

Chapter 7: Descriptive Statistics

Univariate Data

What are we trying to do?

A **univariate data set** records one numerical variable for each observation.

Examples:

- ▶ exam scores for a class,
- ▶ heights of students,
- ▶ waiting times,
- ▶ monthly temperatures.

What are we trying to do?

A **univariate data set** records one numerical variable for each observation.

Examples:

- ▶ exam scores for a class,
- ▶ heights of students,
- ▶ waiting times,
- ▶ monthly temperatures.

Our basic job is to answer three questions:

- ① **What does the distribution look like?**
- ② **Where is the center?**
- ③ **How spread out is it?**

Chapter roadmap

By the end of this chapter, students should be able to:

- ① build and interpret frequency tables and histograms;
- ② recognize skewness and modality;
- ③ compute and interpret mean and median;
- ④ compute range, quartiles, and IQR;
- ⑤ compute variance and standard deviation;
- ⑥ use the Bell Curve Rule and Range Rule;
- ⑦ construct and interpret box plots;
- ⑧ estimate statistics from grouped data.

Most mistakes in this chapter are bookkeeping mistakes, not deep mathematics.

Skill 1: From raw data to a histogram

A **histogram** groups numerical data into equal-width intervals called **bins**.

Skill 1: From raw data to a histogram

A **histogram** groups numerical data into equal-width intervals called **bins**.

A reliable procedure:

- ① Find the smallest and largest values.
- ② Choose reasonable equal-width bins.
- ③ Count how many data values fall in each bin.
- ④ Convert counts to relative frequencies if requested.
- ⑤ Draw one bar over each bin.

Skill 1: From raw data to a histogram

A **histogram** groups numerical data into equal-width intervals called **bins**.

A reliable procedure:

- ① Find the smallest and largest values.
- ② Choose reasonable equal-width bins.
- ③ Count how many data values fall in each bin.
- ④ Convert counts to relative frequencies if requested.
- ⑤ Draw one bar over each bin.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{total number of observations}}.$$

Bin notation: read the endpoints carefully

Suppose a bin is

$(60, 70]$.

Bin notation: read the endpoints carefully

Suppose a bin is

$$(60, 70].$$

This means:

$$60 < x \leq 70.$$

Bin notation: read the endpoints carefully

Suppose a bin is

$$(60, 70].$$

This means:

$$60 < x \leq 70.$$

So:

- ▶ 60 is **not** in the bin,
- ▶ 61 is in the bin,
- ▶ 70 is in the bin.

Bin notation: read the endpoints carefully

Suppose a bin is

$$(60, 70].$$

This means:

$$60 < x \leq 70.$$

So:

- ▶ 60 is **not** in the bin,
- ▶ 61 is in the bin,
- ▶ 70 is in the bin.

A value on a boundary must be counted **exactly once**. Humans have invented many ways to lose points; double-counting a boundary is one of the cheaper ones.

Example 1A: Build a frequency table

Twenty quiz scores are:

52, 55, 57, 61, 63, 63, 68, 71, 72, 74,
74, 76, 79, 81, 82, 84, 86, 88, 91, 96.

Use the bins

$(50, 60]$, $(60, 70]$, $(70, 80]$, $(80, 90]$, $(90, 100]$.

Example 1A: Build a frequency table

Twenty quiz scores are:

52, 55, 57, 61, 63, 63, 68, 71, 72, 74,
74, 76, 79, 81, 82, 84, 86, 88, 91, 96.

Use the bins

$(50, 60]$, $(60, 70]$, $(70, 80]$, $(80, 90]$, $(90, 100]$.

Start by tallying:

Bin	Frequency
$(50, 60]$	_____
$(60, 70]$	_____
$(70, 80]$	_____
$(80, 90]$	_____
$(90, 100]$	_____

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

$(50, 60] : \quad 52, 55, 57 \quad \Rightarrow \quad 3$

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

$(50, 60] : 52, 55, 57 \Rightarrow 3$

$(60, 70] : 61, 63, 63, 68 \Rightarrow 4$

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

$$(50, 60] : \quad 52, 55, 57 \quad \Rightarrow \quad 3$$

$$(60, 70] : \quad 61, 63, 63, 68 \quad \Rightarrow \quad 4$$

$$(70, 80] : \quad 71, 72, 74, 74, 76, 79 \quad \Rightarrow \quad 6$$

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

$$(50, 60] : \quad 52, 55, 57 \quad \Rightarrow \quad 3$$

$$(60, 70] : \quad 61, 63, 63, 68 \quad \Rightarrow \quad 4$$

$$(70, 80] : \quad 71, 72, 74, 74, 76, 79 \quad \Rightarrow \quad 6$$

$$(80, 90] : \quad 81, 82, 84, 86, 88 \quad \Rightarrow \quad 5$$

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

$$(50, 60] : \quad 52, 55, 57 \quad \Rightarrow \quad 3$$

$$(60, 70] : \quad 61, 63, 63, 68 \quad \Rightarrow \quad 4$$

$$(70, 80] : \quad 71, 72, 74, 74, 76, 79 \quad \Rightarrow \quad 6$$

$$(80, 90] : \quad 81, 82, 84, 86, 88 \quad \Rightarrow \quad 5$$

$$(90, 100] : \quad 91, 96 \quad \Rightarrow \quad 2$$

Example 1A: Solution, one bin at a time

52, 55, 57, 61, 63, 63, 68, 71, 72, 74, 74, 76, 79, 81, 82, 84, 86, 88, 91, 96

$$(50, 60] : \quad 52, 55, 57 \quad \Rightarrow \quad 3$$

$$(60, 70] : \quad 61, 63, 63, 68 \quad \Rightarrow \quad 4$$

$$(70, 80] : \quad 71, 72, 74, 74, 76, 79 \quad \Rightarrow \quad 6$$

$$(80, 90] : \quad 81, 82, 84, 86, 88 \quad \Rightarrow \quad 5$$

$$(90, 100] : \quad 91, 96 \quad \Rightarrow \quad 2$$

Check:

$$3 + 4 + 6 + 5 + 2 = 20.$$

Example 1A: Relative frequencies

Since $n = 20$,

$$\text{relative frequency} = \frac{\text{frequency}}{20}.$$

Example 1A: Relative frequencies

Since $n = 20$,

$$\text{relative frequency} = \frac{\text{frequency}}{20}.$$

Bin	Freq.	Rel. Freq.
(50, 60]	3	$3/20 = 0.15$
(60, 70]	4	$4/20 = 0.20$
(70, 80]	6	$6/20 = 0.30$
(80, 90]	5	$5/20 = 0.25$
(90, 100]	2	$2/20 = 0.10$

Example 1A: Relative frequencies

Since $n = 20$,

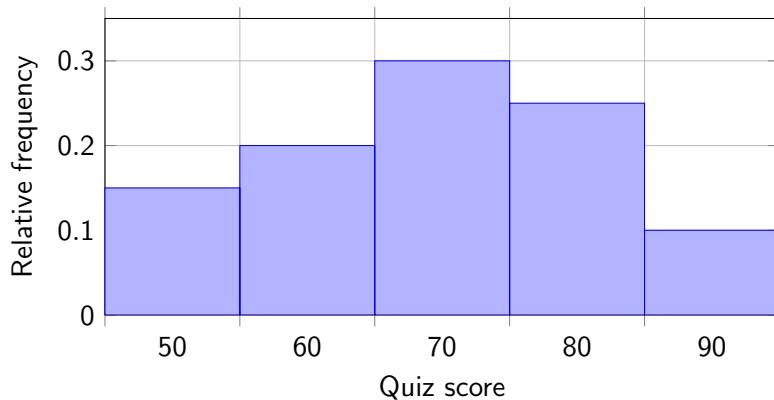
$$\text{relative frequency} = \frac{\text{frequency}}{20}.$$

Bin	Freq.	Rel. Freq.
(50, 60]	3	$3/20 = 0.15$
(60, 70]	4	$4/20 = 0.20$
(70, 80]	6	$6/20 = 0.30$
(80, 90]	5	$5/20 = 0.25$
(90, 100]	2	$2/20 = 0.10$

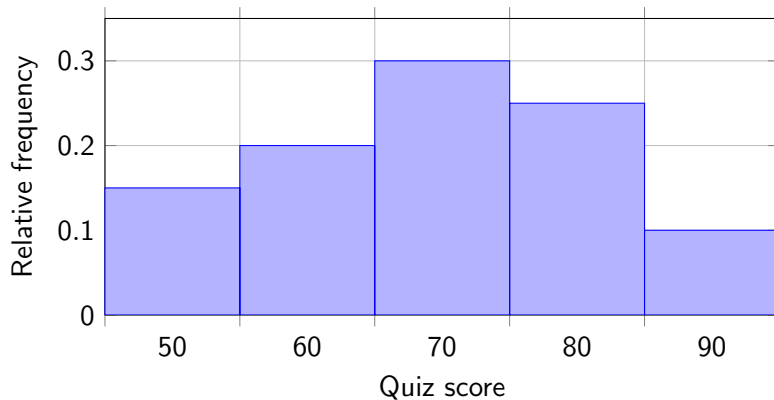
Check:

$$0.15 + 0.20 + 0.30 + 0.25 + 0.10 = 1.$$

Example 1A: Histogram



Example 1A: Histogram



The most common bin is

$(70, 80]$.

Example 1B: Frequency or relative frequency?

Two classes took the same exam.

	60–69	70–79	80–89	90–100
Class A ($n = 20$)	2	6	8	4
Class B ($n = 40$)	8	8	16	8

Which class had the larger *proportion* of students scoring 70–79?

Example 1B: Frequency or relative frequency?

Two classes took the same exam.

	60–69	70–79	80–89	90–100
Class A ($n = 20$)	2	6	8	4
Class B ($n = 40$)	8	8	16	8

Which class had the larger *proportion* of students scoring 70–79?

Class A:

$$\frac{6}{20} = 0.30 = 30\%.$$

Example 1B: Frequency or relative frequency?

Two classes took the same exam.

	60–69	70–79	80–89	90–100
Class A ($n = 20$)	2	6	8	4
Class B ($n = 40$)	8	8	16	8

Which class had the larger *proportion* of students scoring 70–79?

Class A:

$$\frac{6}{20} = 0.30 = 30\%.$$

Class B:

$$\frac{8}{40} = 0.20 = 20\%.$$

Example 1B: Frequency or relative frequency?

Two classes took the same exam.

	60–69	70–79	80–89	90–100
Class A ($n = 20$)	2	6	8	4
Class B ($n = 40$)	8	8	16	8

Which class had the larger *proportion* of students scoring 70–79?

Class A:

$$\frac{6}{20} = 0.30 = 30\%.$$

Class B:

$$\frac{8}{40} = 0.20 = 20\%.$$

So Class A has the larger **relative frequency**, even though Class B has the larger raw count.

Frequency vs. relative frequency

A **frequency histogram** uses counts on the vertical axis.

A **relative-frequency histogram** uses proportions or percentages.

Frequency vs. relative frequency

A **frequency histogram** uses counts on the vertical axis.

A **relative-frequency histogram** uses proportions or percentages.

For one data set, the two histograms have the **same shape**; only the vertical scale changes.

Frequency vs. relative frequency

A **frequency histogram** uses counts on the vertical axis.

A **relative-frequency histogram** uses proportions or percentages.

For one data set, the two histograms have the **same shape**; only the vertical scale changes.

Relative frequency is especially useful when comparing groups of **different sizes**.

Skill 2: Describe the shape

Three words do a lot of work:

Mode

A **mode** is a peak in the distribution.

Skill 2: Describe the shape

Three words do a lot of work:

Mode

A **mode** is a peak in the distribution.

Bimodal

A distribution with two clear peaks is **bimodal**.

Skill 2: Describe the shape

Three words do a lot of work:

Mode

A **mode** is a peak in the distribution.

Bimodal

A distribution with two clear peaks is **bimodal**.

Skewness

A distribution is **right-skewed** if the long tail points right, and **left-skewed** if the long tail points left.

Example 2A: Identify the shape

Suppose a histogram has bar heights

2, 7, 3, 2, 8, 4, 1

from left to right.

Example 2A: Identify the shape

Suppose a histogram has bar heights

2, 7, 3, 2, 8, 4, 1

from left to right.

There is a peak near the second bin.

Example 2A: Identify the shape

Suppose a histogram has bar heights

2, 7, 3, 2, 8, 4, 1

from left to right.

There is a peak near the second bin.

There is another peak near the fifth bin.

Example 2A: Identify the shape

Suppose a histogram has bar heights

2, 7, 3, 2, 8, 4, 1

from left to right.

There is a peak near the second bin.

There is another peak near the fifth bin.

Therefore the distribution is **bimodal**.

Example 2B: Which way is the skew?

Suppose most household incomes in a sample lie between \$40,000 and \$100,000, but a few observations are \$300,000, \$500,000, and \$900,000.

Example 2B: Which way is the skew?

Suppose most household incomes in a sample lie between \$40,000 and \$100,000, but a few observations are \$300,000, \$500,000, and \$900,000.

The unusual values stretch the distribution to the **right**.

Example 2B: Which way is the skew?

Suppose most household incomes in a sample lie between \$40,000 and \$100,000, but a few observations are \$300,000, \$500,000, and \$900,000.

The unusual values stretch the distribution to the **right**.

So the distribution is **right-skewed**.

Example 2B: Which way is the skew?

Suppose most household incomes in a sample lie between \$40,000 and \$100,000, but a few observations are \$300,000, \$500,000, and \$900,000.

The unusual values stretch the distribution to the **right**.

So the distribution is **right-skewed**.

For a right-skewed distribution, we often expect

$$\bar{x} > \text{median}.$$

The reason is that the mean gets pulled toward the large values.

Bin width can change what you see

A histogram is not unique.

Bin width can change what you see

A histogram is not unique.

- ▶ Too few bins can hide important structure.
- ▶ Too many bins can make the graph look noisy and fragmented.

Bin width can change what you see

A histogram is not unique.

- ▶ Too few bins can hide important structure.
- ▶ Too many bins can make the graph look noisy and fragmented.

The chapter recommends using a reasonable number of equal-width bins and choosing convenient endpoints.

The histogram is a summary of the data, not the data itself.

Skill 3: Median

First, **sort the data**.

Skill 3: Median

First, **sort the data**.

If n is odd:

median = middle value.

Skill 3: Median

First, **sort the data**.

If n is odd:

median = middle value.

If n is even:

$$\text{median} = \frac{\text{two middle values added together}}{2}.$$

Skill 3: Median

First, **sort the data**.

If n is odd:

median = middle value.

If n is even:

$$\text{median} = \frac{\text{two middle values added together}}{2}.$$

The median is resistant to extreme values because it depends mainly on **position**, not magnitude.

Example 3A: Median with an odd number of values

Data:

9, 4, 11, 6, 8, 6, 12.

Example 3A: Median with an odd number of values

Data:

9, 4, 11, 6, 8, 6, 12.

Step 1: sort.

4, 6, 6, 8, 9, 11, 12.

Example 3A: Median with an odd number of values

Data:

9, 4, 11, 6, 8, 6, 12.

Step 1: sort.

4, 6, 6, 8, 9, 11, 12.

Step 2: count.

$$n = 7.$$

Example 3A: Median with an odd number of values

Data:

9, 4, 11, 6, 8, 6, 12.

Step 1: sort.

4, 6, 6, 8, 9, 11, 12.

Step 2: count.

$$n = 7.$$

Step 3: locate the middle value.

$$\text{median} = 8.$$

Example 3B: Median with an even number of values

Data:

12, 5, 8, 3, 10, 7.

Example 3B: Median with an even number of values

Data:

12, 5, 8, 3, 10, 7.

Sort:

3, 5, 7, 8, 10, 12.

Example 3B: Median with an even number of values

Data:

12, 5, 8, 3, 10, 7.

Sort:

3, 5, 7, 8, 10, 12.

There are 6 values, so average the third and fourth:

$$\text{median} = \frac{7 + 8}{2}.$$

Example 3B: Median with an even number of values

Data:

12, 5, 8, 3, 10, 7.

Sort:

3, 5, 7, 8, 10, 12.

There are 6 values, so average the third and fourth:

$$\text{median} = \frac{7 + 8}{2}.$$

$\text{median} = 7.5$

Skill 3: Mean

The **mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Skill 3: Mean

The **mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Translation into plain English:

add everything, then divide by how many values there are.

Skill 3: Mean

The **mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Translation into plain English:

add everything, then divide by how many values there are.

Unlike the median, the mean uses the numerical size of **every** observation.

Example 3C: Mean and median agree

Data:

4, 6, 6, 8, 9, 11, 12.

Example 3C: Mean and median agree

Data:

4, 6, 6, 8, 9, 11, 12.

Median:

8

Example 3C: Mean and median agree

Data:

4, 6, 6, 8, 9, 11, 12.

Median:

8

Mean:

$$\bar{x} = \frac{4 + 6 + 6 + 8 + 9 + 11 + 12}{7} = \frac{56}{7}.$$

Example 3C: Mean and median agree

Data:

4, 6, 6, 8, 9, 11, 12.

Median:

8

Mean:

$$\bar{x} = \frac{4 + 6 + 6 + 8 + 9 + 11 + 12}{7} = \frac{56}{7}.$$

$$\bar{x} = 8$$

Here the center is summarized nicely by either measure.

Example 3D: One extreme value changes the story

Data:

4, 6, 6, 8, 9, 11, 40.

Example 3D: One extreme value changes the story

Data:

4, 6, 6, 8, 9, 11, 40.

Median:

8

Example 3D: One extreme value changes the story

Data:

4, 6, 6, 8, 9, 11, 40.

Median:

8

Mean:

$$\bar{x} = \frac{4 + 6 + 6 + 8 + 9 + 11 + 40}{7} = \frac{84}{7} = 12.$$

Example 3D: One extreme value changes the story

Data:

4, 6, 6, 8, 9, 11, 40.

Median:

8

Mean:

$$\bar{x} = \frac{4 + 6 + 6 + 8 + 9 + 11 + 40}{7} = \frac{84}{7} = 12.$$

The single value 40 pulls the mean upward but does not change the median.

For strongly skewed data, the **median** is often a more useful description of a typical value.

Mean from a frequency table

Sometimes the data are presented like this:

x	2	3	4	5
f	1	3	4	2

Mean from a frequency table

Sometimes the data are presented like this:

x	2	3	4	5
f	1	3	4	2

Instead of writing repeated values, use

$$\bar{x} = \frac{\sum fx}{\sum f}.$$

Mean from a frequency table

Sometimes the data are presented like this:

x	2	3	4	5
f	1	3	4	2

Instead of writing repeated values, use

$$\bar{x} = \frac{\sum fx}{\sum f}.$$

Here:

$$n = 1 + 3 + 4 + 2 = 10.$$

Mean from a frequency table

Sometimes the data are presented like this:

x	2	3	4	5
f	1	3	4	2

Instead of writing repeated values, use

$$\bar{x} = \frac{\sum fx}{\sum f}.$$

Here:

$$n = 1 + 3 + 4 + 2 = 10.$$

$$\bar{x} = \frac{1(2) + 3(3) + 4(4) + 2(5)}{10} = \frac{37}{10} = \boxed{3.7}.$$

Skill 4: Range

The **range** measures total spread:

$$\text{range} = \text{maximum} - \text{minimum}.$$

Skill 4: Range

The **range** measures total spread:

$$\text{range} = \text{maximum} - \text{minimum}.$$

Example:

3, 5, 8, 10, 14.

Skill 4: Range

The **range** measures total spread:

$$\text{range} = \text{maximum} - \text{minimum}.$$

Example:

3, 5, 8, 10, 14.

$$\text{range} = 14 - 3 = 11.$$

Skill 4: Range

The **range** measures total spread:

$$\text{range} = \text{maximum} - \text{minimum}.$$

Example:

3, 5, 8, 10, 14.

$$\text{range} = 14 - 3 = 11.$$

The range is one number, not the pair (3, 14).

Skill 4: Quartiles

Quartiles divide ordered data into quarters.

- ▶ Q_1 : first quartile,
- ▶ Q_2 : median,
- ▶ Q_3 : third quartile.

Skill 4: Quartiles

Quartiles divide ordered data into quarters.

- ▶ Q_1 : first quartile,
- ▶ Q_2 : median,
- ▶ Q_3 : third quartile.

We will follow the convention used in the chapter:

- ▶ Find the overall median.
- ▶ Q_1 is the median of the values below the overall median.
- ▶ Q_3 is the median of the values above the overall median.

Example 4A: Quartiles and IQR

Ordered data:

2, 4, 5, 7, 8, 10, 12, 13, 20.

Example 4A: Quartiles and IQR

Ordered data:

2, 4, 5, 7, 8, 10, 12, 13, 20.

There are 9 values, so the median is the fifth value:

$$Q_2 = 8.$$

Example 4A: Quartiles and IQR

Ordered data:

2, 4, 5, 7, 8, 10, 12, 13, 20.

There are 9 values, so the median is the fifth value:

$$Q_2 = 8.$$

Lower half:

2, 4, 5, 7.

Example 4A: Quartiles and IQR

Ordered data:

2, 4, 5, 7, 8, 10, 12, 13, 20.

There are 9 values, so the median is the fifth value:

$$Q_2 = 8.$$

Lower half:

2, 4, 5, 7.

$$Q_1 = \frac{4 + 5}{2} = \boxed{4.5}.$$

Example 4A: Quartiles and IQR

Ordered data:

2, 4, 5, 7, 8, 10, 12, 13, 20.

There are 9 values, so the median is the fifth value:

$$Q_2 = 8.$$

Lower half:

2, 4, 5, 7.

$$Q_1 = \frac{4 + 5}{2} = \boxed{4.5}.$$

Upper half:

10, 12, 13, 20.

Example 4A: Quartiles and IQR

Ordered data:

2, 4, 5, 7, 8, 10, 12, 13, 20.

There are 9 values, so the median is the fifth value:

$$Q_2 = 8.$$

Lower half:

2, 4, 5, 7.

$$Q_1 = \frac{4 + 5}{2} = \boxed{4.5}.$$

Upper half:

10, 12, 13, 20.

$$Q_3 = \frac{12 + 13}{2} = \boxed{12.5}.$$

Example 4A: Finish the spread measures

We found

$$Q_1 = 4.5, \quad Q_2 = 8, \quad Q_3 = 12.5.$$

Example 4A: Finish the spread measures

We found

$$Q_1 = 4.5, \quad Q_2 = 8, \quad Q_3 = 12.5.$$

The **interquartile range** is

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

Example 4A: Finish the spread measures

We found

$$Q_1 = 4.5, \quad Q_2 = 8, \quad Q_3 = 12.5.$$

The **interquartile range** is

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

$$\text{IQR} = 12.5 - 4.5 = \boxed{8}.$$

Example 4A: Finish the spread measures

We found

$$Q_1 = 4.5, \quad Q_2 = 8, \quad Q_3 = 12.5.$$

The **interquartile range** is

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

$$\text{IQR} = 12.5 - 4.5 = \boxed{8}.$$

The full range is

$$20 - 2 = \boxed{18}.$$

Example 4A: Finish the spread measures

We found

$$Q_1 = 4.5, \quad Q_2 = 8, \quad Q_3 = 12.5.$$

The **interquartile range** is

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

$$\text{IQR} = 12.5 - 4.5 = \boxed{8}.$$

The full range is

$$20 - 2 = \boxed{18}.$$

Interpretation:

- ▶ range = 18 describes the full data span;
- ▶ IQR = 8 describes the spread of the middle 50%.

Example 4B: Quartiles when n is even

Ordered data:

3, 4, 4, 5, 7, 9, 10, 11.

Example 4B: Quartiles when n is even

Ordered data:

3, 4, 4, 5, 7, 9, 10, 11.

Overall median:

$$Q_2 = \frac{5 + 7}{2} = 6.$$

Example 4B: Quartiles when n is even

Ordered data:

3, 4, 4, 5, 7, 9, 10, 11.

Overall median:

$$Q_2 = \frac{5 + 7}{2} = 6.$$

Lower half:

$$3, 4, 4, 5 \Rightarrow Q_1 = \frac{4 + 4}{2} = \boxed{4}.$$

Example 4B: Quartiles when n is even

Ordered data:

3, 4, 4, 5, 7, 9, 10, 11.

Overall median:

$$Q_2 = \frac{5 + 7}{2} = 6.$$

Lower half:

$$3, 4, 4, 5 \Rightarrow Q_1 = \frac{4 + 4}{2} = \boxed{4}.$$

Upper half:

$$7, 9, 10, 11 \Rightarrow Q_3 = \frac{9 + 10}{2} = \boxed{9.5}.$$

Example 4B: Quartiles when n is even

Ordered data:

3, 4, 4, 5, 7, 9, 10, 11.

Overall median:

$$Q_2 = \frac{5 + 7}{2} = 6.$$

Lower half:

$$3, 4, 4, 5 \Rightarrow Q_1 = \frac{4 + 4}{2} = \boxed{4}.$$

Upper half:

$$7, 9, 10, 11 \Rightarrow Q_3 = \frac{9 + 10}{2} = \boxed{9.5}.$$

$$\text{IQR} = 9.5 - 4 = \boxed{5.5}.$$

Five-number summary

The information needed for the chapter's basic box plot is:

min, Q_1 , median, Q_3 , max.

Five-number summary

The information needed for the chapter's basic box plot is:

$\text{min}, Q_1, \text{median}, Q_3, \text{max}$.

For

2, 4, 5, 7, 8, 10, 12, 13, 20,

the five-number summary is

2, 4.5, 8, 12.5, 20.

Skill 5: Standard deviation

The sample standard deviation is

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

Skill 5: Standard deviation

The sample standard deviation is

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

Do not try to swallow the formula whole. Use a four-step routine:

- 1 Find the mean.
- 2 Find each deviation $x_i - \bar{x}$.
- 3 Square the deviations and add them.
- 4 Divide by $n - 1$, then take the square root.

Variance

The **variance** is

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

Variance

The **variance** is

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

So:

$$s = \sqrt{s^2}.$$

Variance

The **variance** is

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

So:

$$s = \sqrt{s^2}.$$

Units matter:

- ▶ s has the same units as the original data;
- ▶ s^2 has squared units.

Example 5A: Standard deviation by hand

Data:

2, 4, 4, 4, 6.

Find s and s^2 .

Example 5A: Standard deviation by hand

Data:

2, 4, 4, 4, 6.

Find s and s^2 .

Step 1: mean.

$$\bar{x} = \frac{2 + 4 + 4 + 4 + 6}{5} = \frac{20}{5} = \boxed{4}.$$

Example 5A: Deviations and squares

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-2	4
4	0	0
4	0	0
4	0	0
6	2	4

Example 5A: Deviations and squares

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-2	4
4	0	0
4	0	0
4	0	0
6	2	4

Sum of squared deviations:

$$4 + 0 + 0 + 0 + 4 = 8.$$

Example 5A: Deviations and squares

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-2	4
4	0	0
4	0	0
4	0	0
6	2	4

Sum of squared deviations:

$$4 + 0 + 0 + 0 + 4 = 8.$$

Variance:

$$s^2 = \frac{8}{5 - 1} = \frac{8}{4} = \boxed{2}.$$

Example 5A: Deviations and squares

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-2	4
4	0	0
4	0	0
4	0	0
6	2	4

Sum of squared deviations:

$$4 + 0 + 0 + 0 + 4 = 8.$$

Variance:

$$s^2 = \frac{8}{5 - 1} = \frac{8}{4} = \boxed{2}.$$

Standard deviation:

$$s = \sqrt{2} \approx \boxed{1.41}.$$

Example 5B: Same mean, different spread

Compare:

$A : 4, 4, 4, 4, 4$

and

$B : 0, 2, 4, 6, 8.$

Example 5B: Same mean, different spread

Compare:

$A : 4, 4, 4, 4, 4$

and

$B : 0, 2, 4, 6, 8.$

Both have mean

$$\bar{x} = 4.$$

Example 5B: Same mean, different spread

Compare:

$A : 4, 4, 4, 4, 4$

and

$B : 0, 2, 4, 6, 8.$

Both have mean

$$\bar{x} = 4.$$

For data set A :

$$s_A = 0.$$

Example 5B: Same mean, different spread

Compare:

$$A : 4, 4, 4, 4, 4$$

and

$$B : 0, 2, 4, 6, 8.$$

Both have mean

$$\bar{x} = 4.$$

For data set A :

$$s_A = 0.$$

For data set B :

$$\sum (x_i - \bar{x})^2 = 16 + 4 + 0 + 4 + 16 = 40.$$

Example 5B: Same mean, different spread

Compare:

$$A : 4, 4, 4, 4, 4$$

and

$$B : 0, 2, 4, 6, 8.$$

Both have mean

$$\bar{x} = 4.$$

For data set A :

$$s_A = 0.$$

For data set B :

$$\sum (x_i - \bar{x})^2 = 16 + 4 + 0 + 4 + 16 = 40.$$

$$s_B^2 = \frac{40}{4} = 10, \quad s_B = \sqrt{10} \approx 3.16.$$

Example 5B: Same mean, different spread

Compare:

$$A : 4, 4, 4, 4, 4$$

and

$$B : 0, 2, 4, 6, 8.$$

Both have mean

$$\bar{x} = 4.$$

For data set A :

$$s_A = 0.$$

For data set B :

$$\sum (x_i - \bar{x})^2 = 16 + 4 + 0 + 4 + 16 = 40.$$

$$s_B^2 = \frac{40}{4} = 10, \quad s_B = \sqrt{10} \approx 3.16.$$

Same center. Very different spread.

Example 5C: Standard deviation from a frequency table

x	1	3	5
f	1	2	1

Example 5C: Standard deviation from a frequency table

x	1	3	5
f	1	2	1

Total number of values:

$$n = 1 + 2 + 1 = 4.$$

Example 5C: Standard deviation from a frequency table

x	1	3	5
f	1	2	1

Total number of values:

$$n = 1 + 2 + 1 = 4.$$

Mean:

$$\bar{x} = \frac{1(1) + 2(3) + 1(5)}{4} = \frac{12}{4} = 3.$$

Example 5C: Standard deviation from a frequency table

x	1	3	5
f	1	2	1

Total number of values:

$$n = 1 + 2 + 1 = 4.$$

Mean:

$$\bar{x} = \frac{1(1) + 2(3) + 1(5)}{4} = \frac{12}{4} = 3.$$

Sum of squared deviations:

$$1(1 - 3)^2 + 2(3 - 3)^2 + 1(5 - 3)^2 = 4 + 0 + 4 = 8.$$

Example 5C: Standard deviation from a frequency table

x	1	3	5
f	1	2	1

Total number of values:

$$n = 1 + 2 + 1 = 4.$$

Mean:

$$\bar{x} = \frac{1(1) + 2(3) + 1(5)}{4} = \frac{12}{4} = 3.$$

Sum of squared deviations:

$$1(1 - 3)^2 + 2(3 - 3)^2 + 1(5 - 3)^2 = 4 + 0 + 4 = 8.$$

$$s = \sqrt{\frac{8}{3}} \approx \boxed{1.63}.$$

Useful skill from the exercises: shifting data

Suppose every data value is replaced by

$$y = x + c.$$

Useful skill from the exercises: shifting data

Suppose every data value is replaced by

$$y = x + c.$$

Then:

$$\bar{x}_y = \bar{x}_x + c$$

and

$$\text{median}_y = \text{median}_x + c.$$

Useful skill from the exercises: shifting data

Suppose every data value is replaced by

$$y = x + c.$$

Then:

$$\bar{x}_y = \bar{x}_x + c$$

and

$$\text{median}_y = \text{median}_x + c.$$

But adding the same constant does **not** change the spread:

$$s_y = s_x, \quad s_y^2 = s_x^2.$$

Example 6A: Add a constant

Suppose a data set has

$$\bar{x} = 10, \quad \text{median} = 9, \quad s = 2, \quad s^2 = 4.$$

Every value is increased by 15.

Example 6A: Add a constant

Suppose a data set has

$$\bar{x} = 10, \quad \text{median} = 9, \quad s = 2, \quad s^2 = 4.$$

Every value is increased by 15.

New mean:

$$10 + 15 = \boxed{25}.$$

Example 6A: Add a constant

Suppose a data set has

$$\bar{x} = 10, \quad \text{median} = 9, \quad s = 2, \quad s^2 = 4.$$

Every value is increased by 15.

New mean:

$$10 + 15 = \boxed{25}.$$

New median:

$$9 + 15 = \boxed{24}.$$

Example 6A: Add a constant

Suppose a data set has

$$\bar{x} = 10, \quad \text{median} = 9, \quad s = 2, \quad s^2 = 4.$$

Every value is increased by 15.

New mean:

$$10 + 15 = \boxed{25}.$$

New median:

$$9 + 15 = \boxed{24}.$$

New standard deviation:

$$\boxed{2}.$$

Example 6A: Add a constant

Suppose a data set has

$$\bar{x} = 10, \quad \text{median} = 9, \quad s = 2, \quad s^2 = 4.$$

Every value is increased by 15.

New mean:

$$10 + 15 = \boxed{25}.$$

New median:

$$9 + 15 = \boxed{24}.$$

New standard deviation:

$$\boxed{2}.$$

New variance:

$$\boxed{4}.$$

Scaling data

Suppose every value is replaced by

$$y = ax.$$

Scaling data

Suppose every value is replaced by

$$y = ax.$$

Then:

$$\boxed{\bar{x}_y = a\bar{x}_x}, \quad \boxed{\text{median}_y = a \text{median}_x}.$$

Scaling data

Suppose every value is replaced by

$$y = ax.$$

Then:

$$\boxed{\bar{x}_y = a\bar{x}_x}, \quad \boxed{\text{median}_y = a \text{median}_x}.$$

For spread:

$$\boxed{s_y = |a|s_x}, \quad \boxed{s_y^2 = a^2 s_x^2}.$$

The absolute value in the standard deviation formula matters when $a < 0$.

Example 6B: Multiply by a negative number

Suppose

$$\bar{x} = 3, \quad \text{median} = 4, \quad s = 2, \quad s^2 = 4.$$

Every value is multiplied by -5 .

Example 6B: Multiply by a negative number

Suppose

$$\bar{x} = 3, \quad \text{median} = 4, \quad s = 2, \quad s^2 = 4.$$

Every value is multiplied by -5 .

$$\bar{x}_{\text{new}} = -5(3) = \boxed{-15}.$$

Example 6B: Multiply by a negative number

Suppose

$$\bar{x} = 3, \quad \text{median} = 4, \quad s = 2, \quad s^2 = 4.$$

Every value is multiplied by -5 .

$$\bar{x}_{\text{new}} = -5(3) = \boxed{-15}.$$

$$\text{median}_{\text{new}} = -5(4) = \boxed{-20}.$$

Example 6B: Multiply by a negative number

Suppose

$$\bar{x} = 3, \quad \text{median} = 4, \quad s = 2, \quad s^2 = 4.$$

Every value is multiplied by -5 .

$$\bar{x}_{\text{new}} = -5(3) = \boxed{-15}.$$

$$\text{median}_{\text{new}} = -5(4) = \boxed{-20}.$$

$$s_{\text{new}} = |-5|(2) = \boxed{10}.$$

Example 6B: Multiply by a negative number

Suppose

$$\bar{x} = 3, \quad \text{median} = 4, \quad s = 2, \quad s^2 = 4.$$

Every value is multiplied by -5 .

$$\bar{x}_{\text{new}} = -5(3) = \boxed{-15}.$$

$$\text{median}_{\text{new}} = -5(4) = \boxed{-20}.$$

$$s_{\text{new}} = |-5|(2) = \boxed{10}.$$

$$s_{\text{new}}^2 = (-5)^2(4) = \boxed{100}.$$

Skill 6: The Bell Curve Rule

For data with an approximately bell-shaped histogram:

$\bar{x} \pm s$ contains about 68%

Skill 6: The Bell Curve Rule

For data with an approximately bell-shaped histogram:

$\bar{x} \pm s$ contains about 68%

$\bar{x} \pm 2s$ contains about 95%

Skill 6: The Bell Curve Rule

For data with an approximately bell-shaped histogram:

$\bar{x} \pm s$ contains about 68%

$\bar{x} \pm 2s$ contains about 95%

$\bar{x} \pm 3s$ contains about 99.7%

Skill 6: The Bell Curve Rule

For data with an approximately bell-shaped histogram:

$\bar{x} \pm s$ contains about 68%

$\bar{x} \pm 2s$ contains about 95%

$\bar{x} \pm 3s$ contains about 99.7%

Memory device:

68% – 95% – 99.7% .

Before using the Bell Curve Rule

The rule is designed for distributions that are approximately:

- ▶ bell-shaped,
- ▶ reasonably symmetric,
- ▶ unimodal.

Before using the Bell Curve Rule

The rule is designed for distributions that are approximately:

- ▶ bell-shaped,
- ▶ reasonably symmetric,
- ▶ unimodal.

Do **not** blindly apply it to a strongly skewed or obviously bimodal distribution.

Before using the Bell Curve Rule

The rule is designed for distributions that are approximately:

- ▶ bell-shaped,
- ▶ reasonably symmetric,
- ▶ unimodal.

Do **not** blindly apply it to a strongly skewed or obviously bimodal distribution. A formula cannot rescue a bad assumption. Mathematics is rude that way.

Example 7A: Build the three intervals

Exam scores are approximately bell-shaped with

$$\bar{x} = 70, \quad s = 5.$$

Example 7A: Build the three intervals

Exam scores are approximately bell-shaped with

$$\bar{x} = 70, \quad s = 5.$$

Within one standard deviation:

$$70 \pm 5 = [65, 75].$$

Example 7A: Build the three intervals

Exam scores are approximately bell-shaped with

$$\bar{x} = 70, \quad s = 5.$$

Within one standard deviation:

$$70 \pm 5 = [65, 75].$$

Within two standard deviations:

$$70 \pm 10 = [60, 80].$$

Example 7A: Build the three intervals

Exam scores are approximately bell-shaped with

$$\bar{x} = 70, \quad s = 5.$$

Within one standard deviation:

$$70 \pm 5 = [65, 75].$$

Within two standard deviations:

$$70 \pm 10 = [60, 80].$$

Within three standard deviations:

$$70 \pm 15 = [55, 85].$$

Example 7A: Build the three intervals

Exam scores are approximately bell-shaped with

$$\bar{x} = 70, \quad s = 5.$$

Within one standard deviation:

$$70 \pm 5 = [65, 75].$$

Within two standard deviations:

$$70 \pm 10 = [60, 80].$$

Within three standard deviations:

$$70 \pm 15 = [55, 85].$$

So approximately:

68% are in $[65, 75]$,

95% are in $[60, 80]$,

99.7% are in $[55, 85]$.

Example 7B: Convert percentages into counts

A bell-shaped data set has 200 observations.

Approximately how many lie within one standard deviation of the mean?

Example 7B: Convert percentages into counts

A bell-shaped data set has 200 observations.

Approximately how many lie within one standard deviation of the mean?

$$0.68(200) = \boxed{136}.$$

Example 7B: Convert percentages into counts

A bell-shaped data set has 200 observations.

Approximately how many lie within one standard deviation of the mean?

$$0.68(200) = \boxed{136}.$$

Within two standard deviations:

$$0.95(200) = \boxed{190}.$$

Example 7B: Convert percentages into counts

A bell-shaped data set has 200 observations.

Approximately how many lie within one standard deviation of the mean?

$$0.68(200) = \boxed{136}.$$

Within two standard deviations:

$$0.95(200) = \boxed{190}.$$

Within three standard deviations:

$$0.997(200) = 199.4.$$

Example 7B: Convert percentages into counts

A bell-shaped data set has 200 observations.

Approximately how many lie within one standard deviation of the mean?

$$0.68(200) = \boxed{136}.$$

Within two standard deviations:

$$0.95(200) = \boxed{190}.$$

Within three standard deviations:

$$0.997(200) = 199.4.$$

So we expect essentially $\boxed{199 \text{ or } 200}$ observations.

Example 7C: Is a value unusual?

A bell-shaped distribution has

$$\bar{x} = 50, \quad s = 4.$$

Consider $x = 61$.

Example 7C: Is a value unusual?

A bell-shaped distribution has

$$\bar{x} = 50, \quad s = 4.$$

Consider $x = 61$.

Distance from the mean:

$$61 - 50 = 11.$$

Example 7C: Is a value unusual?

A bell-shaped distribution has

$$\bar{x} = 50, \quad s = 4.$$

Consider $x = 61$.

Distance from the mean:

$$61 - 50 = 11.$$

Number of standard deviations:

$$\frac{11}{4} = 2.75.$$

Example 7C: Is a value unusual?

A bell-shaped distribution has

$$\bar{x} = 50, \quad s = 4.$$

Consider $x = 61$.

Distance from the mean:

$$61 - 50 = 11.$$

Number of standard deviations:

$$\frac{11}{4} = 2.75.$$

Since 61 is more than two standard deviations above the mean, it would be regarded as **unusual**.

Example 7C: Is a value unusual?

A bell-shaped distribution has

$$\bar{x} = 50, \quad s = 4.$$

Consider $x = 61$.

Distance from the mean:

$$61 - 50 = 11.$$

Number of standard deviations:

$$\frac{11}{4} = 2.75.$$

Since 61 is more than two standard deviations above the mean, it would be regarded as **unusual**.

It is not quite three standard deviations away.

Range Rule

For an approximately bell-shaped data set,

$$s \approx \frac{\text{range}}{4}.$$

Range Rule

For an approximately bell-shaped data set,

$$s \approx \frac{\text{range}}{4}.$$

Why?

Almost all values should lie roughly from

$$\bar{x} - 2s \quad \text{to} \quad \bar{x} + 2s.$$

Range Rule

For an approximately bell-shaped data set,

$$s \approx \frac{\text{range}}{4}.$$

Why?

Almost all values should lie roughly from

$$\bar{x} - 2s \quad \text{to} \quad \bar{x} + 2s.$$

That interval has length

$$4s.$$

So:

$$\text{range} \approx 4s.$$

Example 7D: Range Rule

A roughly bell-shaped data set has

$$\min = 48, \quad \max = 88.$$

Example 7D: Range Rule

A roughly bell-shaped data set has

$$\min = 48, \quad \max = 88.$$

Range:

$$88 - 48 = 40.$$

Example 7D: Range Rule

A roughly bell-shaped data set has

$$\min = 48, \quad \max = 88.$$

Range:

$$88 - 48 = 40.$$

Estimate the standard deviation:

$$s \approx \frac{40}{4}.$$

Example 7D: Range Rule

A roughly bell-shaped data set has

$$\min = 48, \quad \max = 88.$$

Range:

$$88 - 48 = 40.$$

Estimate the standard deviation:

$$s \approx \frac{40}{4}.$$

$$\boxed{s \approx 10}.$$

Example 7D: Range Rule

A roughly bell-shaped data set has

$$\min = 48, \quad \max = 88.$$

Range:

$$88 - 48 = 40.$$

Estimate the standard deviation:

$$s \approx \frac{40}{4}.$$

$$\boxed{s \approx 10}.$$

This is only an estimate. It is not a replacement for the actual standard-deviation calculation when exact data are available.

Skill 7: Construct a box plot

A basic box plot uses the five-number summary:

min, Q_1 , median, Q_3 , max.

Skill 7: Construct a box plot

A basic box plot uses the five-number summary:

min, Q_1 , median, Q_3 , max.

Construction:

- 1 Draw a number line.
- 2 Draw a box from Q_1 to Q_3 .
- 3 Draw a line inside the box at the median.
- 4 Draw whiskers to the minimum and maximum.

Skill 7: Construct a box plot

A basic box plot uses the five-number summary:

min, Q_1 , median, Q_3 , max.

Construction:

- 1 Draw a number line.
- 2 Draw a box from Q_1 to Q_3 .
- 3 Draw a line inside the box at the median.
- 4 Draw whiskers to the minimum and maximum.

In the chapter's convention, the whiskers go all the way to the minimum and maximum.

Example 8A: Draw a box plot

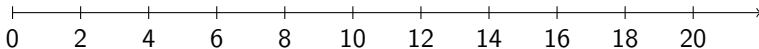
Five-number summary:

$$\min = 2, \quad Q_1 = 4.5, \quad \text{median} = 8, \quad Q_3 = 12.5, \quad \max = 20.$$

Example 8A: Draw a box plot

Five-number summary:

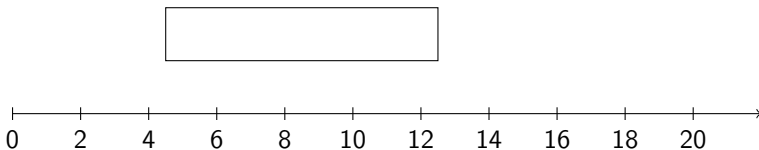
$$\min = 2, \quad Q_1 = 4.5, \quad \text{median} = 8, \quad Q_3 = 12.5, \quad \max = 20.$$



Example 8A: Draw a box plot

Five-number summary:

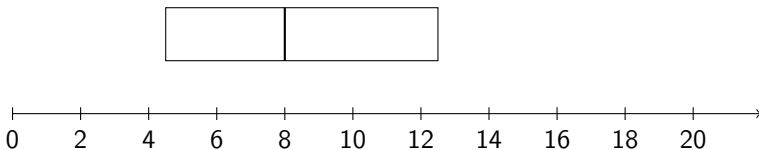
$\min = 2$, $Q_1 = 4.5$, $\text{median} = 8$, $Q_3 = 12.5$, $\max = 20$.



Example 8A: Draw a box plot

Five-number summary:

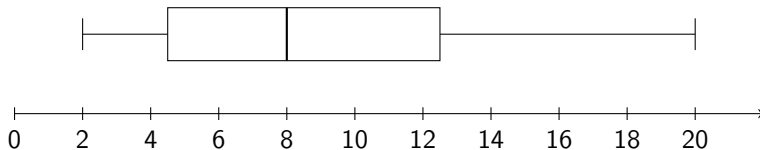
$\min = 2$, $Q_1 = 4.5$, $\text{median} = 8$, $Q_3 = 12.5$, $\max = 20$.



Example 8A: Draw a box plot

Five-number summary:

$\min = 2$, $Q_1 = 4.5$, $\text{median} = 8$, $Q_3 = 12.5$, $\max = 20$.



Reading a box plot

From a box plot you can read:

- ▶ minimum and maximum,
- ▶ median,
- ▶ Q_1 and Q_3 ,
- ▶ range,
- ▶ IQR,
- ▶ rough symmetry or skewness.

Reading a box plot

From a box plot you can read:

- ▶ minimum and maximum,
- ▶ median,
- ▶ Q_1 and Q_3 ,
- ▶ range,
- ▶ IQR,
- ▶ rough symmetry or skewness.

Each interval

$[\min, Q_1]$, $[Q_1, \text{median}]$, $[\text{median}, Q_3]$, $[Q_3, \max]$

contains roughly one quarter of the data.

Reading a box plot

From a box plot you can read:

- ▶ minimum and maximum,
- ▶ median,
- ▶ Q_1 and Q_3 ,
- ▶ range,
- ▶ IQR,
- ▶ rough symmetry or skewness.

Each interval

$[\min, Q_1]$, $[Q_1, \text{median}]$, $[\text{median}, Q_3]$, $[Q_3, \max]$

contains roughly one quarter of the data.

The **lengths** of those intervals need not be equal.

Example 8B: Read a box plot

Suppose a box plot has

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Find the IQR.

Example 8B: Read a box plot

Suppose a box plot has

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Find the IQR.

$$\text{IQR} = 40 - 20 = \boxed{20}.$$

Example 8B: Read a box plot

Suppose a box plot has

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Find the IQR.

$$\text{IQR} = 40 - 20 = \boxed{20}.$$

Find the range.

Example 8B: Read a box plot

Suppose a box plot has

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Find the IQR.

$$\text{IQR} = 40 - 20 = \boxed{20}.$$

Find the range.

$$55 - 10 = \boxed{45}.$$

Example 8B: Read a box plot

Suppose a box plot has

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Find the IQR.

$$\text{IQR} = 40 - 20 = \boxed{20}.$$

Find the range.

$$55 - 10 = \boxed{45}.$$

What percent of the data lie between $Q_1 = 20$ and $Q_3 = 40$?

Example 8B: Read a box plot

Suppose a box plot has

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Find the IQR.

$$\text{IQR} = 40 - 20 = \boxed{20}.$$

Find the range.

$$55 - 10 = \boxed{45}.$$

What percent of the data lie between $Q_1 = 20$ and $Q_3 = 40$?

$$\boxed{50\%}.$$

Example 8C: Use shape information

Again suppose

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Example 8C: Use shape information

Again suppose

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Compare the distances:

$$25 - 20 = 5,$$

but

$$40 - 25 = 15.$$

Example 8C: Use shape information

Again suppose

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Compare the distances:

$$25 - 20 = 5,$$

but

$$40 - 25 = 15.$$

The upper half of the box is much more spread out.

The right whisker is also fairly long.

Example 8C: Use shape information

Again suppose

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Compare the distances:

$$25 - 20 = 5,$$

but

$$40 - 25 = 15.$$

The upper half of the box is much more spread out.

The right whisker is also fairly long.

This suggests a distribution that is **skewed to the right**.

Example 8C: Use shape information

Again suppose

$$\min = 10, \quad Q_1 = 20, \quad \text{median} = 25, \quad Q_3 = 40, \quad \max = 55.$$

Compare the distances:

$$25 - 20 = 5,$$

but

$$40 - 25 = 15.$$

The upper half of the box is much more spread out.

The right whisker is also fairly long.

This suggests a distribution that is **skewed to the right**.

So we would often expect

$$\bar{x} > \text{median}.$$

Example 8D: Compare two box plots

Two data sets have:

	min	Q_1	Med.	Q_3	max
A	50	65	70	75	90
B	58	60	70	80	82

Example 8D: Compare two box plots

Two data sets have:

	min	Q_1	Med.	Q_3	max
A	50	65	70	75	90
B	58	60	70	80	82

Both medians are

70.

Example 8D: Compare two box plots

Two data sets have:

	min	Q_1	Med.	Q_3	max
A	50	65	70	75	90
B	58	60	70	80	82

Both medians are

70.

IQRs:

$$\text{IQR}_A = 75 - 65 = 10, \quad \text{IQR}_B = 80 - 60 = 20.$$

Example 8D: Compare two box plots

Two data sets have:

	min	Q_1	Med.	Q_3	max
A	50	65	70	75	90
B	58	60	70	80	82

Both medians are

70.

IQRs:

$$\text{IQR}_A = 75 - 65 = 10, \quad \text{IQR}_B = 80 - 60 = 20.$$

So the **middle half** of B is more spread out.

Example 8D: Compare two box plots

Two data sets have:

	min	Q_1	Med.	Q_3	max
A	50	65	70	75	90
B	58	60	70	80	82

Both medians are

70.

IQRs:

$$\text{IQR}_A = 75 - 65 = 10, \quad \text{IQR}_B = 80 - 60 = 20.$$

So the **middle half** of B is more spread out.

Ranges:

$$\text{range}(A) = 40, \quad \text{range}(B) = 24.$$

So A has the larger **overall** spread because its extremes are farther apart.

Skill 8: Cumulative relative frequency

The **cumulative relative frequency** tells us the proportion of observations at or below the current bin endpoint.

Skill 8: Cumulative relative frequency

The **cumulative relative frequency** tells us the proportion of observations at or below the current bin endpoint.

Example:

Bin	Freq.	Rel. Freq.
(0, 10]	2	0.10
(10, 20]	4	0.20
(20, 30]	8	0.40
(30, 40]	5	0.25
(40, 50]	1	0.05

Skill 8: Cumulative relative frequency

The **cumulative relative frequency** tells us the proportion of observations at or below the current bin endpoint.

Example:

Bin	Freq.	Rel. Freq.
(0, 10]	2	0.10
(10, 20]	4	0.20
(20, 30]	8	0.40
(30, 40]	5	0.25
(40, 50]	1	0.05

To get the cumulative column, keep a running total.

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	0.10
(10, 20]	4	0.20	

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	0.10
(10, 20]	4	0.20	0.30
(20, 30]	8	0.40	

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	0.10
(10, 20]	4	0.20	0.30
(20, 30]	8	0.40	0.70
(30, 40]	5	0.25	

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	0.10
(10, 20]	4	0.20	0.30
(20, 30]	8	0.40	0.70
(30, 40]	5	0.25	0.95
(40, 50]	1	0.05	

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	0.10
(10, 20]	4	0.20	0.30
(20, 30]	8	0.40	0.70
(30, 40]	5	0.25	0.95
(40, 50]	1	0.05	1.00

Example 9A: Build the cumulative column

Bin	f	Rel. Freq.	Cum. Rel. Freq.
(0, 10]	2	0.10	0.10
(10, 20]	4	0.20	0.30
(20, 30]	8	0.40	0.70
(30, 40]	5	0.25	0.95
(40, 50]	1	0.05	1.00

The final cumulative relative frequency should be 1 (apart from possible rounding error).

Example 9B: Locate quartiles from grouped data

Use the cumulative frequencies:

0.10, 0.30, 0.70, 0.95, 1.00.

Example 9B: Locate quartiles from grouped data

Use the cumulative frequencies:

0.10, 0.30, 0.70, 0.95, 1.00.

Q_1 corresponds to $25\% = 0.25$.

Since

$$0.10 < 0.25 \leq 0.30,$$

Q_1 lies in $(10, 20]$.

Example 9B: Locate quartiles from grouped data

Use the cumulative frequencies:

0.10, 0.30, 0.70, 0.95, 1.00.

Q_1 corresponds to $25\% = 0.25$.

Since

$$0.10 < 0.25 \leq 0.30,$$

Q_1 lies in $(10, 20]$.

The median corresponds to $50\% = 0.50$.

Since

$$0.30 < 0.50 \leq 0.70,$$

median lies in $(20, 30]$.

Example 9B: Locate quartiles from grouped data

Use the cumulative frequencies:

0.10, 0.30, 0.70, 0.95, 1.00.

Q_1 corresponds to $25\% = 0.25$.

Since

$$0.10 < 0.25 \leq 0.30,$$

Q_1 lies in $(10, 20]$.

The median corresponds to $50\% = 0.50$.

Since

$$0.30 < 0.50 \leq 0.70,$$

median lies in $(20, 30]$.

Q_3 corresponds to $75\% = 0.75$.

Since

$$0.70 < 0.75 \leq 0.95,$$

Q_3 lies in $(30, 40]$.

Estimating a mean from grouped data

When only bins are known, we no longer know the exact values.

Estimating a mean from grouped data

When only bins are known, we no longer know the exact values.

We replace every observation in a bin by the **midpoint** of that bin.

Estimating a mean from grouped data

When only bins are known, we no longer know the exact values.

We replace every observation in a bin by the **midpoint** of that bin.

For the bins

$(0, 10]$, $(10, 20]$, $(20, 30]$, $(30, 40]$, $(40, 50]$,

the midpoints are

5, 15, 25, 35, 45.

Estimating a mean from grouped data

When only bins are known, we no longer know the exact values.

We replace every observation in a bin by the **midpoint** of that bin.

For the bins

$(0, 10], (10, 20], (20, 30], (30, 40], (40, 50],$

the midpoints are

5, 15, 25, 35, 45.

Then use a weighted mean:

$$\bar{x} \approx \frac{\sum f(\text{midpoint})}{\sum f}.$$

Example 9C: Estimate the grouped mean

Bin	f	Midpoint
(0, 10]	2	5
(10, 20]	4	15
(20, 30]	8	25
(30, 40]	5	35
(40, 50]	1	45

Example 9C: Estimate the grouped mean

Bin	f	Midpoint
(0, 10]	2	5
(10, 20]	4	15
(20, 30]	8	25
(30, 40]	5	35
(40, 50]	1	45

Total:

$$n = 2 + 4 + 8 + 5 + 1 = 20.$$

Example 9C: Estimate the grouped mean

Bin	f	Midpoint
(0, 10]	2	5
(10, 20]	4	15
(20, 30]	8	25
(30, 40]	5	35
(40, 50]	1	45

Total:

$$n = 2 + 4 + 8 + 5 + 1 = 20.$$

Weighted sum:

$$2(5) + 4(15) + 8(25) + 5(35) + 1(45).$$

Example 9C: Estimate the grouped mean

Bin	f	Midpoint
(0, 10]	2	5
(10, 20]	4	15
(20, 30]	8	25
(30, 40]	5	35
(40, 50]	1	45

Total:

$$n = 2 + 4 + 8 + 5 + 1 = 20.$$

Weighted sum:

$$2(5) + 4(15) + 8(25) + 5(35) + 1(45).$$

$$= 10 + 60 + 200 + 175 + 45 = 490.$$

Example 9C: Estimate the grouped mean

Bin	f	Midpoint
(0, 10]	2	5
(10, 20]	4	15
(20, 30]	8	25
(30, 40]	5	35
(40, 50]	1	45

Total:

$$n = 2 + 4 + 8 + 5 + 1 = 20.$$

Weighted sum:

$$2(5) + 4(15) + 8(25) + 5(35) + 1(45).$$

$$= 10 + 60 + 200 + 175 + 45 = 490.$$

$$\bar{x} \approx \frac{490}{20} = \boxed{24.5}.$$

Estimating standard deviation from grouped data

Use the same midpoint approximation.

Estimating standard deviation from grouped data

Use the same midpoint approximation.

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

where m is the bin midpoint.

Estimating standard deviation from grouped data

Use the same midpoint approximation.

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

where m is the bin midpoint.

The only new idea is the factor f : if a midpoint represents 8 observations, its squared deviation must be counted 8 times.

Example 9D: Estimate grouped standard deviation

From the previous example:

$$\bar{x} \approx 24.5, \quad n = 20.$$

Example 9D: Estimate grouped standard deviation

From the previous example:

$$\bar{x} \approx 24.5, \quad n = 20.$$

Compute the weighted squared-deviation sum:

$$\begin{aligned} &2(5 - 24.5)^2 + 4(15 - 24.5)^2 + 8(25 - 24.5)^2 \\ &\quad + 5(35 - 24.5)^2 + 1(45 - 24.5)^2. \end{aligned}$$

Example 9D: Estimate grouped standard deviation

From the previous example:

$$\bar{x} \approx 24.5, \quad n = 20.$$

Compute the weighted squared-deviation sum:

$$\begin{aligned} &2(5 - 24.5)^2 + 4(15 - 24.5)^2 + 8(25 - 24.5)^2 \\ &+ 5(35 - 24.5)^2 + 1(45 - 24.5)^2. \end{aligned}$$

This equals

$$2095.$$

Example 9D: Estimate grouped standard deviation

From the previous example:

$$\bar{x} \approx 24.5, \quad n = 20.$$

Compute the weighted squared-deviation sum:

$$\begin{aligned} &2(5 - 24.5)^2 + 4(15 - 24.5)^2 + 8(25 - 24.5)^2 \\ &+ 5(35 - 24.5)^2 + 1(45 - 24.5)^2. \end{aligned}$$

This equals

$$2095.$$

So

$$s^2 \approx \frac{2095}{19} \approx 110.26.$$

Example 9D: Estimate grouped standard deviation

From the previous example:

$$\bar{x} \approx 24.5, \quad n = 20.$$

Compute the weighted squared-deviation sum:

$$\begin{aligned} &2(5 - 24.5)^2 + 4(15 - 24.5)^2 + 8(25 - 24.5)^2 \\ &+ 5(35 - 24.5)^2 + 1(45 - 24.5)^2. \end{aligned}$$

This equals

$$2095.$$

So

$$s^2 \approx \frac{2095}{19} \approx 110.26.$$

$$s \approx \sqrt{110.26} \approx \boxed{10.50}.$$

If you have the **original raw data**, compute statistics from the raw data.

Exact vs. estimated

If you have the **original raw data**, compute statistics from the raw data.

If you have only grouped intervals and frequencies:

- ▶ quartiles can usually be located only approximately;
- ▶ mean and standard deviation are estimated using bin midpoints.

Exact vs. estimated

If you have the **original raw data**, compute statistics from the raw data.

If you have only grouped intervals and frequencies:

- ▶ quartiles can usually be located only approximately;
- ▶ mean and standard deviation are estimated using bin midpoints.

Why only estimates?

Because all observations in a bin are temporarily treated as if they were sitting at the midpoint.

Mixed Example 1: What measure of center?

A data set of home prices is strongly right-skewed because a few homes cost several million dollars.

Which is usually the better description of a “typical” home price: mean or median?

Mixed Example 1: What measure of center?

A data set of home prices is strongly right-skewed because a few homes cost several million dollars.

Which is usually the better description of a “typical” home price: mean or median?

Median

Mixed Example 1: What measure of center?

A data set of home prices is strongly right-skewed because a few homes cost several million dollars.

Which is usually the better description of a “typical” home price: mean or median?

Median

Reason: a few extremely expensive homes can pull the mean far to the right.

Mixed Example 2: Work backward from the Bell Curve Rule

A bell-shaped distribution has

$$\bar{x} = 100, \quad s = 12.$$

Approximately 95% of the data should lie in what interval?

Mixed Example 2: Work backward from the Bell Curve Rule

A bell-shaped distribution has

$$\bar{x} = 100, \quad s = 12.$$

Approximately 95% of the data should lie in what interval?

Use

$$\bar{x} \pm 2s.$$

Mixed Example 2: Work backward from the Bell Curve Rule

A bell-shaped distribution has

$$\bar{x} = 100, \quad s = 12.$$

Approximately 95% of the data should lie in what interval?

Use

$$\bar{x} \pm 2s.$$

$$100 \pm 24.$$

Mixed Example 2: Work backward from the Bell Curve Rule

A bell-shaped distribution has

$$\bar{x} = 100, \quad s = 12.$$

Approximately 95% of the data should lie in what interval?

Use

$$\bar{x} \pm 2s.$$

$$100 \pm 24.$$

$$\boxed{[76, 124]}.$$

Mixed Example 3: Box-plot percentages

A box plot has

$$Q_1 = 18, \quad \text{median} = 25, \quad Q_3 = 33.$$

Approximately what percent of observations lie between 18 and 25?

Mixed Example 3: Box-plot percentages

A box plot has

$$Q_1 = 18, \quad \text{median} = 25, \quad Q_3 = 33.$$

Approximately what percent of observations lie between 18 and 25?

That is the interval from Q_1 to the median.

Mixed Example 3: Box-plot percentages

A box plot has

$$Q_1 = 18, \quad \text{median} = 25, \quad Q_3 = 33.$$

Approximately what percent of observations lie between 18 and 25?

That is the interval from Q_1 to the median.

25%.

Mixed Example 3: Box-plot percentages

A box plot has

$$Q_1 = 18, \quad \text{median} = 25, \quad Q_3 = 33.$$

Approximately what percent of observations lie between 18 and 25?

That is the interval from Q_1 to the median.

25%.

What percent lie between 18 and 33?

Mixed Example 3: Box-plot percentages

A box plot has

$$Q_1 = 18, \quad \text{median} = 25, \quad Q_3 = 33.$$

Approximately what percent of observations lie between 18 and 25?

That is the interval from Q_1 to the median.

25%.

What percent lie between 18 and 33?

50%.

Mixed Example 4: Do not confuse range and IQR

Suppose

$$\min = 2, \quad Q_1 = 7, \quad \text{median} = 10, \quad Q_3 = 15, \quad \max = 30.$$

Mixed Example 4: Do not confuse range and IQR

Suppose

$$\min = 2, \quad Q_1 = 7, \quad \text{median} = 10, \quad Q_3 = 15, \quad \max = 30.$$

Range:

$$30 - 2 = \boxed{28}.$$

Mixed Example 4: Do not confuse range and IQR

Suppose

$$\min = 2, \quad Q_1 = 7, \quad \text{median} = 10, \quad Q_3 = 15, \quad \max = 30.$$

Range:

$$30 - 2 = \boxed{28}.$$

IQR:

$$15 - 7 = \boxed{8}.$$

Mixed Example 4: Do not confuse range and IQR

Suppose

$$\min = 2, \quad Q_1 = 7, \quad \text{median} = 10, \quad Q_3 = 15, \quad \max = 30.$$

Range:

$$30 - 2 = \boxed{28}.$$

IQR:

$$15 - 7 = \boxed{8}.$$

They answer different questions:

- ▶ range = spread of **all** observations;
- ▶ IQR = spread of the **middle 50%**.

Mixed Example 5: Catch the wrong standard deviation

For data

1, 3, 5,

the mean is 3.

A student writes

$$s = \sqrt{\frac{(1-3)^2 + (3-3)^2 + (5-3)^2}{3}}.$$

What is wrong?

Mixed Example 5: Catch the wrong standard deviation

For data

1, 3, 5,

the mean is 3.

A student writes

$$s = \sqrt{\frac{(1-3)^2 + (3-3)^2 + (5-3)^2}{3}}.$$

What is wrong?

For the chapter's **sample** standard deviation, divide by

$$n - 1,$$

not n .

Mixed Example 5: Catch the wrong standard deviation

For data

1, 3, 5,

the mean is 3.

A student writes

$$s = \sqrt{\frac{(1-3)^2 + (3-3)^2 + (5-3)^2}{3}}.$$

What is wrong?

For the chapter's **sample** standard deviation, divide by

$$n - 1,$$

not n .

Correct:

$$s = \sqrt{\frac{4 + 0 + 4}{2}} = \sqrt{4} = \boxed{2}.$$

The chapter in one slide

$$\text{relative frequency} = \frac{f}{n}$$

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

$$\text{range} = \text{max} - \text{min}$$

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

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

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

$$68\% - 95\% - 99.7\%$$

$$s \approx \frac{\text{range}}{4} \text{ for bell-shaped data}$$

The chapter in one slide

$$\text{relative frequency} = \frac{f}{n}$$

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

$$\text{range} = \text{max} - \text{min}$$

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

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

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

$$68\% - 95\% - 99.7\%$$

$$s \approx \frac{\text{range}}{4} \text{ for bell-shaped data}$$

And for grouped data:

$$\bar{x} \approx \frac{\sum f(\text{midpoint})}{\sum f}$$

A practical decision tree

Given a data problem, ask:

- 1 Is the data raw, grouped, or graphical?

A practical decision tree

Given a data problem, ask:

- ① Is the data raw, grouped, or graphical?
- ② Am I being asked about **shape**, **center**, or **spread**?

A practical decision tree

Given a data problem, ask:

- ① Is the data raw, grouped, or graphical?
- ② Am I being asked about **shape**, **center**, or **spread**?
- ③ If center: mean or median?

A practical decision tree

Given a data problem, ask:

- ① Is the data raw, grouped, or graphical?
- ② Am I being asked about **shape**, **center**, or **spread**?
- ③ If center: mean or median?
- ④ If spread: range/IQR or standard deviation?

A practical decision tree

Given a data problem, ask:

- ① Is the data raw, grouped, or graphical?
- ② Am I being asked about **shape**, **center**, or **spread**?
- ③ If center: mean or median?
- ④ If spread: range/IQR or standard deviation?
- ⑤ Is the distribution bell-shaped enough for the Bell Curve Rule?

A practical decision tree

Given a data problem, ask:

- ① Is the data raw, grouped, or graphical?
- ② Am I being asked about **shape**, **center**, or **spread**?
- ③ If center: mean or median?
- ④ If spread: range/IQR or standard deviation?
- ⑤ Is the distribution bell-shaped enough for the Bell Curve Rule?
- ⑥ If only grouped data are available, am I finding an **estimate**?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?
- ▶ Did I divide by the correct total for relative frequency?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?
- ▶ Did I divide by the correct total for relative frequency?
- ▶ Did my relative frequencies add to about 1?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?
- ▶ Did I divide by the correct total for relative frequency?
- ▶ Did my relative frequencies add to about 1?
- ▶ Did I use $n - 1$ for sample variance and standard deviation?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?
- ▶ Did I divide by the correct total for relative frequency?
- ▶ Did my relative frequencies add to about 1?
- ▶ Did I use $n - 1$ for sample variance and standard deviation?
- ▶ Did I take the square root at the end when finding s ?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?
- ▶ Did I divide by the correct total for relative frequency?
- ▶ Did my relative frequencies add to about 1?
- ▶ Did I use $n - 1$ for sample variance and standard deviation?
- ▶ Did I take the square root at the end when finding s ?
- ▶ Did I use the Bell Curve Rule only when the shape is appropriate?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I sort before finding the median or quartiles?
- ▶ Did I divide by the correct total for relative frequency?
- ▶ Did my relative frequencies add to about 1?
- ▶ Did I use $n - 1$ for sample variance and standard deviation?
- ▶ Did I take the square root at the end when finding s ?
- ▶ Did I use the Bell Curve Rule only when the shape is appropriate?
- ▶ Did I distinguish range from IQR?

Final practice

Data:

5, 7, 7, 8, 10, 11, 15.

Find the median, mean, range, Q_1 , Q_3 , and IQR.

Final practice

Data:

5, 7, 7, 8, 10, 11, 15.

Find the median, mean, range, Q_1 , Q_3 , and IQR.

First three:

$$\text{median} = 8,$$

$$\bar{x} = \frac{63}{7} = 9,$$

$$\text{range} = 15 - 5 = 10.$$

Final practice

Data:

5, 7, 7, 8, 10, 11, 15.

Find the median, mean, range, Q_1 , Q_3 , and IQR.

First three:

$$\text{median} = 8,$$

$$\bar{x} = \frac{63}{7} = 9,$$

$$\text{range} = 15 - 5 = 10.$$

For the quartiles:

5, 7, 7 | 8 | 10, 11, 15.

Final practice

Data:

5, 7, 7, 8, 10, 11, 15.

Find the median, mean, range, Q_1 , Q_3 , and IQR.

First three:

$$\text{median} = 8,$$

$$\bar{x} = \frac{63}{7} = 9,$$

$$\text{range} = 15 - 5 = 10.$$

For the quartiles:

5, 7, 7 | 8 | 10, 11, 15.

Therefore

$$Q_1 = 7,$$

$$Q_3 = 11,$$

$$\text{IQR} = 11 - 7 = 4.$$