Five-minute review

1. Why does a large sample not automatically eliminate bias?

2. Of 50 students surveyed in an advanced class, 40 like math. Can this represent every student?

Write your answers before we discuss them.

Frequency distributions

What we will learn

Organize raw data into frequency and relative-frequency tables.

Why does it matter?

A list of travel times is hard to compare at a glance. Group it into sensible intervals to see where most trips fall.

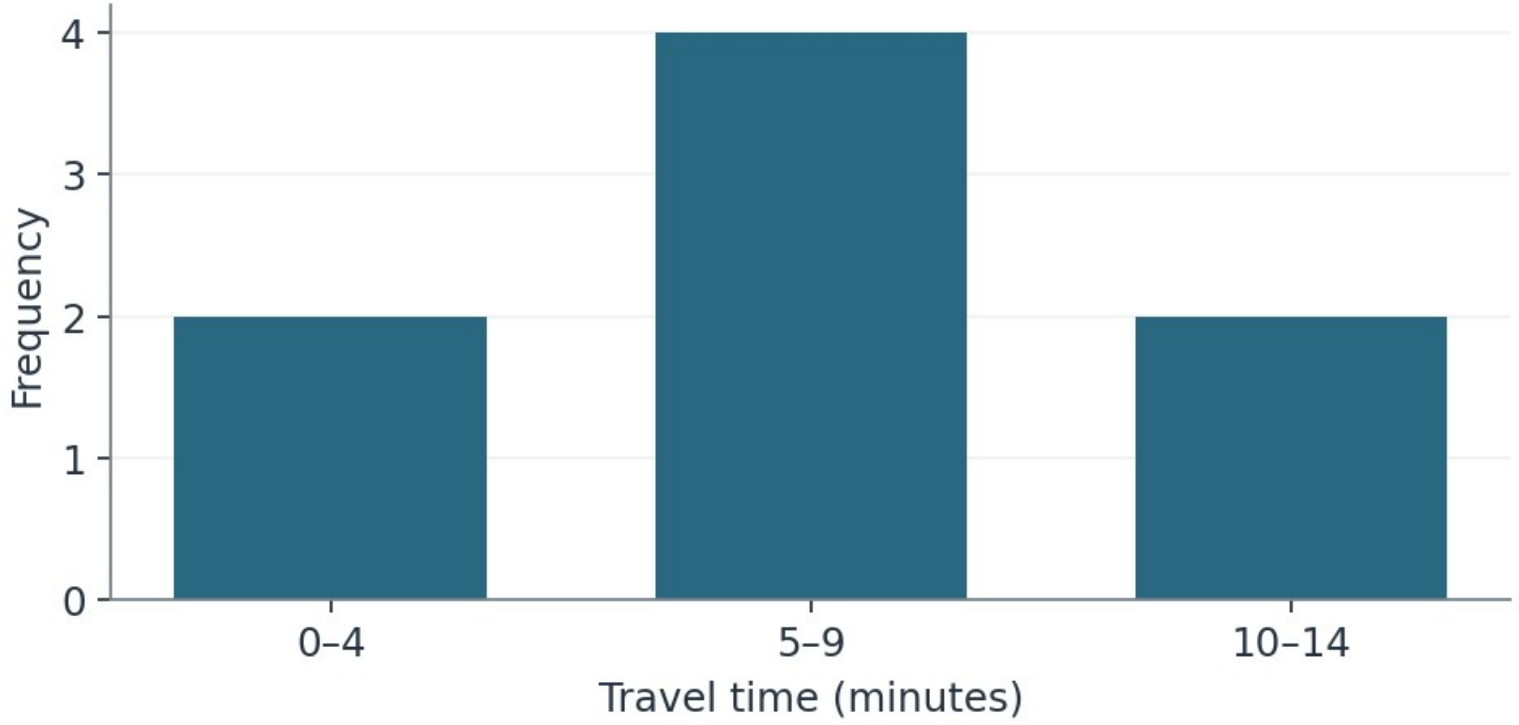
Definitions: Frequency distributions

Frequency: the number of observations in a category or class.

Relative frequency: the frequency divided by the total number of observations.

Cumulative frequency: the total count at or below a class endpoint.

Illustration and interpretation



Notice the pattern

A list of travel times is hard to compare at a glance. Group it into sensible intervals to see where most trips fall.

Illustrative data / model

Formulas

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

$$\text{Midpoint} = \frac{L + U}{2}$$

Important ideas and notation

Frequency f is the count in a class; relative frequency = f/n. Cumulative frequency counts observations up through a class.

Use nonoverlapping classes covering all values. For integer data, width = difference between consecutive lower limits; midpoint = (lower + upper)/2.

Check that frequencies total n and relative frequencies total 1 (up to rounding).

Example 2.1.1

Data: 2, 4, 5, 7, 8, 9, 12, 14. Classes 0–4, 5–9, 10–14. Construct a frequency table, including relative and cumulative frequencies.

Solution:

Frequencies 2, 4, 2

relative frequencies .25, .50, .25

cumulative counts 2, 6, 8.

Example 2.1.2

Classes 10–19, 20–29, 30–39: find width and midpoints.

Solution:

Width 10

midpoints 14.5, 24.5, 34.5.

Example 2.1.3 • Travel times to school

What fraction of the trips took 5–9 minutes?

Minutes	Frequency
0–4	2
5–9	4
10–14	2
Total	8

Illustrative classroom dataset

Example 2.1.3 • Solution

What fraction of the trips took 5–9 minutes?

Solution:

$\frac{4}{8} = 0.50$, or 50%.

Individual practice

1. Data 1, 1, 3, 4, 6, 8. Count 0–2, 3–5, 6–8.

2. If f = 12 and n = 40, find relative frequency.

3. Frequencies 5, 9, 6: find cumulative frequencies.

Group activity

Make a frequency table for 3, 4, 4, 6, 9, 11, 14, 17, 18, 19 using classes of width 5 starting at 0.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. 2, 2, 2.

2. .30.

3. 5, 14, 20.

Group solution

Solution:

Counts 3, 2, 2, 3

relative frequencies .30, .20, .20, .30.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Why must classes not overlap?

2. Frequencies are 3, 5 and 2 in consecutive classes. Find the relative frequency of the second class and its cumulative frequency.

Show your reasoning. Submit before leaving.

Five-minute review

1. Why must classes not overlap?

2. Data 1, 1, 3, 4, 6, 8. Count 0–2, 3–5, 6–8.

Write your answers before we discuss them.

Histograms, dot plots and stems

What we will learn

Construct and interpret numerical data displays.

Why does it matter?

A stem plot keeps individual scores visible, while a histogram makes the overall shape easier to see.

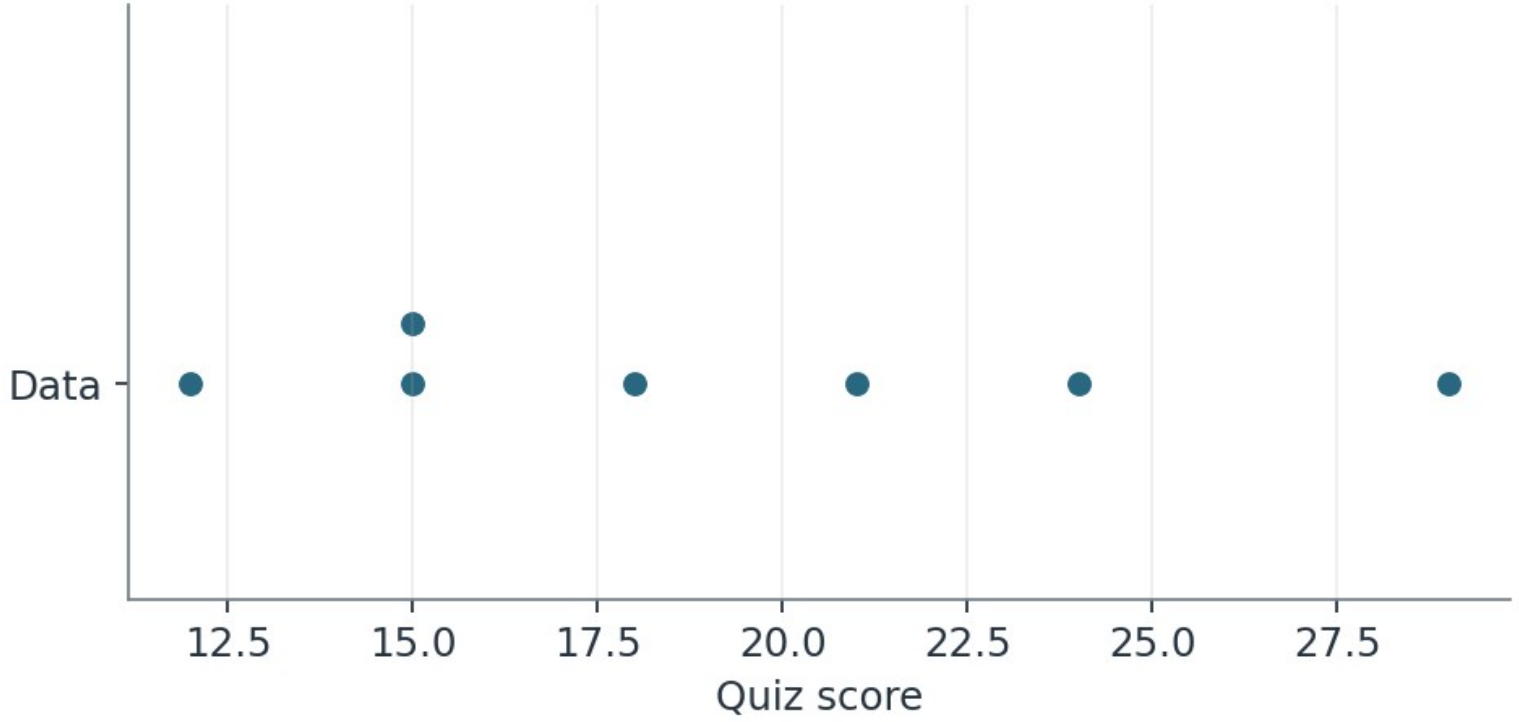
Definitions: Histograms dot plots and stems

Histogram: a graph of numerical intervals whose bars touch (for equal-width bins, height can represent frequency).

Dot plot: dots mark individual observations along a number line.

Stem-and-leaf plot: each value is split into a stem and a leaf; a key specifies its place value.

Illustration and interpretation



Notice the pattern

A stem plot keeps individual scores visible, while a histogram makes the overall shape easier to see.

Illustrative data / model

Important ideas and notation

A histogram shows counts or relative frequencies in adjacent numerical intervals. Equal-width bins allow frequency comparison by bar height.

A dot plot uses one dot per observation; a stem-and-leaf plot preserves exact values with a key and sorted leaves.

Describe shape, center, spread, gaps and unusual values. A skewed distribution has a longer tail in the named direction.

Example 2.2.1

For scores 12, 15, 15, 18, 21, 24, 29, make a stem-and-leaf plot.

Solution:

1 : 2 5 5 8

2 : 1 4 9. Key: 1 : 2 = 12.

Example 2.2.2

Counts by interval 0–9, 10–19, 20–29, 30–39: 9, 6, 3, 1. Describe shape.

Solution:

Right-skewed: larger frequencies on the left, tail to the right.

Example 2.2.3 • Scores on a short quiz

How many scores were in the teens?

Stem	Leaves
1	2 5 5 8
2	1 4 9
Key	1 : 2 means 12

Illustrative classroom dataset

Example 2.2.3 • Solution

How many scores were in the teens?

Solution:

Four: 12, 15, 15 and 18.

Individual practice

1. Draw a dot plot for 2, 2, 3, 5, 5, 5.

2. Make stems for 31, 32, 37, 40, 45.

3. Why are histogram bars adjacent?

Group activity

Create a histogram from classes 0–4, 5–9, 10–14, 15–19 with counts 2, 6, 6, 2. Describe the shape.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. Two dots at 2, one at 3, three at 5.

2. 3 : 1 2 7; 4 : 0 5; key 3 : 1 = 31.

3. Bins cover a numerical continuum with neighboring boundaries.

Group solution

Solution:

Symmetric, not necessarily normal

bars of heights 2, 6, 6, 2.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Does symmetric always mean normal?

2. Construct a stem-and-leaf plot for 21, 24, 24, 31, 35. Include a key.

Show your reasoning. Submit before leaving.

Five-minute review

1. Does symmetric always mean normal?

2. Draw a dot plot for 2, 2, 3, 5, 5, 5.

Write your answers before we discuss them.

Choosing graphs and detecting distortions

What we will learn

Select an appropriate display and critique distorted graphs.

Why does it matter?

A chart can contain correct numbers and still give a distorted impression if the scale is misleading.

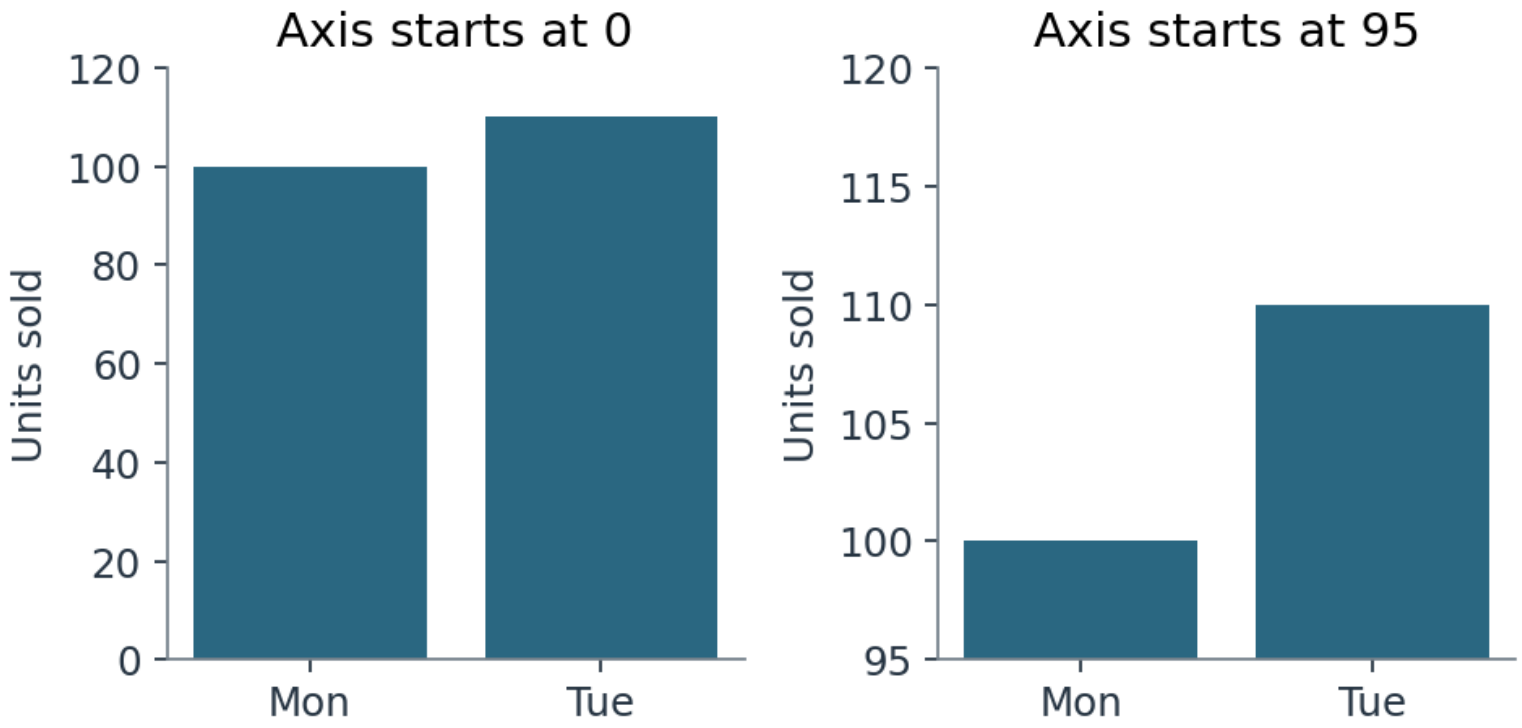
Definitions: Choosing graphs and detecting distortions

Bar graph: compares categories with separated bars.

Pie chart: shows mutually exclusive parts of a whole.

Time-series graph: displays measurements in chronological order.

Illustration and interpretation



Notice the pattern

A chart can contain correct numbers and still give a distorted impression if the scale is misleading.

Illustrative data / model

Important ideas and notation

Bar graphs compare categories; pie charts show parts of a whole; time-series graphs track changes over time. Use clear labels, units and scales.

A frequency polygon joins class midpoint frequencies. A pictograph needs a consistent symbol key.

Truncated bar axes, unequal scales, misleading area and selective time windows can exaggerate differences. Check source and context.

Example 2.3.1

A store sold 100 units Monday and 110 Tuesday. A bar graph begins at 95.

Solution:

Bar lengths appear 5 and 15, suggesting triple rather than a 10% increase. Start at zero.

Example 2.3.2

A survey reports 30 bus, 10 walk, 10 bike. Pie angle for bus?

Solution:

$\frac{30}{50} \times 360° = 216°.$

Example 2.3.3 • Two versions of the same sales chart

What is the actual percentage increase?

Day	Units
Monday	100
Tuesday	110

Illustrative classroom dataset

Example 2.3.3 • Solution

What is the actual percentage increase?

Solution:

$$\frac{110 - 100}{100} = 10\%.$$

Individual practice

1. Best graph for monthly rainfall over one year?

2. Best graph for favorite school subject?

3. Find percent increase from 80 to 88.

Group activity

Sketch two bar graphs of 45 and 50 using axes starting at 0 and 40. Compare impressions.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. Time-series line graph.

2. Bar graph (or pie chart for shares).

3. 10%.

Group solution

Solution:

Zero baseline shows 11.1% increase

truncated bars exaggerate the difference.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Name two graph features to verify.

2. A graph shows rainfall by month. Which graph would best reveal changes over the year?

Show your reasoning. Submit before leaving.

Five-minute review

1. Name two graph features to verify.

2. Best graph for monthly rainfall over one year?

Write your answers before we discuss them.

Mean, median, mode and midrange

What we will learn

Calculate and interpret four measures of center.

Why does it matter?

The mean shares the total equally; the median tells us the middle. A very long reading session can pull the mean upward.

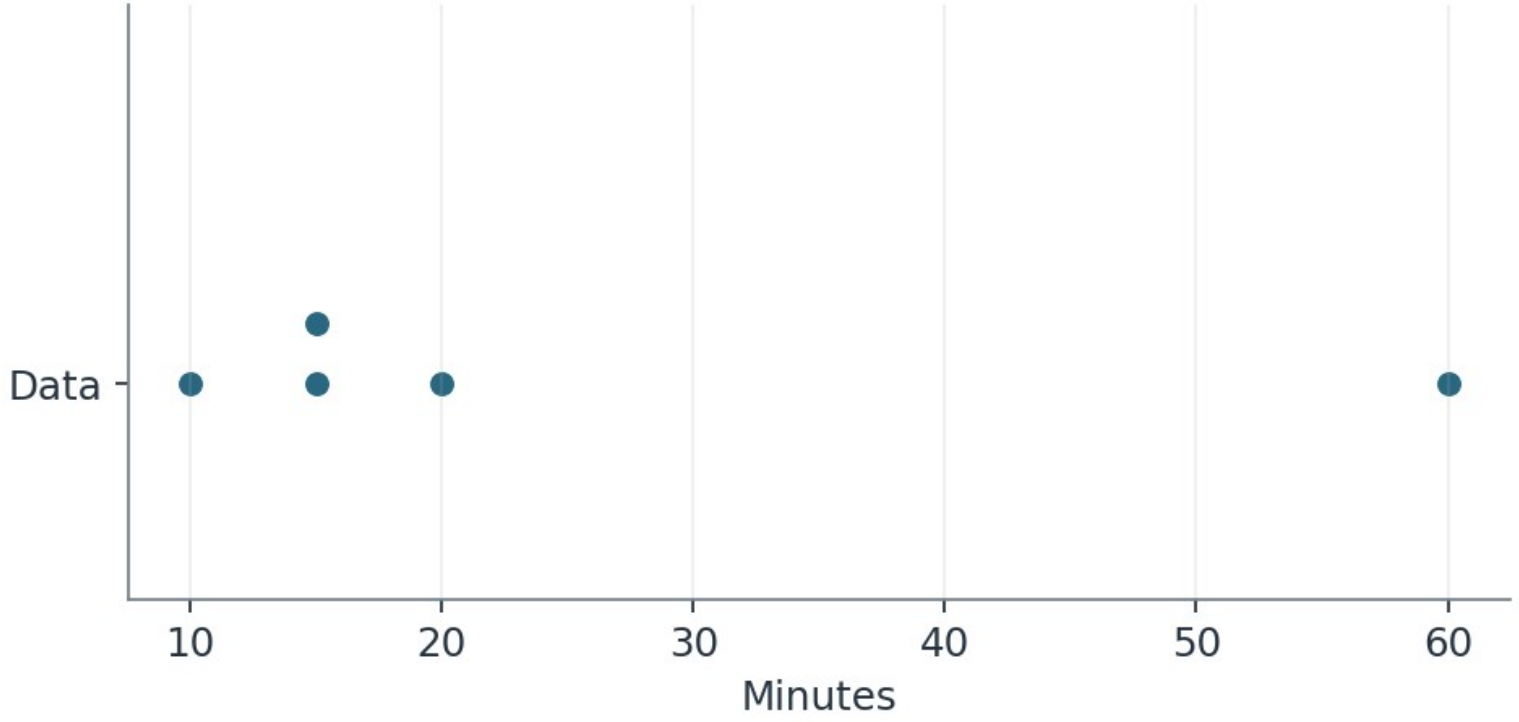
Definitions: Mean median mode and midrange

Mean: the sum of observations divided by their number.

Median: the middle value after ordering, or the average of the two middle values.

Mode: the most frequent value; midrange: the average of minimum and maximum.

Illustration and interpretation



Notice the pattern

The mean shares the total equally; the median tells us the middle. A very long reading session can pull the mean upward.

Illustrative data / model

Formulas

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{Midrange} = \frac{x_{\min} + x_{\max}}{2}$$

Important ideas and notation

Sample mean $\bar{x} = \Sigma x/n$; population mean $\mu = \Sigma x/N$. Mean uses every numerical value.

Median is the middle ordered value or average of the two middle values. Mode is most frequent; may have ties or no mode under the classroom convention.

Midrange = (minimum + maximum)/2. Choose a measure appropriate for the variable and shape.

Example 2.4.1

Data 4, 6, 6, 8, 11. Find all four measures.

Solution:

$\text{Mean} = \frac{35}{5} = 7$

median = 6

mode = 6

midrange = 7.5.

Example 2.4.2

Data 3, 8, 10, 15. Find the mean, median, mode and midrange.

Solution:

$$\text{Mean} = \frac{36}{4} = 9$$

$$\text{median} = \frac{8+10}{2} = 9$$

no mode

midrange = 9.

Example 2.4.3 • Typical daily reading time

Find the median and explain whether it is typical.

Day	Minutes
Mon	10
Tue	15
Wed	15
Thu	20
Fri	60

Illustrative classroom dataset

Example 2.4.3 • Solution

Find the median and explain whether it is typical.

Solution:

15 minutes

the mean is 24, pulled up by the 60-minute value.

Individual practice

1. Find mean, median, mode of 2, 4, 4, 6, 9.

2. Find median of 10, 3, 8, 7.

3. Find midrange of 12, 15, 21, 24.

Group activity

Find a five-value data set with mean 10 and median 8.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. 5, 4, 4.

2. 7.5.

3. 18.

Group solution

Solution:

Example 4, 6, 8, 12, 20

total 50.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Which measure can describe favorite colors?

2. Find the mean, median, mode and midrange of 2, 4, 4, 6, 9.

Show your reasoning. Submit before leaving.

Five-minute review

1. Which measure can describe favorite colors?

2. Find mean, median, mode of 2, 4, 4, 6, 9.

Write your answers before we discuss them.

Grouped mean and effects of unusual values

What we will learn

Estimate a grouped mean and explain sensitivity to outliers.

Why does it matter?

Suppose the original travel times were deleted and only the grouped counts remain. The midpoint gives a reasonable estimate, not the exact mean.

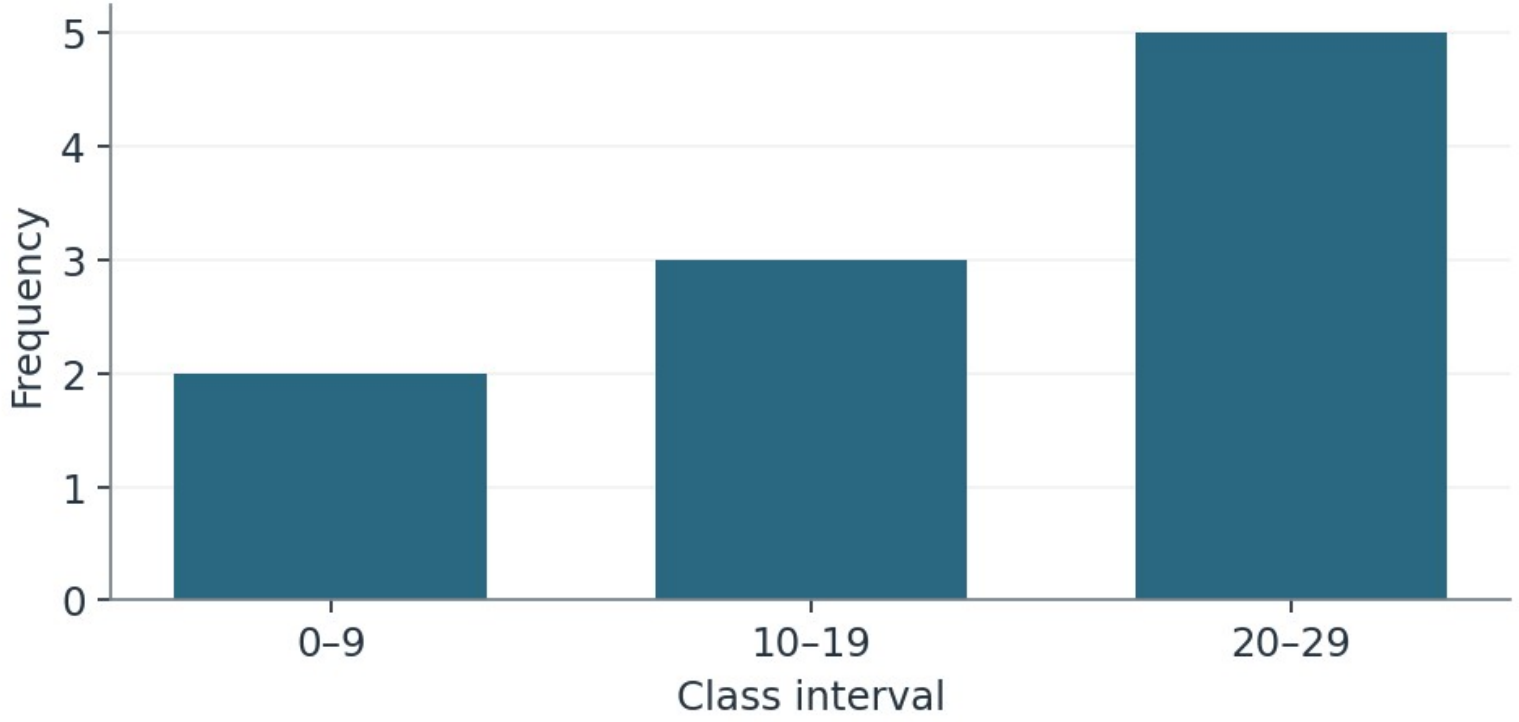
Definitions: Grouped mean and effects of unusual values

Grouped data: observations summarized in intervals rather than listed individually.

Class midpoint: the average of a class's lower and upper limits.

Weighted mean: an average that accounts for how often each value occurs.

Illustration and interpretation



Notice the pattern

Suppose the original travel times were deleted and only the grouped counts remain. The midpoint gives a reasonable estimate, not the exact mean.

Illustrative data / model

Formulas

$$\bar{x} \approx \frac{\sum f_i m_i}{\sum f_i}$$

Important ideas and notation

Grouped mean ≈ Σ(fm)/Σf, where m is a class midpoint. It is approximate because individual values within bins are unknown.

The mean and midrange are sensitive to extremes; median is more resistant. Skewness can pull the mean toward the long tail, but mean–median–mode order is not a universal rule.

Using the method correctly

Adding c to every value shifts the mean and median by c; multiplying by a positive c scales both. Shape and context determine suitable summaries.

Example 2.5.1

Classes 0–9, 10–19, 20–29; f = 2, 3, 5. Estimate the mean using class midpoints.

Solution:

Midpoints 4.5, 14.5, 24.5

weighted sum = 175

grouped mean ≈17.5.

Example 2.5.2

Data 2, 4, 6, 8, 10: replace 10 by 50. Compare the original and revised mean and median.

Solution:

Mean goes from 6 to 14

median stays 6. The mean is more affected.

Example 2.5.3 • When individual values are missing

Why is the grouped mean only an estimate?

Minutes	Midpoint	Count
0–9	4.5	2
10–19	14.5	3
20–29	24.5	5

Illustrative classroom dataset

Example 2.5.3 • Solution

Why is the grouped mean only an estimate?

Solution:

Each observation is replaced by its class midpoint, so original values are unknown.

Individual practice

1. Classes 10–19 and 20–29; counts 4 and 6.

2. Add 3 to each value when mean = 12 and median = 10.

3. For strongly right-skewed income data, which center is often better?

Group activity

Compare 5, 6, 7, 8, 9 with 5, 6, 7, 8, 49. Calculate center and recommend a measure.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. Grouped mean ≈20.5.

2. New mean 15; median 13.

3. Median.

Group solution

Solution:

Means 7 and 15

medians both 7. Median better describes typical value in second set.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Why is grouped mean an estimate?

2. Midpoints 5 and 15 have frequencies 3 and 7. Estimate the mean.

Show your reasoning. Submit before leaving.

Five-minute review

1. Why is grouped mean an estimate?

2. Classes 10–19 and 20–29; counts 4 and 6.

Write your answers before we discuss them.

Range, MAD and variance

What we will learn

Measure spread and distinguish sample and population variance.

Why does it matter?

Two groups can have identical averages but very different levels of consistency. Deviations show how far each observation is from the center.

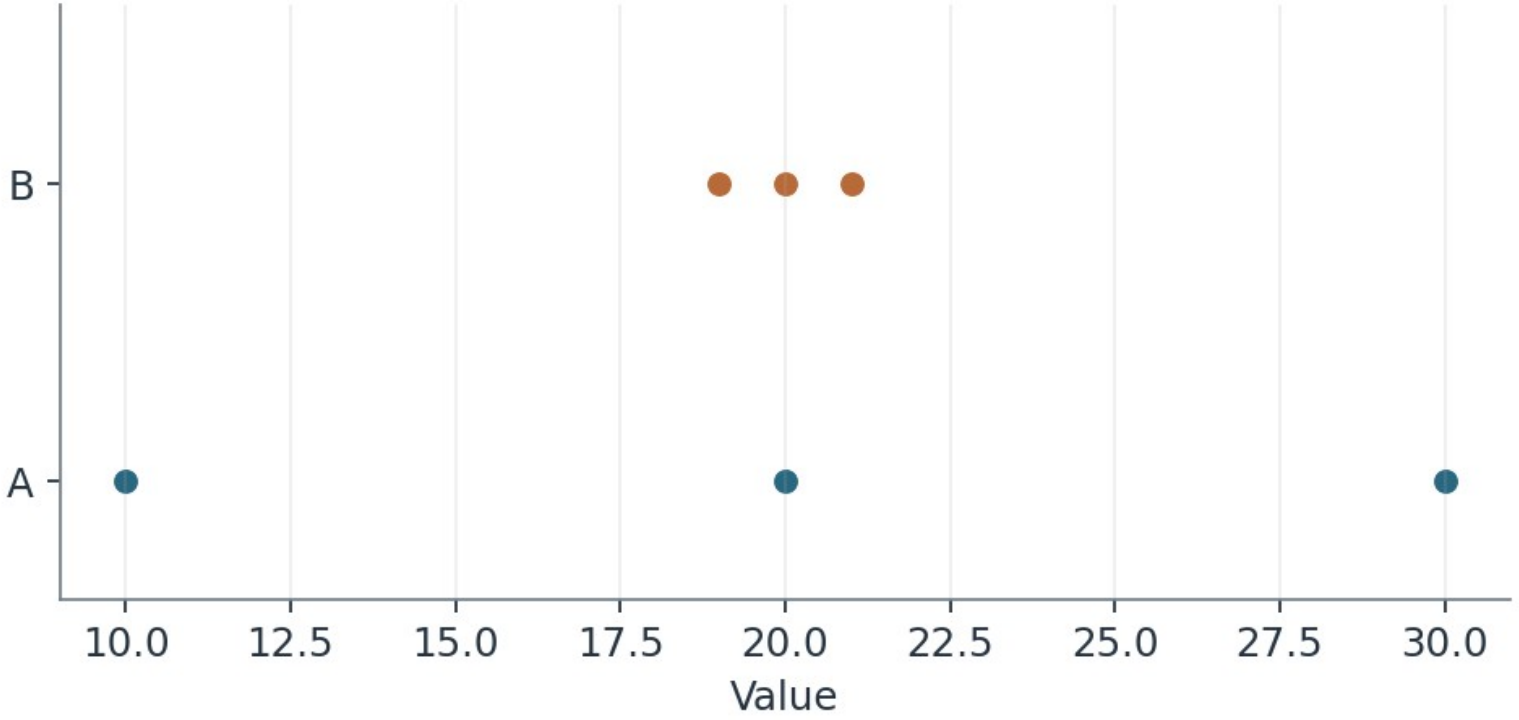
Definitions: Range mad and variance

Range: maximum minus minimum.

Mean absolute deviation (MAD): the mean of the absolute distances from the mean.

Variance: average squared deviation for a population; sample variance divides the squared-deviation sum by n−1.

Illustration and interpretation



Notice the pattern

Two groups can have identical averages but very different levels of consistency. Deviations show how far each observation is from the center.

Illustrative data / model

Formulas

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}, \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Important ideas and notation

Range = max − min. Mean absolute deviation (MAD) = Σabs(x−mean)/n; it is the average absolute distance from the mean.

Population variance $\sigma^2 = \Sigma(x-\mu)^2/N$. Sample variance $s^2 = \Sigma(x-\bar{x})^2/(n-1)$, for $n > 1$.

Squared deviations prevent cancellation. Variance has squared units; standard deviation has original units.

Example 2.6.1

Data 2, 4, 6 as a population. Find the mean, range, MAD and population variance.

Solution:

Mean 4

range 4

$$\text{MAD} = \frac{2+0+2}{3} = \frac{4}{3}$$

$$\sigma^2 = \frac{4+0+4}{3} = \frac{8}{3}.$$

Example 2.6.2

Treat 2, 4, 6 as a sample. Calculate the sample variance.

Solution:

$\bar{x} = 4$

squared deviations total 8

$s^2 = \frac{8}{2} = 4.$

Example 2.6.3 • Same center, different spread

Which has larger range and MAD?

Group	Scores	Mean
A	10, 20, 30	20
B	19, 20, 21	20

Illustrative classroom dataset

Example 2.6.3 • Solution

Which has larger range and MAD?

Solution:

A: range 20, MAD $\frac{20}{3}$

B: range 2, MAD $\frac{2}{3}$.

Individual practice

1. Find range of 6, 8, 10, 18.

2. Find MAD of 3, 5, 7, 9.

3. Sample 4, 6, 8: find variance.

Group activity

Compare A: 10, 20, 30 and B: 19, 20, 21 using sample variance.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. 12.

2. Mean 6; MAD 2.

3. $s^2 = 4$.

Group solution

Solution:

A: 100

B: 1. B is more consistent.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Why do squared deviations not cancel?

2. Treat 3, 5, 7 as the entire population. Find the range, MAD and variance.

Show your reasoning. Submit before leaving.

Quick review — 5 minutes

1. Find the mean and range of 6, 8, 10, 18.

2. Why do we square deviations when calculating variance?

Lesson 2.7: Standard deviation

What we will learn

- Calculate population and sample standard deviation.
- Interpret standard deviation in the units of the data.
- Compare two data sets with the same mean.

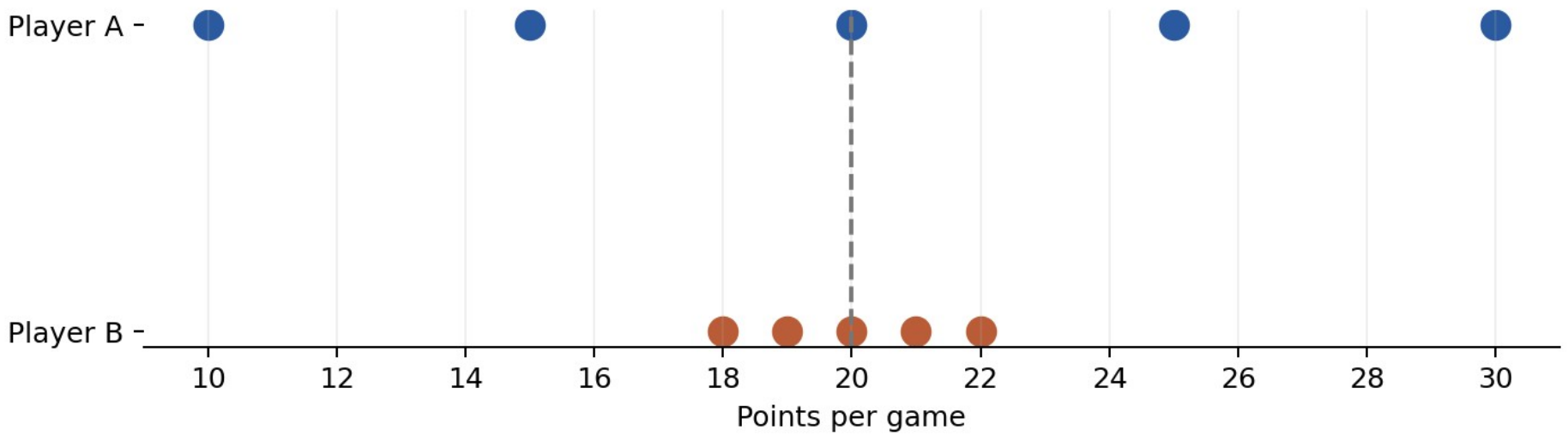
Why does it matter?

Two basketball players average the same points per game. Which player is more consistent?

Same mean, different variability

The points scored in five games are shown below. Both players average 20 points.

Player	Game 1	Game 2	Game 3	Game 4	Game 5
A	10	15	20	25	30
B	18	19	20	21	22



Definition: Standard deviation

Standard deviation measures the spread of data values around their mean. It is the positive square root of variance (or zero if all values are equal).

Population standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Sample standard deviation

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

N = population size; n = sample size. Standard deviation has the same units as the original data.

Example 2.7.1: Calculating sample standard deviation

Five quiz scores are 5, 12, 6, 8 and 14. Find the sample variance and sample standard deviation.

Solution: Step 1. Find the sample mean.

$$\bar{x} = \frac{5 + 12 + 6 + 8 + 14}{5} = \frac{45}{5} = 9$$

Step 2. Subtract the mean from each observation, then square the result.

Example 2.7.1: Squared deviations

Score	Deviation from mean	Squared deviation
5	$5 - 9 = -4$	16
12	$12 - 9 = 3$	9
6	$6 - 9 = -3$	9
8	$8 - 9 = -1$	1
14	$14 - 9 = 5$	25
Total	0	60

The squared deviations add to 60. The ordinary deviations add to zero.

Example 2.7.1: Variance and standard deviation

Step 3. Divide the sum of squared deviations by n − 1.

$$s^2 = \frac{60}{5-1} = \frac{60}{4} = 15$$

Step 4. Take the square root of the variance.

$$s = \sqrt{15} \approx 3.87 \text{ points}$$

Interpretation: the quiz scores typically vary from their mean on a scale of about 3.87 points.

Example 2.7.2: Comparing the two players

Treat the five games for each player as a sample. Both means equal 20.

$$s_A = \sqrt{\frac{100 + 25 + 0 + 25 + 100}{4}} = \sqrt{62.5} \approx 7.91$$

$$s_B = \sqrt{\frac{4 + 1 + 0 + 1 + 4}{4}} = \sqrt{2.5} \approx 1.58$$

Player B is more consistent because the sample standard deviation is smaller.

Example 2.7.3: Coefficient of variation

Group A has mean 50 and standard deviation 5. Group B has mean 100 and standard deviation 8. Which has smaller relative spread?

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

$$CV_A = \frac{5}{50} \times 100\% = 10\%$$

$$CV_B = \frac{8}{100} \times 100\% = 8\%$$

Group B has smaller relative variability. Use CV for suitable positive ratio-scale data.

Individual practice

1. Find the sample variance and standard deviation of 3, 5, 7, 9, 11.

2. A sample variance is 36. Find its standard deviation.

3. If all scores increase by 5, what happens to standard deviation?

Group activity

Two teams record weekly practice hours: A: 2, 4, 6, 8, 10; B: 4, 5, 6, 7, 8.

Calculate each sample mean and standard deviation. Which team has a more consistent schedule?

Solutions

1. Sample variance = 10; standard deviation ≈ 3.16.

2. Standard deviation = 6.

3. Standard deviation is unchanged.

Group: means = 6; sample standard deviations ≈ 3.16 and 1.58. Team B is more consistent.

Exit ticket — 5 minutes

1. Can standard deviation be zero? Explain.

2. A sample variance is 49 square centimeters. Find the standard deviation and give its units.

Five-minute review

1. Can SD be negative or zero?

2. If $s^2 = 36$, find s.

Write your answers before we discuss them.

Grouped variability and technology

What we will learn

Estimate grouped variance and verify calculations with technology.

Why does it matter?

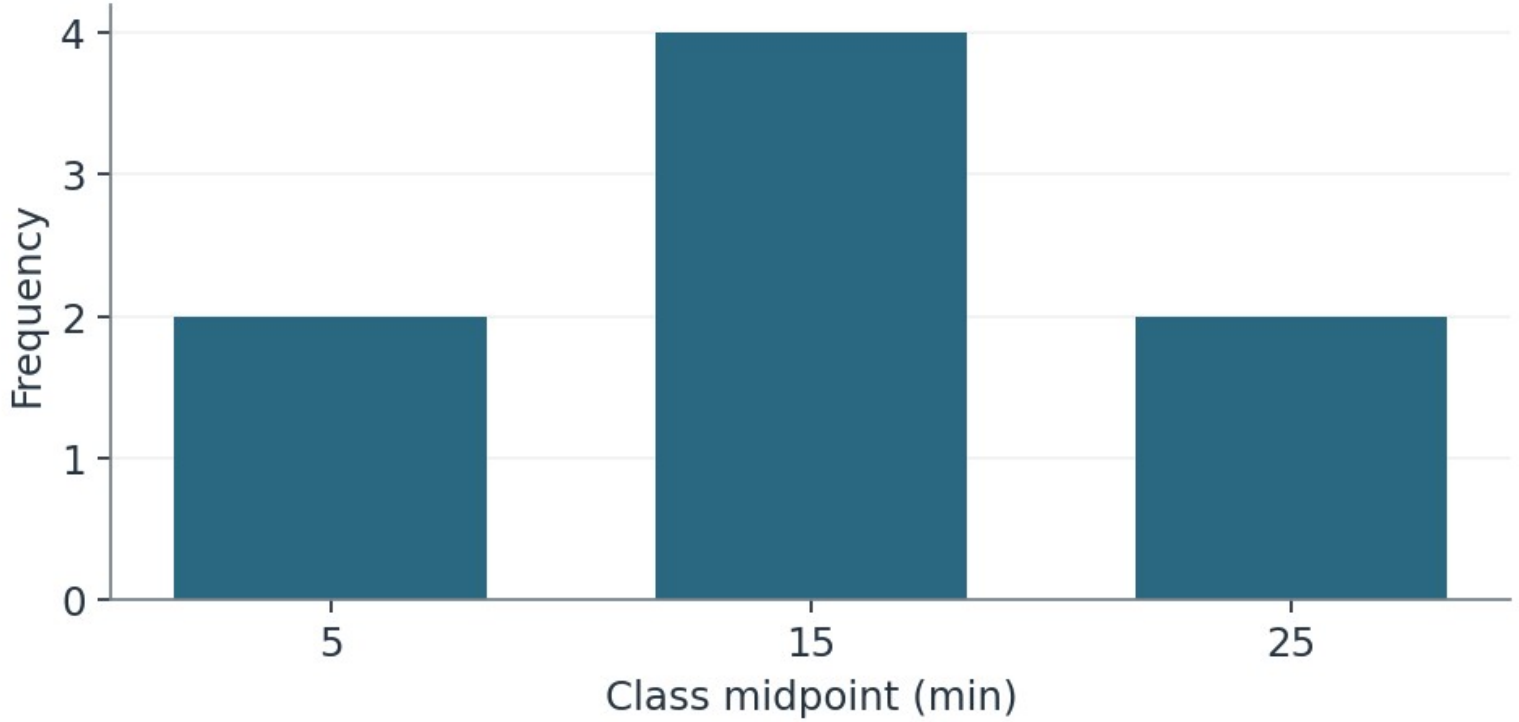
When only counts are available, use the midpoints as approximate observations. Technology makes the repeated calculations manageable.

Definitions: Grouped variability and technology

Grouped standard deviation: an approximate standard deviation calculated by treating every value in a class as its midpoint.

Frequency weight: the number of observations represented by a midpoint.

Illustration and interpretation



Notice the pattern

When only counts are available, use the midpoints as approximate observations. Technology makes the repeated calculations manageable.

Illustrative data / model

Formulas

$$\bar{x} \approx \frac{\sum f_i m_i}{n}$$

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

$$s^2 \approx \frac{\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{n}}{n - 1}$$

Important ideas and notation

Use midpoints m and frequencies f: $\bar{x} \approx \Sigma fm/n$; $s^2 \approx \Sigma f(m-\bar{x})^2/(n-1)$. For population variance divide by N.

Equivalent sample formula: $s^2 \approx [\Sigma fm^2 - (\Sigma fm)^2/n]/(n-1)$. Grouped results are approximate.

Spreadsheet functions AVERAGE, STDEV.S and STDEV.P calculate raw-data summaries. Use frequency weights, not unweighted midpoints, for grouped data.

Example 2.8.1

Midpoints 5, 15, 25; frequencies 2, 4, 2. Estimate the sample variance and standard deviation.

Solution:

$n=8$

$\bar{x}=15$

$\Sigma f(m-15)^2=400$

$s^2\approx 57.14$

$s\approx 7.56.$

Example 2.8.2

Raw data 2, 4, 6, 8: compare spreadsheet sample and population SD.

Solution:

STDEV.S ≈2.58

STDEV.P ≈2.24.

Example 2.8.3 • A grouped delivery-time report

How many midpoint observations would a spreadsheet contain?

Midpoint (min)	Frequency
5	2
15	4
25	2

Illustrative classroom dataset

Example 2.8.3 • Solution

How many midpoint observations would a spreadsheet contain?

Solution:

Eight (2 + 4 + 2).

Individual practice

1. Midpoints 10, 20; f=2, 2. Find sample variance.

2. For the same values treated as population, variance?

3. Why is using STDEV.S on two midpoints alone wrong?

Group activity

Build a spreadsheet expanding midpoint 10 three times and 20 three times; compute sample SD.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. $\frac{100}{3} \approx 33.33$.

2. 25.

3. It ignores the frequencies.

Group solution

Solution:

Mean 15

$$s=\sqrt{(\frac{150}{5})}=\sqrt{30}\approx 5.48.$$

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. What is lost when data are grouped?

2. Midpoints 10 and 20 each have frequency 2. Estimate the population SD.

Show your reasoning. Submit before leaving.

Five-minute review

1. What is lost when data are grouped?

2. Midpoints 10, 20; f=2, 2. Find sample variance.

Write your answers before we discuss them.

Percentiles, quartiles and IQR

What we will learn

Compute quartiles, percentile rank and interquartile range.

Why does it matter?

Quartiles divide ordered data into four sections and help describe the middle half of the distribution.

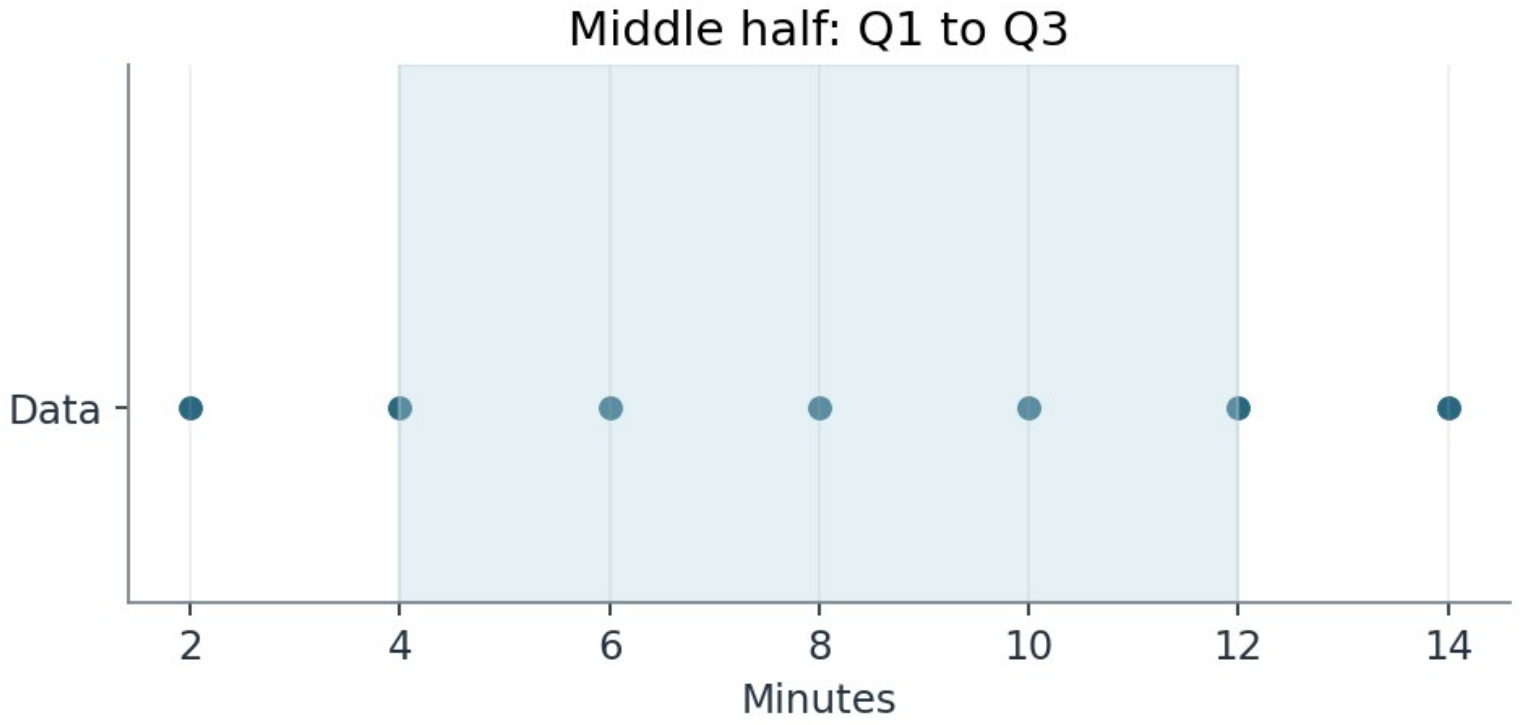
Definitions: Percentiles quartiles and iqr

Percentile rank: the percentage of observations at or below a specified value under our stated convention.

Quartiles: Q1, Q2 and Q3 mark the lower quarter, median and upper quarter positions.

Interquartile range (IQR): Q3 − Q1, the spread of the middle half.

Illustration and interpretation



Notice the pattern

Quartiles divide ordered data into four sections and help describe the middle half of the distribution.

Illustrative data / model

Formulas

$$IQR = Q_3 - Q_1$$

$$\text{Percentile rank} = 100 \times \frac{\text{Number at or below } x}{n}$$

Important ideas and notation

Order data first. In this course Q1 and Q3 are medians of lower and upper halves, excluding the overall median for odd n. Software conventions may differ.

IQR = Q3 − Q1. It measures spread of the middle half of the data. Percentile rank convention here = 100 × (number at or below x)/n.

A percentile rank is relative position, not percent correct. Ties can affect percentile definitions.

Example 2.9.1

Ordered data 2, 4, 6, 8, 10, 12, 14. Find the median, Q1, Q3 and interquartile range.

Solution:

Median=8

lower half 2,4,6 gives Q1=4

upper half 10,12,14 gives Q3=12

IQR=8.

Example 2.9.2

In 20 ordered scores, 15 are at or below 82.

Solution:

$$\text{Percentile rank} = 100 \times \frac{15}{20} = 75\text{th by stated convention.}$$

Example 2.9.3 • Ranking practice times

What is the middle 50% spread using the median-of-halves convention?

Part	Values (min)
Sorted times	2, 4, 6, 8, 10, 12, 14
Lower half	2, 4, 6
Upper half	10, 12, 14

Illustrative classroom dataset

Example 2.9.3 • Solution

What is the middle 50% spread using the median-of-halves convention?

Solution:

$Q3 - Q1 = 12 - 4 = 8$ minutes.

Individual practice

1. Find Q1, median, Q3 of 1, 3, 5, 7, 9, 11.

2. If Q1=25 and Q3=60, find IQR.

3. Of 40 values, 32 are at or below 18.

Group activity

Compare quartiles for 1, 2, 3, 4, 5, 6, 7 and 1, 2, 3, 4, 5, 6, 70.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. 3, 6, 9.

2. 35.

3. 80th percentile rank.

Group solution

Solution:

Both Q1=2, median=4, Q3=6, IQR=4

max changes.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Why specify a quartile convention?

2. Find Q1, Q3 and IQR for 3, 5, 7, 9, 11, 13, 15, 17.

Show your reasoning. Submit before leaving.

Five-minute review

1. Why specify a quartile convention?

2. Find Q1, median, Q3 of 1, 3, 5, 7, 9, 11.

Write your answers before we discuss them.

Outliers, five-number summaries and box plots

What we will learn

Construct modified box plots and compare distributions.

Why does it matter?

A single long trip can change the range dramatically. A modified box plot shows it separately when it passes the IQR rule.

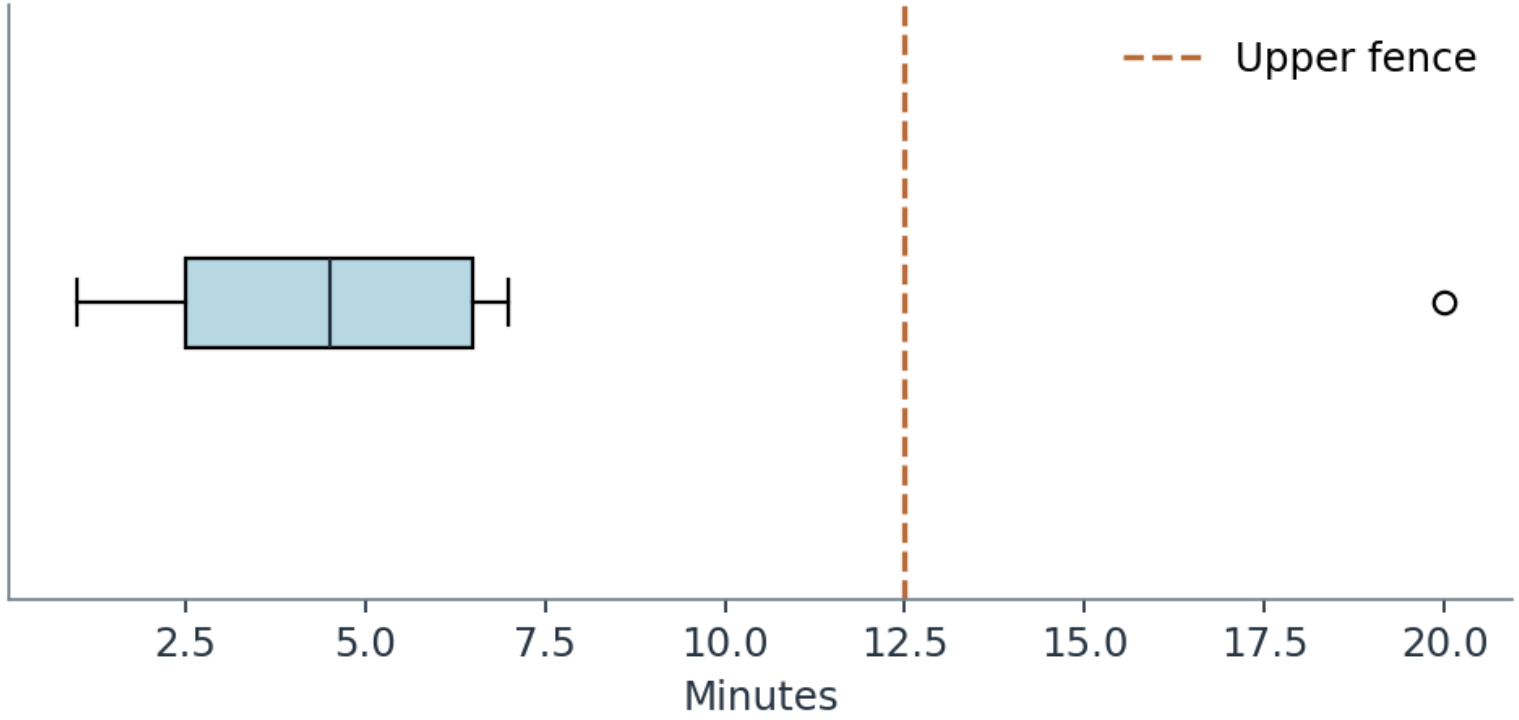
Definitions: Outliers five number summaries and box plots

Five-number summary: minimum, Q1, median, Q3 and maximum.

Outlier flag: an observation below $Q1 - 1.5IQR$ or above $Q3 + 1.5IQR$.

Modified box plot: box and median line, whiskers to extreme non-outliers, with flagged values shown separately.

Illustration and interpretation



Notice the pattern

A single long trip can change the range dramatically. A modified box plot shows it separately when it passes the IQR rule.

Illustrative data / model

Formulas

$$L = Q_1 - 1.5IQR, \quad U = Q_3 + 1.5IQR$$

Important ideas and notation

Five-number summary: min, Q1, median, Q3, max. Fences: Q1−1.5IQR and Q3+1.5IQR; values strictly beyond fences are flagged.

Modified box plot: box from Q1 to Q3; median line; whiskers to extreme non-outliers; plot outliers separately.

Compare center, IQR, overall spread and asymmetry. Investigate outliers rather than automatically deleting them.

Example 2.10.1

Data 1, 2, 3, 4, 5, 6, 7, 20. Find the five-number summary, IQR, fences and modified box plot whiskers.

Solution:

Five-number summary: 1, 2.5, 4.5, 6.5, 20

IQR = 6.5 − 2.5 = 4

Fences: −3.5 and 12.5

Flagged value: 20. Modified whiskers: 1 and 7.

Example 2.10.2

A: Q1=10, med=15, Q3=20; B: Q1=8, med=15, Q3=26. Compare the median and the spread of the middle half.

Solution:

Same median

IQR A=10, B=18

B is more variable in the middle half.

Example 2.10.3 • An unusually long commute

Does the largest value belong beyond the upper fence?

Statistic	Minutes
Q1	2.5
Median	4.5
Q3	6.5
Upper fence	12.5
Largest value	20

Illustrative classroom dataset

Example 2.10.3 • Solution

Does the largest value belong beyond the upper fence?

Solution:

Yes. $20 > 12.5$

the largest non-outlier is 7.

Individual practice

1. Q1=12, Q3=20: find fences.

2. Is 32 an outlier for these fences?

3. What does the box length represent?

Group activity

Draw modified plots of 2, 3, 4, 5, 6, 7, 8 and 2, 3, 4, 5, 6, 7, 30.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. IQR 8; fences 0 and 32.

2. No; it is on the upper fence.

3. IQR.

Group solution

Solution:

Both Q1=3, median=5, Q3=7

fences −3,13

second has outlier 30 and upper whisker 7

first upper whisker 8.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Where does the upper whisker end when there is an outlier?

2. If Q1=12 and Q3=20, find both fences and determine whether 33 is flagged.

Show your reasoning. Submit before leaving.

Five-minute review

1. Where does the upper whisker end when there is an outlier?

2. Q1=12, Q3=20: find fences.

Write your answers before we discuss them.

Scatterplots and correlation

What we will learn

Describe two-variable relationships and interpret correlation.

Why does it matter?

Each dot in a scatterplot represents a pair of measurements for the same student. Keep each pair together.

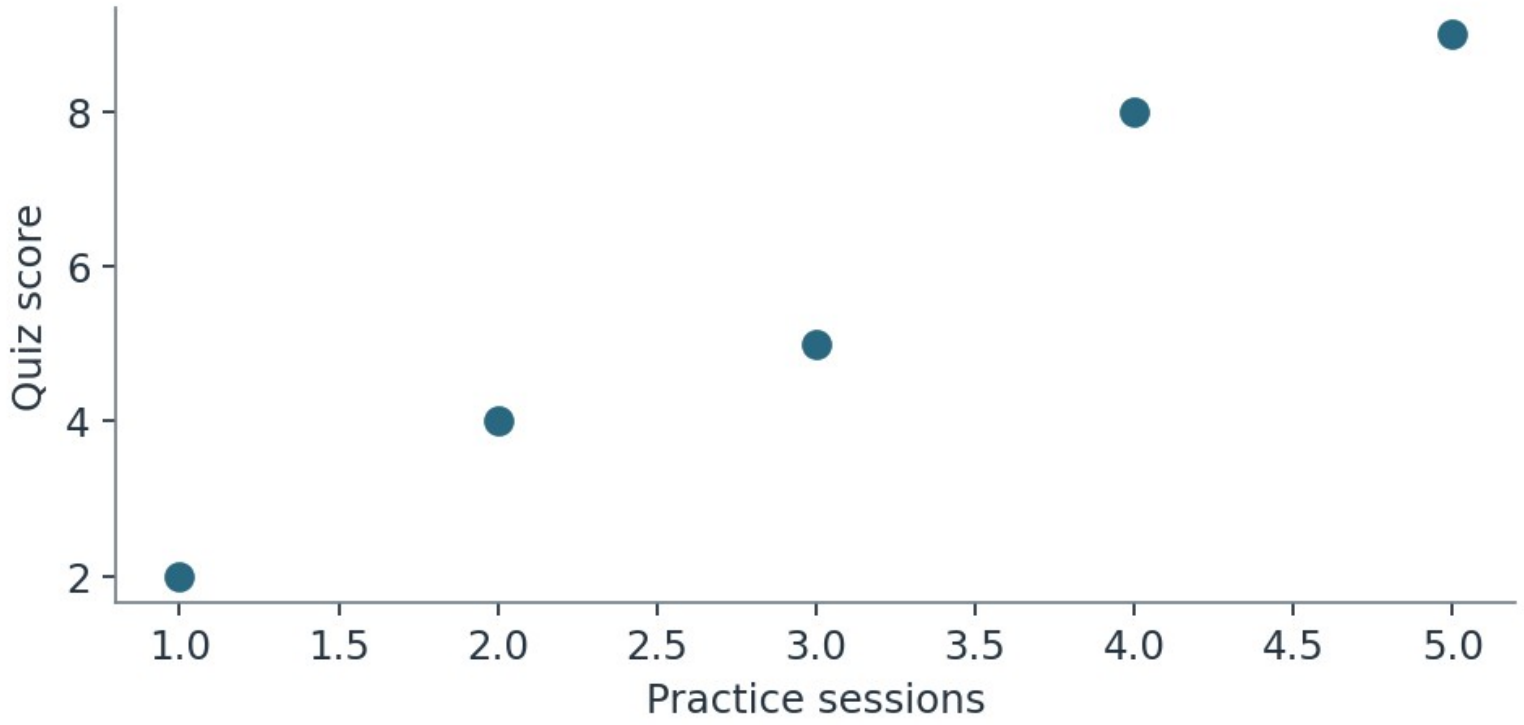
Definitions: Scatterplots and correlation

Scatterplot: each point represents a pair of numerical measurements.

Correlation coefficient r: measures the direction and strength of a linear relationship, from −1 to +1.

Association: a pattern relating two variables; it need not be causal.

Illustration and interpretation



Notice the pattern

Each dot in a scatterplot represents a pair of measurements for the same student. Keep each pair together.

Illustrative data / model

Formulas

$$-1 \leq r \leq 1$$

Important ideas and notation

A scatterplot displays paired numerical values (x,y). Describe direction, form, strength and unusual points.

Pearson correlation r is between −1 and 1; sign gives direction and magnitude describes linear association. r≈0 does not rule out nonlinear association.

Correlation is not causation. Check confounding, outliers and restricted ranges; use technology for r.

Example 2.11.1

Pairs (1,3),(2,5),(3,7),(4,9). Describe the association and find r.

Solution:

Perfect positive linear relationship

r=1.

Example 2.11.2

Pairs (1,8),(2,6),(3,4),(4,2). Describe the association and find r.

Solution:

Perfect negative linear relationship

$r=-1$.

Example 2.11.3 • Practice sessions and quiz scores

Describe direction and form without calculating r.

Sessions	Score
1	2
2	4
3	5
4	8
5	9

Illustrative classroom dataset

Example 2.11.3 • Solution

Describe direction and form without calculating r.

Solution:

Positive, approximately linear
scores tend to rise with sessions.

Individual practice

1. What direction is $r=-.82$?

2. Is $r=1.2$ possible?

3. What can $r=0$ conceal?

Group activity

Plot (1,2),(2,4),(3,5),(4,8),(5,9). Use CORREL on the two columns.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. Strong negative linear association.

2. No.

3. A nonlinear relationship.

Group solution

Solution:

Positive, strong linear association

r≈.988. No causal claim from this alone.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Why can an extreme point change r?

2. The points (1,6), (2,4), (3,2) fall exactly on a line. What is r?

Show your reasoning. Submit before leaving.

Five-minute review

1. Why can an extreme point change r?

2. What direction is r=−.82?

Write your answers before we discuss them.

Regression and predictions

What we will learn

Fit a regression line, predict and interpret residuals.

Why does it matter?

A regression line summarizes a trend. A prediction is not a promise; actual results can be above or below the line.

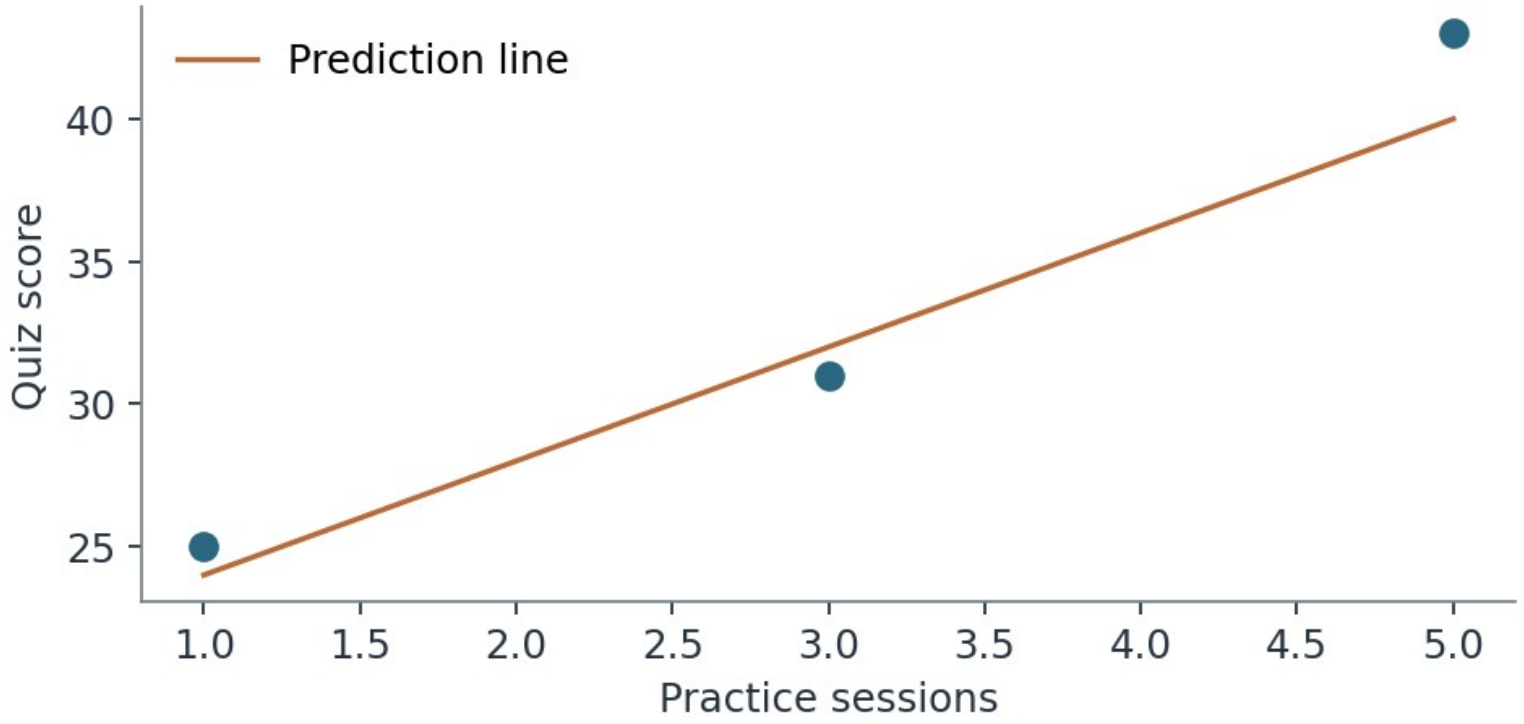
Definitions: Regression and predictions

Regression line: a fitted line used to predict the response y from explanatory variable x.

Slope: predicted change in y for one unit increase in x.

Residual: observed y minus predicted y.

Illustration and interpretation



Notice the pattern

A regression line summarizes a trend. A prediction is not a promise; actual results can be above or below the line.

Illustrative data / model

Formulas

$$\hat{y} = a + bx, \quad e = y - \hat{y}$$

Important ideas and notation

Least-squares line $\hat{y} = a + bx$ minimizes the sum of squared vertical residuals. b is predicted change in y for a one-unit increase in x.

Residual = observed y − predicted $\hat{y}$. Use spreadsheet SLOPE, INTERCEPT and CORREL or graphing technology.

Interpolation predicts inside observed x range; extrapolation outside it is riskier. High abs(r) indicates strong linear association, not guaranteed accuracy or causation.

Example 2.12.1

Line $\hat{y} = 20+4x$ predicts quiz score from practice sessions. For $x=5$, actual 43. Find the predicted score and residual.

Solution:

Prediction 40

residual $43-40=3$ points.

Example 2.12.2

Pairs (1,3),(2,5),(3,7),(4,9). Find regression line.

Solution:

$\hat{y}=1+2x$

$r=1$

predicted y at $x=2.5$ is 6.

Example 2.12.3 • Predicting a quiz score

Find the residual for five sessions.

x (sessions)	Observed score	Predicted score
1	25	24
3	31	32
5	43	40

Illustrative classroom dataset

Example 2.12.3 • Solution

Find the residual for five sessions.

Solution:

$43 - 40 = 3$ points, so the line underestimated.

Individual practice

1. For $\hat{y}=10+3x$, find prediction at $x=4$.

2. Actual 19, predicted 22. Residual?

3. Training x values 1–6; predicting $x=20$ is what?

Group activity

Use the pairs (1,2), (2,4), (3,5), (4,8), (5,9). Calculate the regression slope and intercept using a spreadsheet; predict at x=3.5.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. 22.

2. −3.

3. Extrapolation, potentially unreliable.

Group solution

Solution:

Least-squares slope = 1.8

intercept = 0.2

prediction at 3.5 = 6.5.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Does a small residual prove a causal relationship?

2. A model is $\hat{y} = 12 + 3x$. At $x=4$, the actual response is 22. Find the prediction and residual.

Show your reasoning. Submit before leaving.

Five-minute review

1. Does a small residual prove a causal relationship?

2. For $\hat{y}=10+3x$, find prediction at $x=4$.

Write your answers before we discuss them.

How transformations change statistics

What we will learn

Predict how adding, multiplying or changing an observation affects center and spread.

Why does it matter?

If every score receives the same bonus, the scores move together. Their distances from the mean do not change.

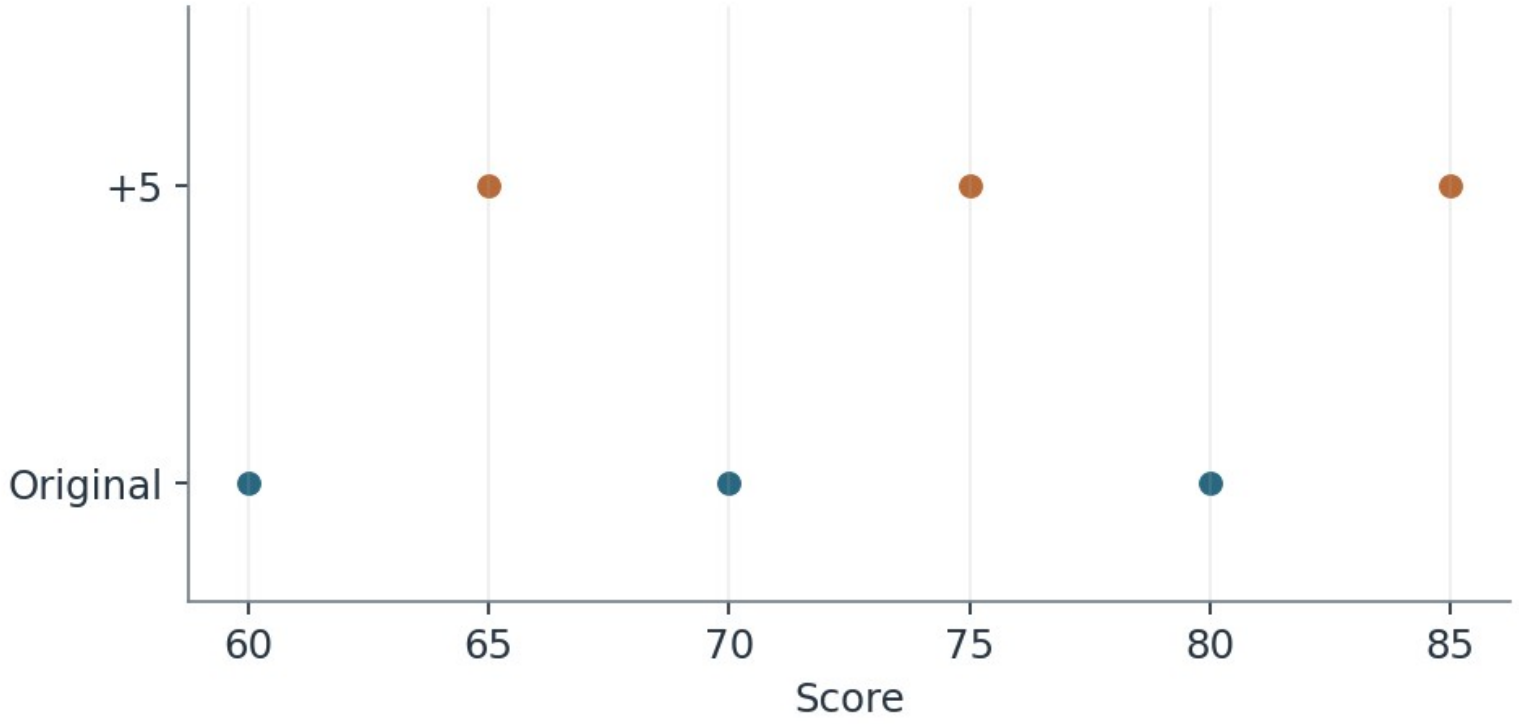
Definitions: How transformations change statistics

Linear transformation: changing every data value by $x' = a + bx$.

Location measures: summaries such as mean and median that describe typical position.

Spread measures: summaries such as range, IQR and SD that describe variability.

Illustration and interpretation



Notice the pattern

If every score receives the same bonus, the scores move together. Their distances from the mean do not change.

Illustrative data / model

Formulas

$$\bar{y} = a + k\bar{x}, \quad s_y = |k|s_x, \quad s_y^2 = k^2 s_x^2$$

Important ideas and notation

Adding a constant c to every observation adds c to mean, median and quartiles; range, IQR and standard deviation stay unchanged.

Multiplying every observation by a positive k multiplies mean, median, quartiles, range, IQR and standard deviation by k; variance by k^2.

Changing only one observation is different: the mean can shift substantially, while the median may be stable. Recalculate instead of applying a whole-data transformation rule.

Example 2.13.1

Scores 60, 70, 80 have mean 70 and range 20. Add 5 to every score. Find the new mean, range and describe the change in SD.

Solution:

65, 75, 85

mean 75

range 20. Standard deviation is unchanged.

Example 2.13.2

A sample has mean 12, SD 3 and variance 9. Convert every value from feet to inches. Find the new mean, SD and variance.

Solution:

Multiply by 12: mean 144 inches, SD 36 inches, variance 1296 square inches.

Example 2.13.3 • Adjusting a grading scale

Which transformation changes the spread?

Original	Add 5	Multiply by 2
60	65	120
70	75	140
80	85	160

Illustrative classroom dataset

Example 2.13.3 • Solution

Which transformation changes the spread?

Solution:

Multiplying by 2 doubles the range and SD

adding 5 leaves them unchanged.

Individual practice

1. Mean = 24, SD = 4. Add 10 to every value. Find new mean and SD.

2. Median = 15, IQR = 6. Multiply every value by 2.

3. Data 2, 4, 6, 8, 10. Replace 10 with 50. Compare means and medians.

Group activity

Two sensors measure the same five temperatures. Sensor B always reports 2 degrees higher. If A reports 10, 12, 14, 16, 18, compare means, ranges and SDs. Explain calibration.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. 34 and 4.

2. Median 30; IQR 12.

3. Old mean 6; new mean 14; both medians 6.

Group solution

Solution:

A mean 14, B mean 16

both ranges 8, both population SD $\sqrt{8} \approx 2.83$. A constant offset changes location, not spread.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. If every value is doubled, what happens to variance?

2. A dataset has mean 10 and SD 2. Transform every value using $x' = 3x+5$.

Show your reasoning. Submit before leaving.

Five-minute review

1. If every value is doubled, what happens to variance?

2. Mean = 24, SD = 4. Add 10 to every value. Find new mean and SD.

Write your answers before we discuss them.

Regression residuals and model reliability

What we will learn

Compute residuals and judge the reliability and limitations of linear predictions.

Why does it matter?

A good-looking regression line can hide patterns. Residuals help us see when a straight line is a poor fit.

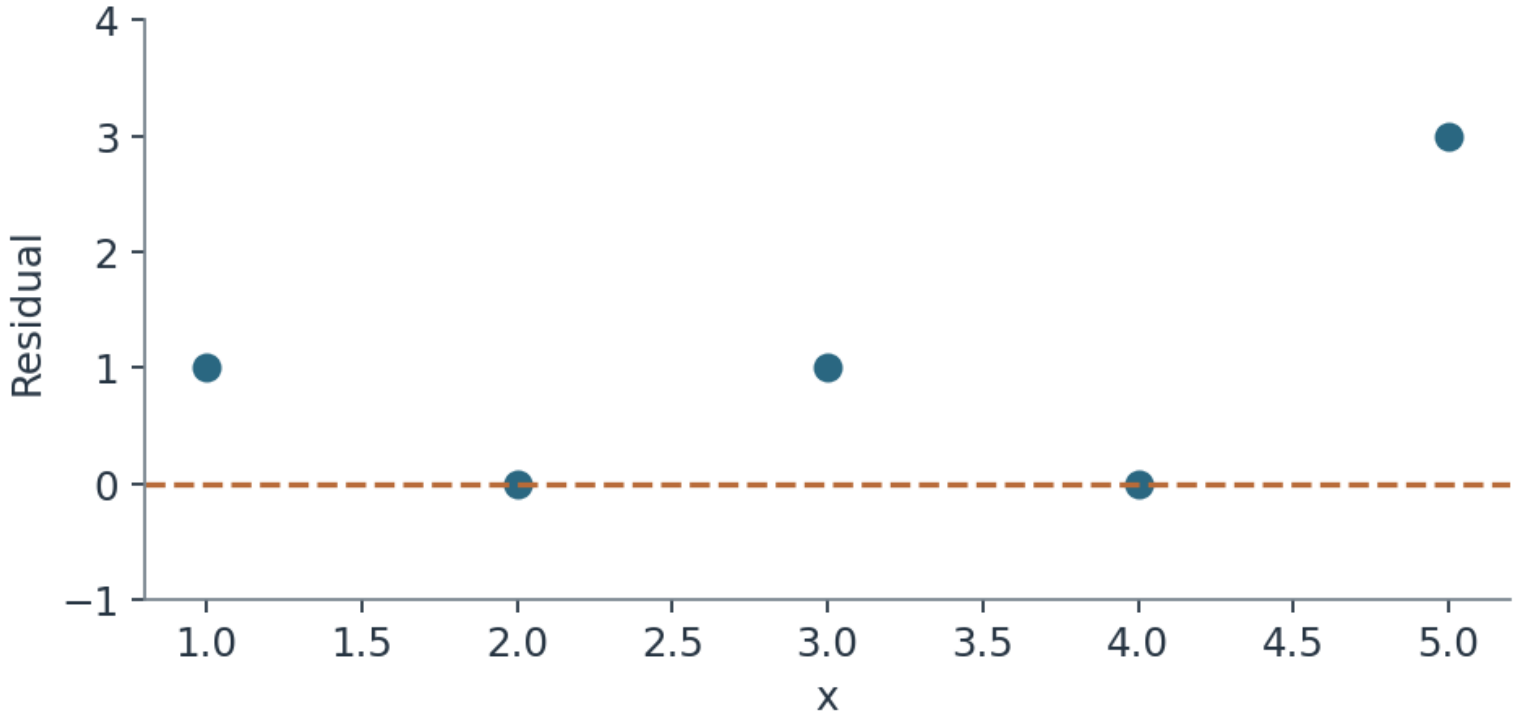
Definitions: Regression residuals and model reliability

Residual plot: a graph of residuals against fitted values or the explanatory variable.

Interpolation: prediction within the observed range of x.

Extrapolation: prediction beyond the observed range of x.

Illustration and interpretation



Notice the pattern

A good-looking regression line can hide patterns. Residuals help us see when a straight line is a poor fit.

Illustrative data / model

Formulas

$$e = y - \hat{y}$$

Important ideas and notation

Residual = observed y − predicted y. A positive residual means the model underestimated the response; a negative residual means it overestimated.

Use scatterplots and the correlation coefficient r to judge linear association. A residual plot with no obvious pattern supports a linear model; a curve suggests a different model.

Using the method correctly

Interpolation predicts inside the observed x-range; extrapolation predicts outside it and is riskier. Even a strong correlation cannot establish causation.

Example 2.14.1

The prediction line is $\hat{y} = 20 + 3x$. At $x = 4$, observed $y = 35$. Find the residual.

Solution:

Predicted $y = 32$

residual $= 35 - 32 = 3$. The model underestimated by 3.

Example 2.14.2

A model was fitted for study times from 1 to 5 hours. r = .92. Is a prediction at 12 hours reliable?

Solution:

It is extrapolation. Strong association within the observed range does not establish reliability at 12 hours.

Example 2.14.3 • Checking a prediction model

Which observation has the largest absolute residual?

x	Observed	Predicted
1	4	3
2	5	5
3	8	7
4	9	9
5	14	11

Illustrative classroom dataset

Example 2.14.3 • Solution

Which observation has the largest absolute residual?

Solution:

At $x=5$, residual $= 14-11 = 3$.

Individual practice

1. $\hat{y} = 8 + 2x$. At $x = 6$, $y = 17$. Find the residual.

2. Residuals are positive at both ends of the x-range and negative in the middle. What does this suggest?

3. Two correlations are $-.85$ and $.35$. Which has stronger linear association?

Group activity

For x = 1, 2, 3, 4, 5, observed y = 4, 5, 8, 9, 14. Use $\hat{y} = 1 + 2x$. List residuals, then plot residuals against x and discuss fit.

Show your calculation or graph. Write a sentence explaining the result in context.

Check your answers

1. $17 - 20 = -3$; overestimate by 3.

2. Curvature; the linear model may not be appropriate.

3. $-.85$, because its absolute value is larger.

Group solution

Solution:

Predictions 3, 5, 7, 9, 11

residuals 1, 0, 1, 0, 3. All are nonnegative, suggesting the supplied line underpredicts

a fitted line may improve the model.

Compare your conclusion with the question. Check your units and explain what the numbers represent.

Exit ticket • 5 minutes

1. Does r = .98 prove that increasing x causes y to increase?

2. A prediction is 50 and the actual response is 44. What does the residual tell us?

Show your reasoning. Submit before leaving.