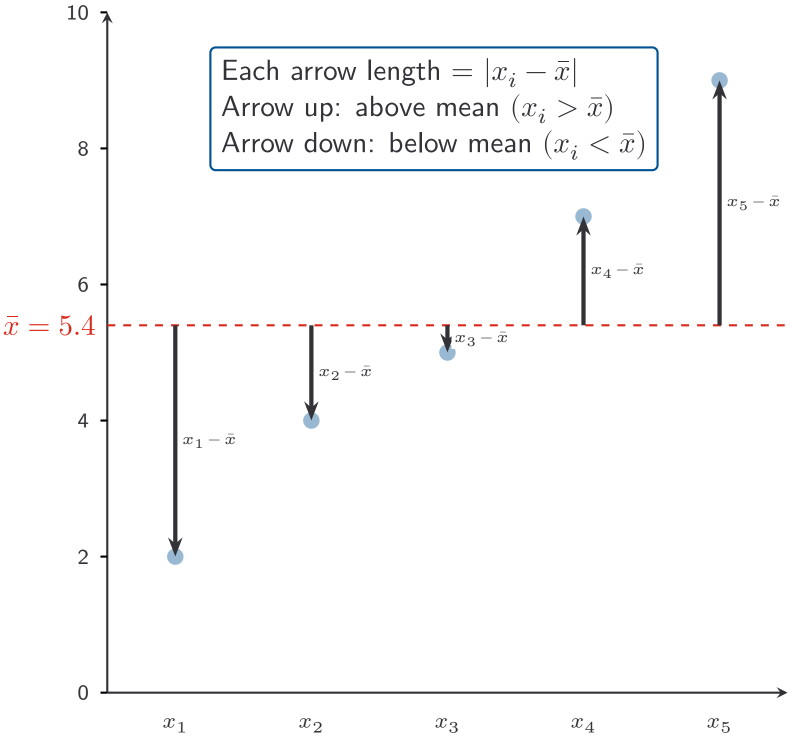
Taking the square root of the variance returns us to the original units of measurement, making the standard deviation directly interpretable.

Intuition: If heights are measured in centimetres, the variance is measured in square centimetres, which is not a meaningful unit for describing height. Taking the square root gives the standard deviation, which is again measured in centimetres and can be interpreted as the typical distance from the mean.

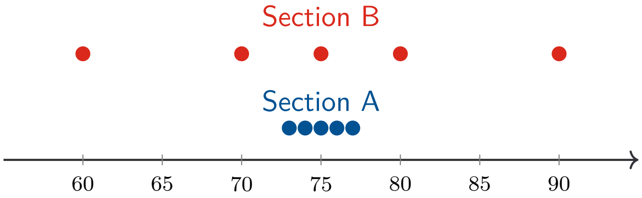
For example, a standard deviation of 5 cm means that heights typically differ from the average by about 5 cm. Whether a standard deviation is considered “large” or “small” depends on the context and the mean, but in practice we usually judge it by comparing it with the standard deviation of another distribution.

Example 2.35 *Computing the Standard Deviation*
Find the standard deviation for the dataset:

$$2 \quad 4 \quad 5 \quad 7 \quad 9$$



Example 2.36 *Standard Deviation with Two Groups*
Two sections of the same statistics course both have a mean exam score of 75. Section A has scores: $\{73, 74, 75, 76, 77\}$ and Section B has scores: $\{60, 70, 75, 80, 90\}$. Calculate the standard deviation for each section and interpret the difference.



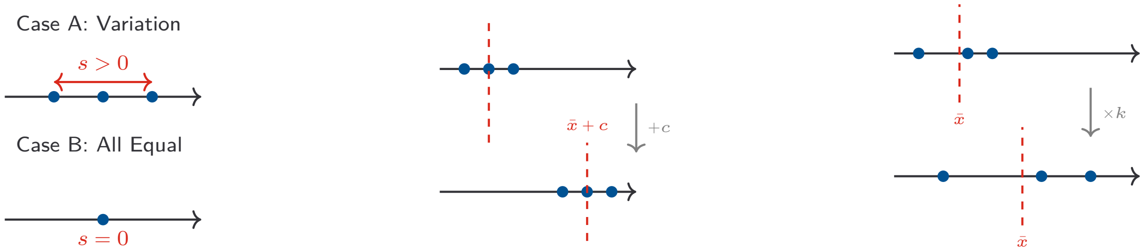
Example 2.37 *Effect of Adding a Constant*
A professor curves all exam scores by adding 10 points to everyone's score. The original scores were: $\{65, 70, 75, 80, 85\}$.

(a) Calculate the standard deviation of the original scores.

(b) Calculate the standard deviation after adding 10 points to each score. How did the standard deviation change?

Key point: Properties of standard deviation:

- $s \geq 0$, with $s = 0$ only when all values are identical
- Adding or subtracting a constant does not change s
- Multiplying all values by k changes s to $|k|s$

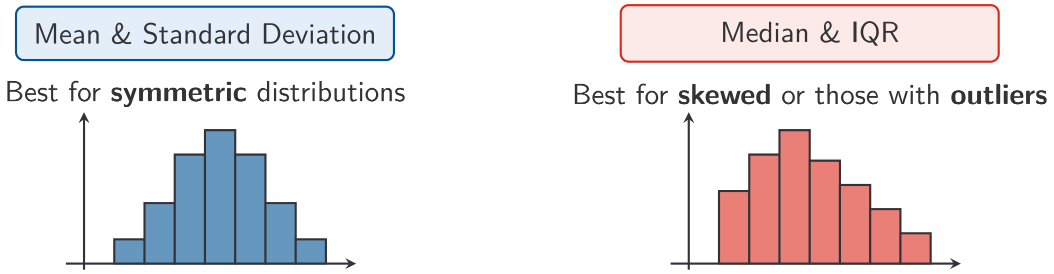


We divide by $n-1$ (Bessel's correction) instead of n to get a better (unbiased) estimate of population variability from sample data. The reasons for it are a bit technical, but it essentially corrects for the fact that using the sample mean underestimates variability.

2.8. Choosing Measures of Centre and Variability

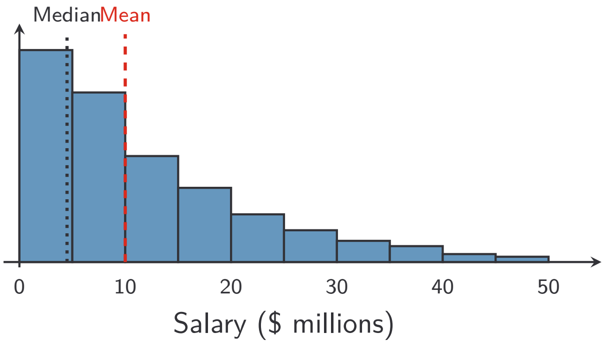
We have now built a complete toolkit: the mean and standard deviation measure centre and spread using all data points, while the median and IQR provide resistant alternatives that ignore extreme values. But how do we decide which pair to use? The choice depends on the **shape** of the distribution and the presence of **outliers**.

These two pairs of summary statistics serve different purposes:


Mean & Standard Deviation
Best for symmetric distributions
Median & IQR
Best for skewed or those with outliers

The following examples illustrate how to apply these guidelines using real data.

Example 2.38 *NBA Player Salaries (2024–25)*
The distribution of NBA player salaries for the 2024–25 season ($n = 450$ players) is shown below.

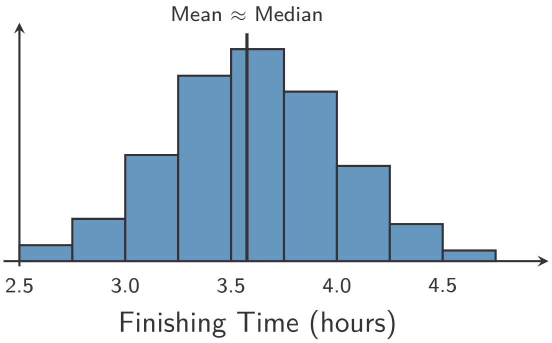


Summary Statistics: Mean = \$10.2M, Median = \$4.5M, SD = \$11.8M, IQR = \$12.1M

(a) Is this distribution symmetric or skewed?

(b) Which measures of centre and spread should be reported?

Example 2.39 *Boston Marathon Finishing Times (2024)*
Finishing times (hours) for 500 randomly sampled runners from the 2024 Boston Marathon are summarized below.

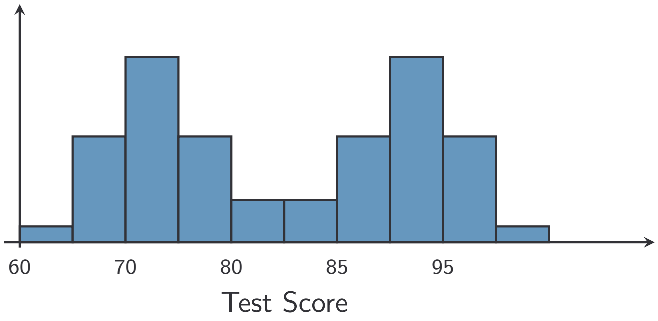


Summary Statistics: Mean = 3.72 hrs, Median = 3.68 hrs, SD = 0.48 hrs, IQR = 0.65 hrs

(a) Is this distribution symmetric or skewed?

(b) Which measures of centre and spread should be reported?

Example 2.40 *Bimodal Symmetric Distribution*
A combined sample of test scores (out of 100) from two sections of a course is shown below.



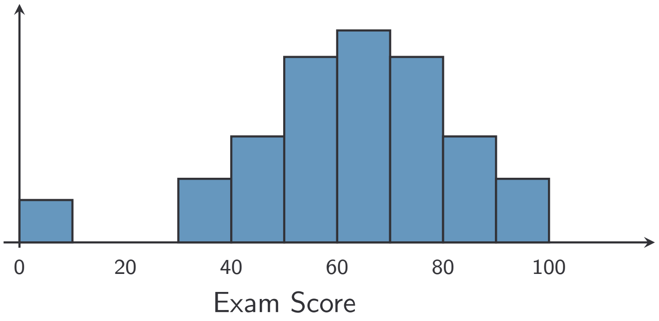
Summary Statistics: Mean = 80.0, Median = 80.0, Mode = 72 and 90 (bimodal), SD = 9.4

(a) Describe the shape of the distribution.

(b) What might explain this pattern?

(c) Which statistics are appropriate?

Example 2.41 *Symmetric Distribution with Outliers*
A class of students took an exam, and their scores (out of 100) are shown below.



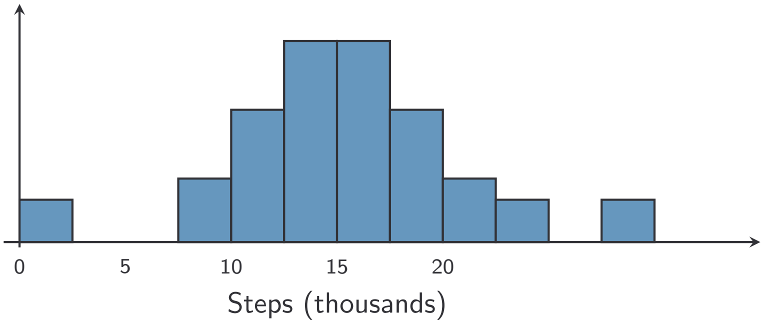
Summary Statistics: Mean = 81.0, Median = 86.0, SD = 19.8, IQR = 16.0

(a) Describe the shape of the distribution.

(b) Identify the outlier.

(c) Which statistics should we report in summarizing the data?

Example 2.42 *Symmetric Distribution with Outliers on Both Sides*
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?

Key point: Even when a distribution appears symmetric, the presence of outliers means we should use resistant measures (median and IQR) rather than non-resistant measures (mean and SD).

Example 2.43 *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 × IQR rule.

(c) Which statistics best describe “typical” home prices?

Example 2.44 *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 for this data?

Chapter Summary

Measures of Centre

- **Mean ($\bar{x}$):** Arithmetic average; balance point of the distribution. **Not resistant** to outliers.
- **Median (M):** Middle value of sorted data. **Resistant** to outliers.
- **Mode:** Most frequent value. Only measure usable for categorical data. **Resistant** to outliers.

Measures of Spread

- **Range:** Max − Min. Simple but **not resistant**.
- **IQR:** $Q_3 - Q_1$; spread of middle 50%. **Resistant** to outliers.
- **Variance (s^2):** Average squared deviation from mean.
- **Standard Deviation (s):** Square root of variance; typical distance from mean. **Not resistant**.

Graphical Summaries

- **Five-Number Summary:** Min, Q_1, Median, Q_3, Max
- **Standard Boxplot:** Whiskers extend to absolute min/max
- **Modified Boxplot:** Whiskers extend to most extreme non-outlier values; outliers plotted individually
- 1.5× **IQR Rule:** Values beyond $Q_1 - 1.5 \times \text{IQR}$ or $Q_3 + 1.5 \times \text{IQR}$ are potential outliers

Choosing Summary Statistics

- **Symmetric distributions:** Use Mean and Standard Deviation
- **Skewed distributions or outliers:** Use Median and IQR
- **Mean > Median:** Typically indicates right-skewed distribution
- **Mean < Median:** Typically indicates left-skewed distribution

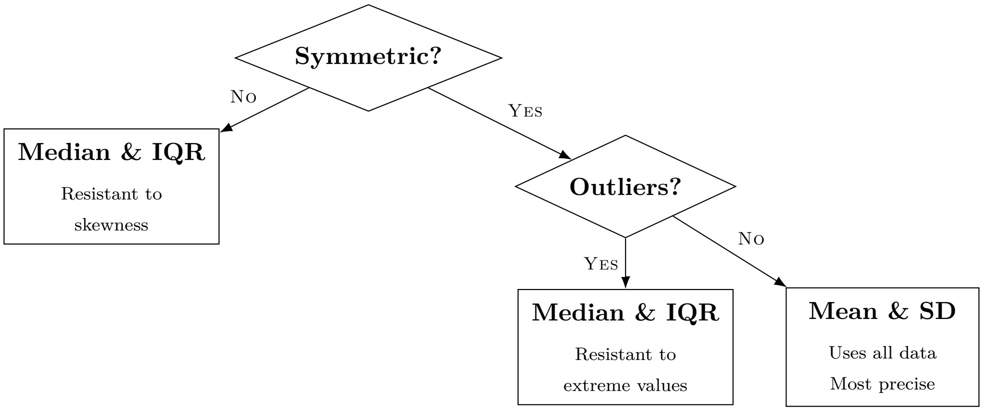


Figure: Decision flowchart for choosing appropriate summary statistics based on distribution shape and presence of outliers.

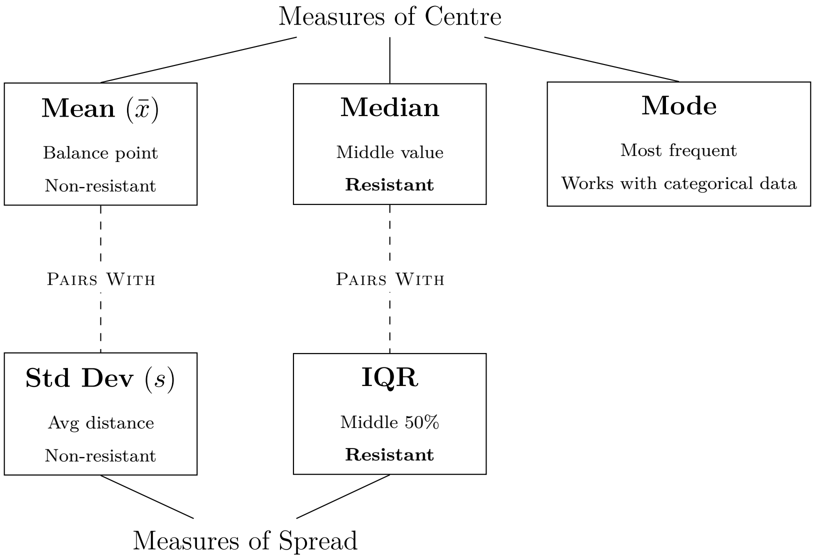


Figure: Decision flowchart for choosing appropriate summary statistics based on distribution shape and presence of outliers. Figure: Comparison of measures. The mean pairs with standard deviation for symmetric data without outliers. The median pairs with IQR for skewed data or data with outliers.

2.10. Practice Exercises

Exercise 2.45 *Mean and Median Comparison*
For the dataset

$$15 \quad 22 \quad 29 \quad 31 \quad 35 \quad 42 \quad 88$$

(a) Compute the mean and median.

(b) Which represents the "typical" value better? Explain.

Exercise 2.46 *Mode Identification*
For each dataset, identify the mode(s) or state that there is no mode:

(a) Data:

$$3 \quad 5 \quad 5 \quad 7 \quad 8 \quad 8 \quad 8 \quad 10$$

(b) Data:

$$12 \quad 15 \quad 18 \quad 21 \quad 24$$

(c) Data:

$$4 \quad 4 \quad 6 \quad 6 \quad 8 \quad 8$$

(d) Survey responses:

Agree Disagree Agree Neutral Agree

Exercise 2.47 *Range and Five-Number Summary*
For the dataset $\{10, 15, 20, 25, 30, 35\}$:

(a) Compute the five-number summary.

(b) Compute the standard deviation.

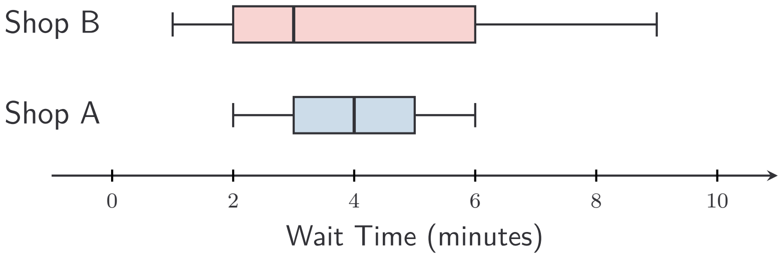
Exercise 2.48
Consider the dataset: $\{12, 14, 15, 16, 17, 18, 19, 20, 45\}$.

(a) Calculate the IQR and the upper/lower fences.

(b) Identify any outliers.

(c) Sketch the modified boxplot for this data.

Exercise 2.49 *Visual Interpretation of Boxplots*
The boxplots below show the distribution of wait times (in minutes) for two different coffee shops, A and B.



(a) Which shop has the lower median wait time?

(b) Which shop has more consistent wait times (lower variability)? Explain how you know.

(c) Describe the shape of the distribution for Shop B based on the boxplot.

Exercise 2.50 *Standard Deviation Calculation*

Calculate the standard deviation for the dataset: $\{8, 10, 10, 12, 15\}$.

Exercise 2.51 *Data Transformations*
A university class has a mean test score of 70 and a standard deviation of 10.

(a) The professor decides to curve the exam by adding 5 points to every student's score. What are the new mean and standard deviation?

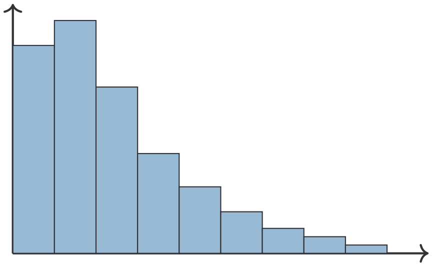
(b) Instead of adding points, the professor decides to grade the exam out of 200 instead of 100, effectively multiplying every score by 2. What are the new mean and standard deviation?

Exercise 2.52 *Critique*
A company reports: “Average salary is $95k, standard deviation $45k.”

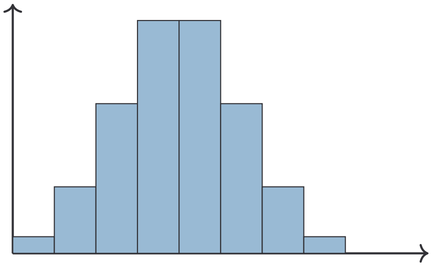
(a) If the CEO earns $2M, how does this affect your interpretation?

(b) What statistic would be more informative?

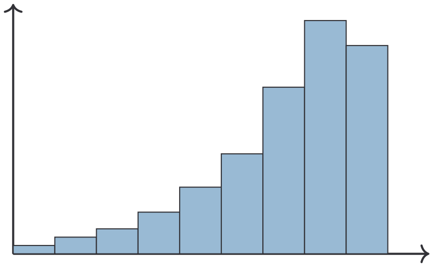
Exercise 2.53 *Visualizing Mean vs. Median*
Below are histograms for three different distributions of household incomes in three different towns.



Town A



Town B

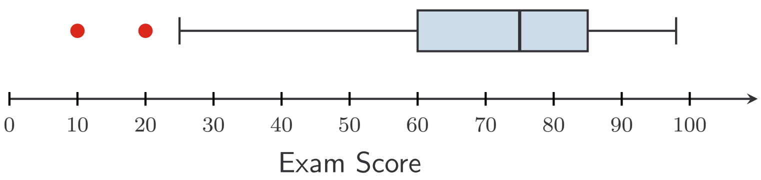


Town C

Below are three pairs of summary statistics calculated from these towns (in thousands of dollars). Match each town to its correct pair of statistics, and explain your reasoning based on the shape of the distribution.

- Pair 1: Mean = \$65k, Median = \$85k
- Pair 2: Mean = \$120k, Median = \$80k
- Pair 3: Mean = \$75k, Median = \$75k

Exercise 2.54 *Reverse Engineering a Boxplot*
Consider the modified boxplot below, which represents scores on a difficult physics exam.



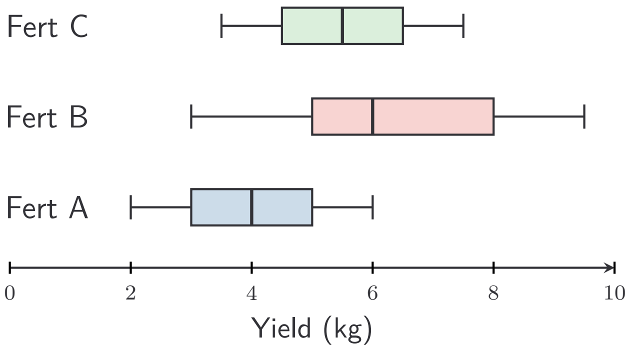
(a) Estimate the five-number summary from the plot.

(b) Calculate the IQR based on your estimates.

(c) Calculate the lower fence used to identify outliers.

(d) Explain numerically why the scores 10 and 20 are classified as outliers, while the score 25 is not.

Exercise 2.55 *Comparing Multiple Groups*
An agricultural researcher tested three different fertilizers (A, B, and C) on tomato plants and recorded the yield (in kg) per plant. The results are summarized in the boxplots below.



(a) Which fertilizer produced the highest typical yield?

(b) Which fertilizer produced the most consistent results (lowest variability)?

(c) Describe the shape of the distribution for Fertilizer B.

(d) If you wanted to guarantee a yield of at least 4 kg per plant, which fertilizer would you choose and why?

Exercise 2.56 *Reverse Engineering Quartiles from Outlier Rules*

A dataset has a median of 60. You are told that the value **112** is the *smallest possible integer* that classifies as a high outlier according to the 1.5 × IQR rule (i.e., $x >$ Upper Fence).

Using this information, determine the theoretical maximum possible value for the third quartile (Q_3).

Hint: Express the upper fence in terms of Q_3 *and* Q_1*, and remember the constraint* $Q_1 \leq \textit{Median} \leq Q_3$.

Additional practice: Odd problems in the text, 2.15 - 2.53 (except 2.35)