



Descriptive Statistics and Distributions (AP Statistics Foundations)

Four focused lessons, worked examples, a mistake clinic, and 60 practice problems with full explanations.

In this booster

31.1 Exploring One-Variable Data

31.2 Graphs and Comparing Distributions

31.3 Exploring Two-Variable Data

31.4 Categorical Data and Association

Every section teaches the idea, walks through the most common problem types, flags the traps students fall into, and ends with fifteen practice problems. Full answers and explanations are at the back.

31.1 Exploring One-Variable Data

Learn to describe a distribution with shape, outliers, center, and spread. Use the five-number summary, the 1.5 IQR rule, and standard deviation to summarize data and judge which statistics to trust.

Why This Matters

Every statistics course begins the same way: you are handed a list of numbers and asked, "What do you see?" Quiz scores, commute times, battery lifetimes, points per game. A list of 40 numbers tells you almost nothing until you describe it. This section gives you the vocabulary and the tools to do that precisely, and AP Statistics expects you to use them in every free-response answer.

Students get stuck in predictable places. They report a number without saying what it means in context. They use the mean when an outlier makes the median the honest choice. They forget the difference between a **resistant** statistic and a **non-resistant** one. And they mix up the formulas for the outlier fences and the standard deviation. Learn the logic here and these errors disappear. The same ideas power everything later: normal curves, sampling distributions, confidence intervals, and tests.

Core Ideas

A **quantitative variable** takes numerical values for which arithmetic makes sense (minutes, points, dollars). To describe its distribution, use the four-part checklist **SOCS**: **S**hape, **O**utliers (and other unusual features), **C**enter, **S**pread. Always write your description in the context of the problem.

Measures of center

The **mean** is $\bar{x} = \frac{\sum x_i}{n}$, the balance point of the data. The **median** is the middle value of the ordered list (the average of the two middle values when n is even).

A statistic is **resistant** if extreme values barely change it. The median and IQR are resistant. The mean, standard deviation, and range are not.

Shape	Mean versus median	Best summary
Roughly symmetric	About equal	Mean and standard deviation
Skewed right (long tail to high values)	Mean is greater than median	Median and IQR
Skewed left (long tail to low values)	Mean is less than median	Median and IQR

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

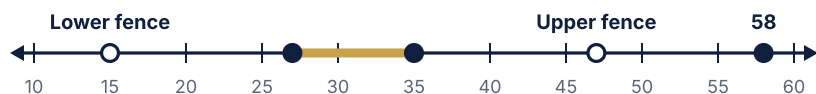
For spread, start with the **five-number summary**: minimum, Q_1 , median, Q_3 , maximum. Here Q_1 is the median of the lower half of the ordered data and Q_3 the median of the upper half; when n is odd, leave the overall median out of both halves. The **range** is max minus min, and the **interquartile range** is $IQR = Q_3 - Q_1$, the width of the middle half of the data.

The 1.5 IQR rule for outliers

A value is flagged as an outlier if it is **below** $Q_1 - 1.5(IQR)$ or **above** $Q_3 + 1.5(IQR)$.

A value exactly equal to a fence is not flagged. Outliers should be investigated, not automatically deleted.

The number line below shows the rule at work for a set of 10 players' scoring totals. The shaded stretch is the middle half of the data (from $Q_1 = 27$ to $Q_3 = 35$). The fences sit at 15 and 47, and the value 58 lies beyond the upper fence.



Middle half (27 to 35), fences at 15 and 47, and the outlier 58.

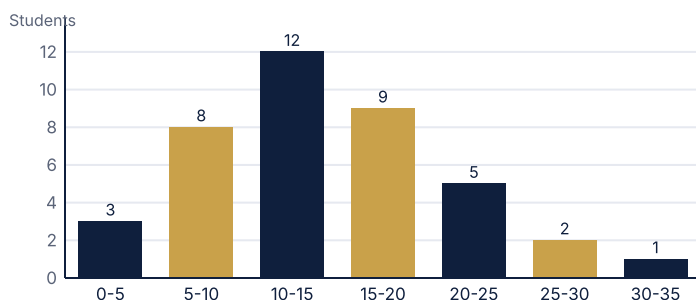
Standard deviation

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

It measures the typical distance of the values from the mean. Larger s means more spread. $s = 0$ only when all values are identical. The units of s match the units of the data.

To find s by hand: (1) compute $\bar{x}$; (2) subtract it from each value; (3) square each deviation; (4) add and divide by $n - 1$; (5) take the square root. On the exam you will usually use technology, but you must know what the number means.

Linear transformations. Adding a constant c to every value shifts the center ($\bar{x}$, median, quartiles) by c but leaves spread (s , IQR, range) unchanged. Multiplying every value by b multiplies center and spread by $|b|$. Shape does not change when $b > 0$.



Minutes to travel to school for 40 students (each bar counts values from the left edge up to, but not including, the right edge).

Worked Examples

EXAMPLE 1: MEAN, MEDIAN, AND THE EFFECT OF AN OUTLIER

Seven students counted the text messages they sent yesterday: 8, 11, 11, 14, 15, 19, 52. Find the mean and median. Which better describes a typical student?

SOLUTION

Sum: $8 + 11 + 11 + 14 + 15 + 19 + 52 = 130$, so $\bar{x} = 130/7 \approx 18.6$ texts.

The data are already ordered. The middle (4th) value is 14, so the median is 14 texts.

The value 52 pulls the mean above all but one of the students. The median, 14, is closer to what most students sent. Because the data are skewed right, the median is the better summary of a typical student.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

EXAMPLE 2: FIVE-NUMBER SUMMARY AND OUTLIERS

A coach records points scored by a player in 10 games: 22, 25, 27, 28, 30, 31, 33, 35, 36, 58. Find the five-number summary and decide whether any game is an outlier.

SOLUTION

Here $n = 10$. The median is the average of the 5th and 6th values: $(30 + 31)/2 = 30.5$.

Lower half: 22, 25, 27, 28, 30, so $Q_1 = 27$. Upper half: 31, 33, 35, 36, 58, so $Q_3 = 35$. Five-number summary: 22, 27, 30.5, 35, 58.

$IQR = 35 - 27 = 8$ and $1.5(8) = 12$. Fences: $27 - 12 = 15$ and $35 + 12 = 47$.

Since $58 > 47$, the 58-point game is an outlier. No game is below 15.

EXAMPLE 3: STANDARD DEVIATION BY HAND

Find the sample standard deviation of 3, 5, 6, 8, 13.

SOLUTION

Mean: $(3 + 5 + 6 + 8 + 13)/5 = 35/5 = 7$.

Deviations: $-4, -2, -1, 1, 6$. Squares: 16, 4, 1, 1, 36, sum = 58.

$s^2 = 58/(5 - 1) = 14.5$, so $s = \sqrt{14.5} \approx 3.81$.

Interpretation: a typical value in this list is about 3.8 units from the mean of 7.

EXAMPLE 4: LINEAR TRANSFORMATIONS

The heights of students in a class have mean 66 inches and standard deviation 3 inches. (a) The teacher also reports heights in centimeters (1 inch = 2.54 cm). Find the new mean and standard deviation. (b) Instead, every height is increased by 2 inches to account for shoes. What happens to the mean and standard deviation?

SOLUTION

(a) Multiplying by 2.54: mean = $66(2.54) = 167.64$ cm and standard deviation = $3(2.54) = 7.62$ cm. Both center and spread are scaled.

(b) Adding 2 shifts the center: new mean = 68 inches. Spread is unchanged: standard deviation stays 3 inches.

EXAMPLE 5: DESCRIBING A HISTOGRAM IN CONTEXT

The histogram of travel times (minutes) for 40 students is shown in the Core Ideas section. Describe the distribution and say where the median lies.

SOLUTION

Shape: unimodal and skewed right, with a tail toward the longer commutes.

Outliers: the few trips of 25 minutes or more form the tail, with no clear gap separating them, so there is no obvious outlier.

Center: cumulative counts are 3, 11, 23, so the 20th and 21st values lie in the 10 to 15 minute interval. The median is between 10 and 15 minutes, and the mean is somewhat larger because of the right skew.

Spread: times range from about 0 to 35 minutes, with most students between 5 and 20 minutes.

EXAMPLE 6: REMOVING AN OUTLIER

Use the 10 scoring totals from Example 2. Compare the mean, standard deviation, and IQR with and without the 58-point game. Which statistics are resistant?

SOLUTION

With all 10 games: $\bar{x} = 32.5$, $s \approx 9.97$, $IQR = 8$.

Without 58 (9 games: 22 to 36): $\bar{x} \approx 29.7$, $s \approx 4.64$. With $n = 9$, $Q_1 = (25 + 27)/2 = 26$ and $Q_3 = (33 + 35)/2 = 34$, so $IQR = 8$.

The single outlier raised the mean by about 2.8 and more than doubled s , but the IQR did not move. The IQR (and median) are resistant; the mean and standard deviation are not.

Coach's Playbook

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

- 1 **Sort first.** Median and quartiles only make sense on ordered data. Write the sorted list before anything else.
- 2 **Use SOCS in context.** Name the variable and units: "The distribution of travel times is skewed right with a median of about 12 minutes." Never write a bare number.
- 3 **Match the summary to the shape.** Symmetric with no outliers: mean and standard deviation. Skewed or with outliers: median and IQR.
- 4 **Check the fences with arithmetic.** Compute IQR, then 1.5 times IQR, then both fences. Compare each extreme value to the fence.
- 5 **Sanity check s .** For mound-shaped data, s is roughly one quarter of the range. If your answer is far from that, recheck.

Common traps

1) Computing the median of unordered data. 2) Using n instead of $n - 1$ when finding the sample standard deviation, or forgetting the square root. 3) Writing "1.5" instead of $1.5 \times \text{IQR}$ in the outlier rule. 4) Saying "the mean is higher so the data are skewed" without a graph or the median to compare.

Mistake Clinic

Spot the error 1

Problem: Find the median of 9, 3, 12, 5, 8. Wrong work: "There are 5 values, so the median is the middle one, 12." What went wrong: The student took the middle of the list as written. The list was not in order. Right solution: Sort: 3, 5, 8, 9, 12. The middle value is 8, so the median is 8.

Spot the error 2

Problem: A data set has $Q_1 = 20$ and $Q_3 = 32$. What are the outlier fences? What the student wrote: "Upper fence = $32 + 1.5 = 33.5$ and lower fence = $20 - 1.5 = 18.5$." What went wrong: The student added 1.5, not 1.5 times the IQR. Right solution: $\text{IQR} = 12$ and $1.5(12) = 18$. Upper fence = $32 + 18 = 50$; lower fence = $20 - 18 = 2$.

Spot the error 3

Problem: Find the sample standard deviation of 1, 2, 3, 4, 5. Wrong work: "Mean 3. Squared deviations: 4, 1, 0, 1, 4, sum 10. Divide by 4: the standard deviation is 2.5." What went wrong: 2.5 is the variance s^2 . The student forgot the square root. Right solution: $s = \sqrt{2.5} \approx 1.58$.

Real-World Connection

When news reports say "the median home price rose," they use the median because a few mansions would drag the mean far upward. Suppose five homes sold for 240, 255, 260, 285, and 1200 thousand dollars. The mean is 448 thousand but the median is only 260 thousand. The median describes what a typical buyer sees; the mean describes the total market value divided by 5.

Challenge Corner

Problem: Five positive whole numbers $a \leq b \leq c \leq d \leq e$ have mean 10, median 9, a single mode of 7, and range 8. Find them. Solution: The median is $c = 9$. The mode 7 must appear at least twice. Since $d, e \geq 9$, the two 7s must be a and b . The range gives $e = a + 8 = 15$. The sum is $5(10) = 50$, so $d = 50 - 7 - 7 - 9 - 15 = 12$. The set is 7, 7, 9, 12, 15. Check: mean 10, median 9, mode 7 (only repeated value), range 8.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

Practice

PRACTICE Exploring One-Variable Data

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- Household incomes in a large city have a long tail toward very high values. Which pair of summaries is best for describing the distribution?
 (A) Mean and standard deviation (B) Median and IQR
 (C) Mean and IQR (D) Median and standard deviation
- What is the IQR of the data 3, 5, 7, 9, 11, 13, 15, 17?
 (A) 6 (B) 12 (C) 14 (D) 8
- A delivery service reports a mean arrival time of 40 minutes with a standard deviation of 6 minutes. Which interpretation is best?
 (A) Every delivery arrives between 34 and 46 minutes.
 (B) The slowest delivery was 6 minutes later than the fastest.
 (C) A typical delivery time is about 6 minutes from the mean of 40.
 (D) Six percent of deliveries are late.
- A data set has $Q_1 = 34$ and $Q_3 = 46$. Which value would be flagged as an outlier by the 1.5 IQR rule?
 (A) 16 (B) 64 (C) 15 (D) 63
- A set of values has mean 10 and standard deviation 4. Every value is multiplied by 3 and then 2 is subtracted. What is the new standard deviation?
 (A) 10 (B) 14 (C) 12 (D) 30
- Which pair of (mean, median) values is most consistent with a distribution skewed to the left?
 (A) (50, 50) (B) (44, 52) (C) (55, 48)
 (D) (52, 52.1)
- Data set: 10, 12, 14, 16, 18. A new value, 100, is added. Which statistic changes the least?
 (A) Mean (B) Standard deviation (C) Range
 (D) Median
- What is the sample standard deviation of 2, 4, 6?
 (A) 2 (B) 1.63 (C) 2.83 (D) 4
- A data set has 20 values with $Q_1 = 12$ and $Q_3 = 20$. About how many values lie between 12 and 20?
 (A) 5 (B) 10 (C) 15 (D) 20
- Find the mean and median of 6, 8, 9, 9, 12, 31, and say which is the better measure of center and why.
- Find the five-number summary of 14, 16, 17, 19, 20, 22, 23, 25, 26, 48 and use the 1.5 IQR rule to decide whether 48 is an outlier.
- Find the sample standard deviation of 2, 5, 5, 8, 10. Round to two decimal places.
- Daily high temperatures in a city have mean 20 degrees Celsius and standard deviation 3 degrees. Fahrenheit temperature is $F = 1.8C + 32$. Find the mean and standard deviation in Fahrenheit.
- A histogram of 40 quiz scores has intervals 0 to 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50 with counts 4, 9, 15, 8, 4. (a) In which interval does the median lie? (b) Estimate the mean using interval midpoints.
- The mean of 12 values is 20. What value must a 13th observation have to raise the mean to exactly 22? Then find the new mean if the 13th value had been 33.

31.2 Graphs and Comparing Distributions

Build and read dot plots, stemplots, histograms, and box plots, then write comparisons that address shape, center, spread, and outliers in context.

Why This Matters

A graph can answer in one second what a table of numbers hides for a minute. Statisticians draw a picture before they compute anything, because the shape of a distribution decides which numbers are honest. A mean means little for a lopsided histogram, and a single outlier can fool a calculator but never fools a box plot.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

Students get stuck on three skills. They confuse a histogram with a bar chart. They read a box plot as if the box length counted people instead of showing spread. And they write comparisons that never mention both groups, or that give numbers without context. In this section you will learn to build and read each display and to write the comparison sentences that earn full credit.

Core Ideas

Choose a display to match the data and the question. All of the displays below show one quantitative variable.

Display	Best for	What it shows well
Dot plot	Small data sets with few distinct values	Every value, gaps, clusters, and shape
Stemplot	Small to medium data sets	Every value kept, plus shape; stems and leaves must be ordered and need a key
Histogram	Medium to large data sets, any size	Shape, center, spread, and peaks; values grouped into equal-width classes
Box plot	Comparing several groups	Five-number summary and outliers; hides peaks and gaps

Histogram versus bar chart

A **histogram** shows a quantitative variable. Bars touch, are equal in width, and each bar's height is the count (or percent) of values in that interval. A bar chart shows categories and its bars may be reordered.

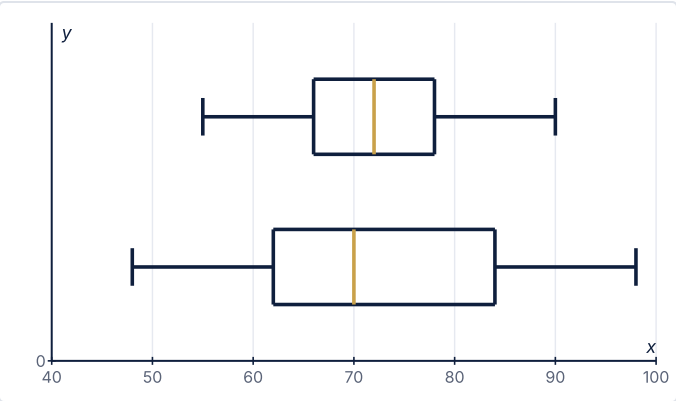
A value on a boundary goes in the interval to its right, unless the problem says otherwise. **Relative frequency** = $\frac{\text{count in class}}{n}$.

The box plot

A box plot is drawn from the five-number summary: a box from Q_1 to Q_3 , a line at the median, and whiskers to the extremes. Each of the four sections (whisker, lower half of the box, upper half of the box, whisker) holds **about 25%** of the data.

In a **modified box plot**, values beyond the 1.5 IQR fences are plotted as separate points, and each whisker stops at the most extreme value that is **not** an outlier.

The box plots below compare two classes. The gold line in each box is the median. Notice that the two medians are close, but Class B has a much longer box and longer whiskers.



Side-by-side box plots of test scores. Class A (top): 55, 66, 72, 78, 90. Class B (bottom): 48, 62, 70, 84, 98.

Comparing distributions

A good comparison addresses **shape, center, spread, and outliers** for both groups, uses comparison words ("greater than," "more variable than"), and names the variable in context.

Wrong: "Class A has a median of 72 and Class B has 70." Right: "Class A's median score (72) is slightly higher than Class B's (70)."

Shape vocabulary. A distribution is **symmetric** if the left and right sides mirror each other, **skewed right** if the tail extends toward larger values, and **skewed left** if it extends toward smaller values. It is **unimodal** with one peak, **bimodal** with two, or **uniform** if the bars are roughly equal. A bimodal shape is a clue that two different groups may have been mixed together.

Worked Examples

Want a plan built around **your** student?
Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

EXAMPLE 1: BUILD AND READ A STEM PLOT

Points scored by a team in 14 games: 24, 27, 31, 33, 33, 35, 38, 41, 42, 46, 48, 52, 55, 63. Make a stemplot, find the median and IQR, and describe the shape.

SOLUTION

Use tens digits as stems and units digits as leaves, in order. Stem 2: 4 7. Stem 3: 1 3 3 5 8. Stem 4: 1 2 6 8. Stem 5: 2 5. Stem 6: 3. Key: 3 | 1 means 31 points.

Median: $(38 + 41)/2 = 39.5$. $Q_1 = 33$ (median of the lower 7 values) and $Q_3 = 48$, so $IQR = 15$.

Fences: $33 - 22.5 = 10.5$ and $48 + 22.5 = 70.5$. No outliers.

Shape: roughly symmetric to slightly skewed right, centered near 40 points, with scores from 24 to 63.

Stem	Leaves
2	4 7
3	1 3 3 5 8
4	1 2 6 8
5	2 5
6	3

EXAMPLE 2: READING A HISTOGRAM

The histogram below shows daily hours of screen time for 50 students. (a) What percent use at least 3 hours? (b) In which interval is the median? (c) Describe the shape.

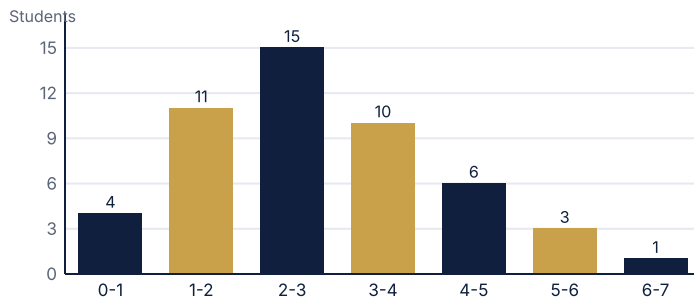
SOLUTION

Counts are 4, 11, 15, 10, 6, 3, 1 (total 50).

(a) Intervals 3 to 4, 4 to 5, 5 to 6, 6 to 7: $10 + 6 + 3 + 1 = 20$, and $20/50 = 40\%$.

(b) Cumulative counts: 4, 15, 30. The 25th and 26th values are in the 2 to 3 hour interval.

(c) Unimodal, skewed right, with a typical value around 2 to 3 hours and a tail up to about 7 hours.



Daily hours of screen time for 50 students.

EXAMPLE 3: A DOT PLOT

Fifteen students reported how many siblings they have. The dot plot is shown. Find the mean and median and describe the shape.

SOLUTION

Counts by value: 0 siblings (1), 1 (3), 2 (5), 3 (3), 4 (2), 6 (1). Total = 15.

Median is the 8th value: sorted list is 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 6, so the median is 2.

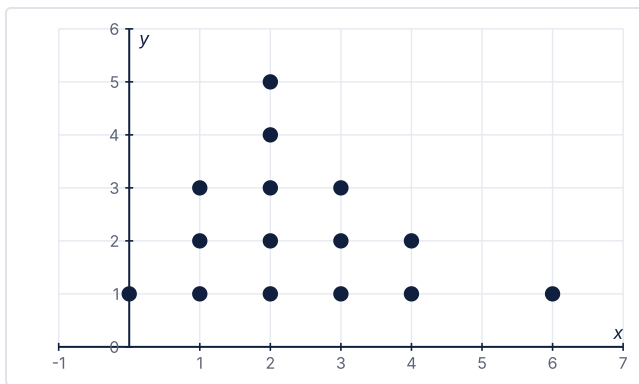
Mean: $(0 + 3 + 10 + 9 + 8 + 6)/15 = 36/15 = 2.4$.

The mean (2.4) is a bit above the median (2) and the tail stretches toward 6, so the distribution is slightly skewed right. For this set $Q_1 = 1$, $Q_3 = 3$, $IQR = 2$, and the upper fence is $3 + 3 = 6$, so a value of 6 is on the fence and is **not** an outlier.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →



Dot plot of the number of siblings for 15 students.

EXAMPLE 4: MODIFIED BOX PLOT

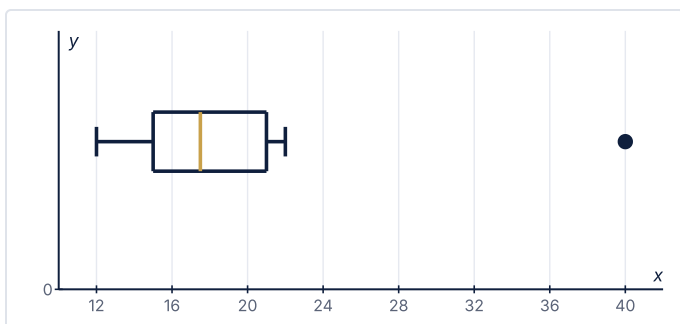
Minutes to finish a puzzle for 10 students: 12, 14, 15, 15, 17, 18, 19, 21, 22, 40. Draw the modified box plot.

SOLUTION

Median = $(17 + 18)/2 = 17.5$. $Q_1 = 15$, $Q_3 = 21$, IQR = 6.

Fences: $15 - 9 = 6$ and $21 + 9 = 30$. The value 40 is above 30, so it is an outlier and is plotted as a separate point.

The upper whisker stops at 22, the largest value that is not an outlier. The lower whisker goes to 12.



Modified box plot of puzzle times. The separate point is the outlier 40.

EXAMPLE 5: COMPARING TWO DISTRIBUTIONS

Use the box plots of Class A and Class B in Core Ideas. Compare the test scores.

SOLUTION

Medians: A is 72 and B is 70, so the centers are similar. Spread: A has IQR = $78 - 66 = 12$ and range 35; B has IQR = $84 - 62 = 22$ and range 50.

Write it in context: "The median score in Class A (72) is a little higher than in Class B (70), but Class B's scores are more variable, with an IQR of 22 compared with 12 for Class A. Class B also has both the lowest score (48) and the highest (98)."

Neither plot shows an outlier, because every whisker end is inside its fences.

EXAMPLE 6: BIMODAL SHAPE

A cafe records arrival times of 65 customers. Counts by hour: 6-7: 5, 7-8: 16, 8-9: 9, 9-10: 3, 10-11: 2, 11-12: 7, 12-13: 17, 13-14: 6. Describe the shape and explain it.

SOLUTION

The histogram has two peaks, at 7 to 8 and at 12 to 13: it is **bimodal**.

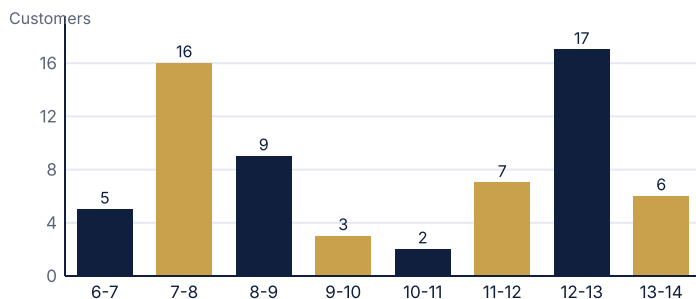
A single mean or median would hide this. The data likely mix two groups of customers, a morning rush and a lunch rush.

Context sentence: "Arrival times are bimodal with peaks near 7:30 and 12:30, suggesting two different customer groups."

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →



Customer arrival times by hour (24-hour clock).

Coach's Playbook

- 1 Pick the display from the data.** Few values: dot plot or stemplot. Many values: histogram. Several groups to compare: side-by-side box plots.
- 2 Label everything.** Axis titles with units, class widths, a stemplot key, and a title in context.
- 3 Read counts before percents.** Add the bar heights to get n , then divide the counts you need by n .
- 4 Use the 25% rule.** Each part of a box plot holds about a quarter of the data, so "at least 75% scored above 62" can be read straight from Q_1 .
- 5 Write comparisons in pairs.** Every sentence mentions both groups: "A is more variable than B, with IQR 12 compared with 22."

Common traps

1) Leaving a gap between bars in a histogram, or using unequal class widths without adjusting. 2) Reading a longer box as more people instead of more spread. 3) Skipping an empty stem in a stemplot, which hides gaps. 4) Giving numbers for one group without comparing them to the other, or omitting units and context.

Mistake Clinic

Spot the error 1

Problem: A histogram of 50 values has counts 12, 6, 5, 9, 10, 8 for the intervals 0-1, 1-2, 2-3, 3-4, 4-5, 5-6. Where is the median? Wrong work: "The tallest bar is 0-1, so the median is between 0 and 1." What went wrong: The tallest bar is the modal class, not the median. Right solution: Cumulative counts are 12, 18, 23, 32. The 25th and 26th values are in the 3 to 4 interval, so the median is there.

Spot the error 2

Problem: Class B's box plot has a longer box than Class A's. Interpret. Wrong work: "Class B has more students than Class A." What went wrong: The length of the box shows spread (the IQR), not how many values there are. Right solution: "Class B's scores are more variable, since its IQR is larger." Group sizes cannot be read from a box plot.

Spot the error 3

Problem: Make a stemplot of 12, 15, 21, 24, 25, 43, 47. Wrong work: Rows for stems 1, 2, 4 only. What went wrong: Stem 3 has no leaves but must still appear. Leaving it out hides the gap between 25 and 43. Right solution: Rows: 1 | 2 5; 2 | 1 4 5; 3 | (empty); 4 | 3 7, with a key such as 4 | 3 means 43.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

Real-World Connection

A coach tracks lap times of two swimmers over a season using box plots. Swimmer P has a lower median but a long box, so some races are far slower than others. Swimmer Q has a slightly higher median but a short box, so Q is dependable. Which swimmer you start in a relay depends on whether you value the best case or consistency, and box plots make that trade-off visible.

Challenge Corner

Problem: Team A has 10 players with a mean height of 70 inches. Team B has 15 players with a mean height of 74 inches. Find the mean height of all 25 players, and explain why you cannot find the combined median from the two team medians. Solution: Total height = $10(70) + 15(74) = 700 + 1110 = 1810$, so the combined mean is $1810/25 = 72.4$ inches. Means can be pooled because the total is recoverable. Medians cannot: the combined median depends on how the values interleave, which the two medians do not reveal. It only has to fall between the two team medians.

Practice**PRACTICE** **Graphs and Comparing Distributions**

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- In a box plot, about what percent of the data lies between the median and Q_3 ?
 (A) 10% (B) 25% (C) 50% (D) 75%
- Which display is best for showing the shape of the distribution of 200 measurements of a continuous variable?
 (A) Stemplot (B) Pie chart (C) Histogram (D) Bar chart
- A stemplot uses the key 2|8 means 2.8 pounds. The row 4|055 represents which values?
 (A) 40, 45, 45 (B) 4.0, 4.5, 4.5 (C) 0.4, 0.5, 0.5 (D) 4.05, 4.5
- A unimodal histogram has a long tail toward high values. Which statement is most likely true?
 (A) The mean is less than the median. (B) The mean and median are equal.
 (C) The standard deviation is zero. (D) The mean is greater than the median.
- Group A has summary (min, Q_1 , median, Q_3 , max) = (20, 30, 40, 50, 70) and Group B has (35, 45, 55, 65, 80). Which statement must be true?
 (A) Every value in B exceeds every value in A.
 (B) B has a larger IQR than A.
 (C) At least 75% of B's values are greater than A's median.
 (D) Group A has more values than Group B.
- A data set has $Q_1 = 20$, $Q_3 = 30$, and maximum 52. The largest value that is not an outlier is 41. In a modified box plot, where does the upper whisker end?
 (A) 41 (B) 45 (C) 52 (D) 30
- A histogram of 80 values has 28 values in the interval 20 to 30. What is the relative frequency of that interval?
 (A) 0.28 (B) 0.65 (C) 2.86 (D) 0.35
- A histogram of heights of all students in a school's volleyball team and chess club combined has two clear peaks. What is the best explanation?
 (A) The data contain an error. (B) The mean is too small.
 (C) The data may combine two different groups. (D) The distribution is uniform.
- In a box plot, the median line is near the right end of the box and the left whisker is much longer than the right whisker. How is the distribution best described?
 (A) Skewed left (B) Skewed right (C) Symmetric (D) Uniform

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

10. A stemplot of 13 exam scores is shown. Key: 3 | 0 means 30. Find the median and the IQR.

Stem	Leaves
1	2 5 8
2	0 1 1 4 6 9
3	0 3 7
4	5

11. A histogram of 50 values has intervals 0 to 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50 with counts 5, 12, 18, 10, 5. What percent of the values are less than 30? What is the relative frequency of the 20 to 30 interval?
12. Ten values: 12, 14, 15, 15, 17, 18, 19, 21, 22, 40. Compute the fences. Which value is an outlier, and where does the upper whisker of the modified box plot end?
13. Class P has median 68, IQR 10, and range 40. Class Q has median 75, IQR 24, and range 55. Write one sentence, in context, comparing the test scores of the two classes.
14. The number of pets owned by 13 students is 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 3, 4, 7. Find the mean and median, state whether 7 is an outlier by the 1.5 IQR rule, and describe the shape.
15. A data set has 48 values with $Q_1 = 30$, median 42, $Q_3 = 60$, minimum 12, and maximum 100. (a) About how many values are greater than 60? (b) Is 100 an outlier? Justify with the 1.5 IQR rule.

31.3 Exploring Two-Variable Data

Describe scatterplots, interpret correlation, find and use the least-squares line, and read residuals and r squared with careful, context-based wording.

Why This Matters

Does more studying lead to higher scores? Do taller players score more? Does a bigger engine mean worse fuel economy? Whenever two quantitative variables are measured on the same individuals, we want to know whether they are related, how strongly, and whether we can use one to predict the other. That is the heart of data analysis in science, business, and medicine.

Students get stuck on interpretation more than on calculation. They say "correlation proves causation," they interpret the slope without the words "predicted" and "on average," they mix up a residual and a predicted value, and they trust predictions far outside the data. This section teaches you to describe a scatterplot, find and use the least-squares line, and defend your interpretations in context.

Core Ideas

In a two-variable data set, the **explanatory variable** x is used to predict or explain the **response variable** y . A **scatterplot** shows each individual as one point. Describe it with four words: **direction** (positive or negative), **form** (linear or curved), **strength** (how closely points follow the form), and **unusual features** (outliers, clusters). Always say it in context.

The correlation coefficient r

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right), \quad -1 \leq r \leq 1.$$

r measures the strength and direction of a **linear** relationship only. Its sign gives the direction; the closer $|r|$ is to 1, the stronger the linear pattern. r has no units, does not change if you change units or swap x and y , and is not resistant to outliers.

Value of r	Typical description
r near 1	Strong positive linear association
r near -1	Strong negative linear association
$ r $ about 0.5	Moderate linear association
r near 0	Little or no linear association (a curved pattern is still possible)

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

The least-squares regression line

$$\hat{y} = a + bx, \quad b = r \frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}.$$

It minimizes the sum of the squared residuals. It always passes through $(\bar{x}, \bar{y})$. The **slope** b is the predicted change in y for each one-unit increase in x . The **intercept** a is the predicted y when $x = 0$ (meaningful only if $x = 0$ makes sense).

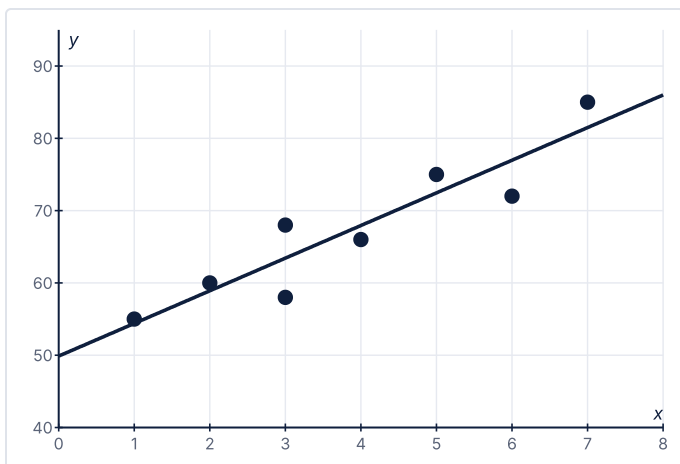
A **residual** is the actual value minus the predicted value: $\text{residual} = y - \hat{y}$. A positive residual means the point lies above the line (the model underestimated). Residuals add to zero. A **residual plot** graphs the residuals against x . A random scatter with no pattern says a line is appropriate; a curved pattern says it is not.

The coefficient of determination

r^2 is the proportion of the variation in y that is accounted for by the linear relationship with x .

Template: "About [r^2 as a percent] of the variation in [y in context] is accounted for by the linear relationship with [x in context]."

The scatterplot below shows hours studied and quiz score for 8 students, together with the least-squares line $\hat{y} = 49.88 + 4.515x$ (for these data $r \approx 0.925$ and $r^2 \approx 0.856$).



Quiz score (y) against hours studied (x) for 8 students, with the least-squares line.

Cautions. Association is not causation: only a well-designed experiment can show that changing x causes a change in y . **Extrapolation**, predicting outside the range of the x data, is unreliable. An **influential point** (often an outlier in the x direction) can change the slope and r dramatically.

Worked Examples

EXAMPLE 1: DESCRIBING A SCATTERPLOT

Describe the scatterplot of hours studied and quiz score shown above.

SOLUTION

Direction: positive. Students who studied more tend to have higher scores.

Form: linear, with no curvature.

Strength: strong, because the points hug the line ($r \approx 0.925$).

Unusual features: no obvious outliers. In context: "There is a strong, positive, linear association between hours studied and quiz score."

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

EXAMPLE 2: COMPUTING R AND THE LINE BY HAND

Five trials give $x = 1, 2, 3, 4, 5$ and $y = 10, 8, 9, 5, 3$. Find r and the least-squares line.

SOLUTION

$\bar{x} = 3$, $\bar{y} = 7$. Deviations of x : $-2, -1, 0, 1, 2$. Deviations of y : $3, 1, 2, -2, -4$.

$$\sum(x - \bar{x})(y - \bar{y}) = -6 - 1 + 0 - 2 - 8 = -17. \quad \sum(x - \bar{x})^2 = 10. \quad \sum(y - \bar{y})^2 = 9 + 1 + 4 + 4 + 16 = 34.$$

$$b = \frac{-17}{10} = -1.7 \text{ and } a = 7 - (-1.7)(3) = 12.1, \text{ so } \hat{y} = 12.1 - 1.7x.$$

$$r = \frac{-17}{\sqrt{10 \cdot 34}} \approx -0.922. \text{ The association is strong, negative, and linear.}$$

EXAMPLE 3: LINE FROM SUMMARY STATISTICS

For a class, hours studied has mean 3.5 and standard deviation 1.8. Quiz score has mean 70 and standard deviation 11. The correlation is $r = 0.88$. Find the regression line and r^2 .

SOLUTION

$$b = r \frac{s_y}{s_x} = 0.88 \cdot \frac{11}{1.8} \approx 5.378.$$

$$a = \bar{y} - b\bar{x} = 70 - 5.378(3.5) \approx 51.18, \text{ so } \hat{y} = 51.18 + 5.378x.$$

$$r^2 = 0.88^2 = 0.7744. \text{ About 77\% of the variation in quiz score is accounted for by the linear relationship with hours studied.}$$

EXAMPLE 4: INTERPRETING AND PREDICTING

Using $\hat{y} = 49.88 + 4.515x$ from the 8-student data (hours studied 1 to 7), interpret the slope and intercept. Predict the score for 5 hours and for 15 hours.

SOLUTION

Slope: for each additional hour studied, the predicted quiz score increases by about 4.5 points, on average.

Intercept: a student who studied 0 hours is predicted to score about 49.9. This is only slightly outside the data, so it is a rough but reasonable value.

5 hours: $\hat{y} = 49.88 + 4.515(5) \approx 72.5$. This is interpolation, so it is trustworthy.

15 hours: $\hat{y} \approx 117.6$. That is above any possible score. The data only cover 1 to 7 hours, so this extrapolation is meaningless.

EXAMPLE 5: RESIDUALS AND RESIDUAL PLOTS

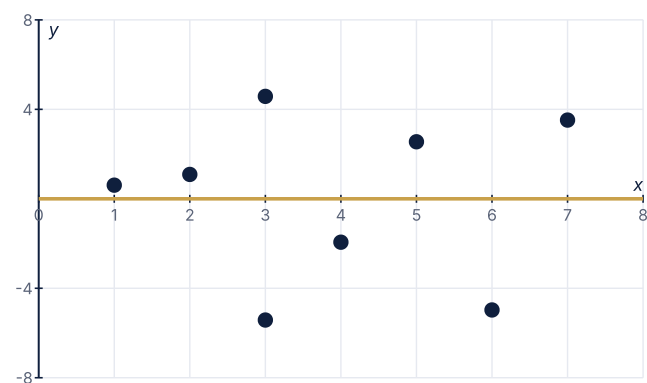
The student who studied 3 hours scored 58. Find and interpret the residual. Then comment on the residual plot below.

SOLUTION

$$\hat{y} = 49.88 + 4.515(3) = 63.425. \text{ Residual} = 58 - 63.425 \approx -5.4.$$

The student scored about 5.4 points **below** what the line predicts for 3 hours of study.

The residual plot (all eight residuals) shows no curve and no fan shape, just random scatter around 0. A linear model is appropriate.



Residual plot for the 8 students. The scatter is random, so a line is a good model.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

EXAMPLE 6: AN INFLUENTIAL POINT

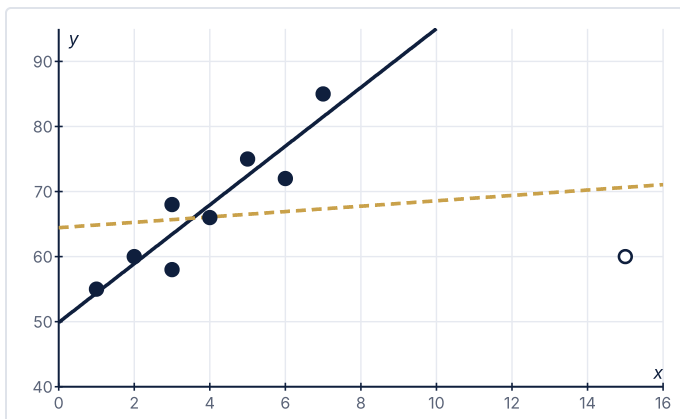
A ninth student studied for 15 hours but scored only 60 because of illness. Predict the effect on the correlation and the slope. The new values are $r \approx 0.18$ and $b \approx 0.41$.

SOLUTION

The new point is far from the others in the x direction and far from the original line, so it has high leverage and is **influential**.

It pulls the line down at the right end: the slope falls from about 4.5 to about 0.4 and r falls from 0.925 to 0.18.

Lesson: report an influential point, and compute the model both with and without it.



The open point (15 hours, score 60) pulls the line from the solid navy line to the dashed gold line.

Coach's Playbook

- 1 Identify the roles.** Which variable is explanatory (x) and which is the response (y)? Read the question: "predict y from x ."
- 2 Plot first.** If the pattern is curved, r and a line are the wrong tools.
- 3 Use the line's anchor.** The line passes through $(\bar{x}, \bar{y})$. If you are given the means and one other fact, you can usually find a .
- 4 Write interpretations with three ingredients.** "Predicted," "on average," and the units. Slope: "For each additional ___, the predicted ___ increases by ___ on average."
- 5 Check units and ranges.** If a prediction is outside the range of the data, say it is extrapolation and unreliable.

Common traps

1) Saying a high r proves that x causes y . 2) Computing a residual as predicted minus actual. 3) Interpreting r^2 as "the percent of points on the line." 4) Interpreting the intercept when $x = 0$ is not in the data or has no meaning.

Mistake Clinic**Spot the error 1**

Problem: $\hat{y} = 12 + 3.5x$, where x is hours of practice and y is points scored. Interpret the slope. Wrong work: "If you practice 1 hour, you score 3.5 points." What went wrong: The slope is a rate of change, not a prediction at $x = 1$. The wording also omits "predicted" and "on average." Right solution: For each additional hour of practice, the predicted number of points increases by 3.5, on average.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

Spot the error 2

Problem: A point has $x = 3$, actual $y = 58$, and predicted $\hat{y} = 63.4$. Find the residual. Wrong work: "Residual = $63.4 - 58 = 5.4$." What went wrong: Residual means actual minus predicted. Right solution: $58 - 63.4 = -5.4$. The negative sign means the actual score is below the prediction.

Spot the error 3

Problem: $r = -0.80$ between outdoor temperature and the number of hot drinks sold. Interpret r^2 . Wrong work: " $r^2 = -0.64$, so 64% of hot drink sales are explained by cold weather and the rest by chance." What went wrong: r^2 is never negative, and the wording must refer to variation, not to causes. Right solution: $r^2 = 0.64$. About 64% of the variation in hot drinks sold is accounted for by the linear relationship with temperature.

Real-World Connection

A farmer records rainfall (inches) and corn yield (bushels per acre) over many seasons and finds $r \approx 0.7$ with slope about 6 bushels per inch within the observed range of 10 to 30 inches. That supports a planning estimate, but not a prediction for a 60-inch flood, which is far outside the data, and the line alone cannot prove that rain caused the yield, since sunlight and soil also vary.

Challenge Corner

Problem: A regression of y on x gives $\hat{y} = 8 + 2x$. The data have $\bar{x} = 5$ and $s_x = 3$, and $s_y = 7.5$. (a) Find $\bar{y}$. (b) Find r . (c) If you regress x on y instead, what is the slope? Solution: (a) The line passes through $(\bar{x}, \bar{y})$, so $\bar{y} = 8 + 2(5) = 18$. (b) $b = r s_y / s_x$ gives $2 = r(7.5/3) = 2.5r$, so $r = 0.8$. (c) Regressing x on y gives slope $r s_x / s_y = 0.8(3/7.5) = 0.32$, which is not $1/2$. The two regression lines are different, but r is the same.

Practice**PRACTICE Exploring Two-Variable Data**

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- A scatterplot has $r = -0.9$. Which description is best?

(A) Strong negative linear association	(B) Weak negative linear association
(C) Strong positive linear association	(D) No association
- Which correlation indicates the strongest linear relationship?

(A) 0.70	(B) 0.40	(C) -0.20	(D) -0.85
----------	----------	-----------	-----------
- In $\hat{y} = 12 + 3.5x$, x is hours of practice and y is points scored. Which interpretation of the slope is correct?

(A) Each hour of practice produces exactly 3.5 points.
(B) The predicted score with no practice is 3.5.
(C) Scores increase by 3.5 percent per hour.
(D) For each additional hour of practice, the predicted score increases by 3.5 points on average.
- A point has actual value 80, and the regression line predicts 74.5. What is the residual, and what does it mean?

(A) -5.5; the model overestimated	(B) 5.5; the model overestimated
(C) 5.5; the model underestimated	(D) -5.5; the model underestimated
- For a regression, $r^2 = 0.64$. Which statement is correct?

(A) 64% of the points lie on the line.
(B) 64% of the variation in y is accounted for by the linear relationship with x .
(C) The correlation is 0.64.
(D) The slope is 0.64.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

6. Every x value is multiplied by 2.54 (inches to centimeters) and the y values are unchanged. What happens to r ?
 (A) It is multiplied by 2.54. (B) It is divided by 2.54.
 (C) It stays the same. (D) It changes sign.
7. A regression line is $\hat{y} = 5 + 2x$ and $\bar{x} = 4$. What is $\bar{y}$?
 (A) 8 (B) 13 (C) 10 (D) 18
8. A line fitted for ages 10 to 15 is used to predict a 40-year-old's value. This is an example of
 (A) extrapolation (B) interpolation (C) a residual
 (D) random sampling
9. A residual plot shows a clear U-shaped curve. What should you conclude?
 (A) A linear model is not appropriate; a curve may fit better.
 (B) The correlation must be zero.
 (C) The linear model is appropriate.
 (D) The slope must be negative.
10. Find the least-squares line when $\bar{x} = 12$, $s_x = 3$, $\bar{y} = 80$, $s_y = 15$, and $r = -0.6$. Then predict y at $x = 12$.
11. A regression line is $\hat{y} = 20.5 + 2.2x$. For $x = 15$ the actual value is 58. Find the predicted value and the residual.
12. For the data $x = 1, 2, 3, 4$ and $y = 3, 5, 4, 8$, find r and the least-squares line. Round r to three decimal places.
13. The correlation between the number of study hours and exam score in a class is $r = -0.78$. (a) Find r^2 as a percent. (b) Explain why the sign of r is lost in r^2 .
14. A phone's battery percentage is modeled by $\hat{y} = 98 - 7.5x$, where x is hours of video streaming (observed from 0 to 8 hours). Interpret the slope and intercept, and explain what is wrong with predicting at 10 hours.
15. Without one point, a data set has $r = 0.925$ and slope 4.5. After adding a point far to the right and well below the line, r drops to about 0.18. Name this type of point, explain why r changed so much, and say what an analyst should do.

31.4 Categorical Data and Association

Read and build two-way tables, find marginal, joint, and conditional percents, decide whether variables are associated, and recognize Simpson's paradox.

Why This Matters

Not all data are numbers. Favorite subject, grade level, whether a patient improved, whether a team won: these are **categorical variables**, and the question is usually the same. Are two categorical variables related? Does the choice of lunch depend on grade? Is a treatment tied to improvement? Two-way tables answer such questions, and they power surveys, medical studies, and market research.

Students get stuck on one idea: **which total do I divide by?** Choosing the wrong denominator is the most common error on this topic. A second sticking point is Simpson's paradox, where a trend that appears in every subgroup reverses when the groups are combined. This section gives you a clear rule for the denominator and a method for spotting the paradox.

Core Ideas

A **two-way table** counts individuals by two categorical variables at once. The row and column totals sit in the margins, and the grand total is in the corner. Here is a survey of 250 students: grade level (rows) and where they eat lunch (columns).

Grade	School lunch	Packed lunch	Off campus	Total
9th	45	35	10	90
10th	30	30	20	80
11th	20	25	35	80
Total	95	90	65	250

Marginal, joint, and conditional

A **marginal distribution** uses only a total row or column: for example, 95 of 250 students (38%) eat school lunch. It ignores the other variable.

A **joint relative frequency** divides a single cell by the grand total: 10 of 250 (4%) are 9th graders who eat off campus.

A **conditional distribution** restricts attention to one row or column and divides by that **row or column total**.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

The denominator rule

For "What percent of **those who** are ___ are ___?", the denominator is the count of the group named after "of those who" (the given condition).

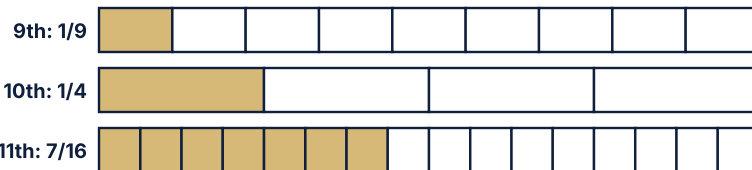
Reading "of the 11th graders" tells you to divide by the 11th-grade total, 80. Reading "of the off-campus students" tells you to divide by 65.

Association

Two categorical variables are **associated** if the conditional distributions of one variable differ across the categories of the other. If all conditional distributions are (nearly) identical, there is no association, and the variables are **independent**.

Independence means $P(A | B) = P(A)$. Always compare percents, not raw counts, when group sizes differ.

Now compare the grades. The fraction of each grade that eats off campus is 10 of 90 (9th), 20 of 80 (10th), and 35 of 80 (11th). The fraction bars show these conditional proportions.



Fraction of each grade that eats off campus (10/90, 20/80, 35/80). The shaded part grows with grade level, which signals an association.

Simpson's paradox

Simpson's paradox occurs when a comparison that holds in every subgroup reverses (or disappears) after the subgroups are combined.

It happens because of a **lurking variable** and unequal group sizes: the combined rate is a weighted average of the subgroup rates, and the weights differ between the two groups being compared.

Worked Examples

EXAMPLE 1: MARGINAL AND JOINT PERCENTS

Use the lunch table. (a) What percent of all students eat packed lunch? (b) What percent of all students are 10th graders who eat school lunch?

SOLUTION

(a) Marginal: the packed-lunch total is 90, so $90/250 = 0.36 = 36\%$.

(b) Joint: the cell for 10th grade and school lunch is 30, so $30/250 = 0.12 = 12\%$.

EXAMPLE 2: CONDITIONAL PERCENTS

Use the lunch table. (a) Of the 11th graders, what percent eat off campus? (b) Of the students who eat off campus, what percent are 11th graders? Are the two answers the same?

SOLUTION

(a) The condition is "11th graders," so divide by 80: $35/80 = 0.4375 = 43.75\%$.

(b) The condition is "eat off campus," so divide by 65: $35/65 \approx 53.8\%$.

They differ because the denominators differ. $P(A | B)$ is generally not equal to $P(B | A)$.

EXAMPLE 3: IS THERE AN ASSOCIATION?

Compare the lunch choices of 9th graders and 11th graders. Is lunch choice associated with grade?

SOLUTION

9th graders: school $45/90 = 50\%$, packed $35/90 \approx 38.9\%$, off campus $10/90 \approx 11.1\%$.

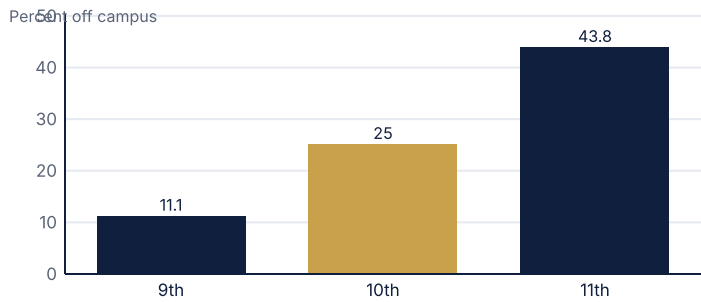
11th graders: school $20/80 = 25\%$, packed $25/80 = 31.25\%$, off campus $35/80 = 43.75\%$.

The conditional distributions are very different: 9th graders mostly eat school lunch, while 11th graders mostly eat off campus. There is an association between grade and lunch choice.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →



Percent of each grade that eats off campus.

EXAMPLE 4: WRITING THE CONCLUSION

Write a complete conclusion about the relationship between grade and lunch choice, citing evidence.

SOLUTION

"The percent of students who eat off campus increases with grade level: about 11% of 9th graders, 25% of 10th graders, and 44% of 11th graders. Because the conditional distributions of lunch choice differ among grades, there is an association between grade level and lunch choice in this sample."

Notice the three ingredients: conditional percents, a comparison, and the word "association" in context. Do not claim that grade causes the choice.

EXAMPLE 5: COMPLETING A TWO-WAY TABLE

In a club of 60 students, 36 are in band. Of the 24 students not in band, 9 play a sport. In total 25 play a sport. Fill in the table.

SOLUTION

Not in band and sport: 9. Not in band and no sport: $24 - 9 = 15$.
Sport total is 25, so band and sport: $25 - 9 = 16$. Band and no sport: $36 - 16 = 20$.
No sport total: $20 + 15 = 35$, and $25 + 35 = 60$ matches the grand total.
Table: band, sport 16; band, no sport 20; not band, sport 9; not band, no sport 15.

EXAMPLE 6: SIMPSON'S PARADOX

Two drivers delivered packages on time as follows. Driver A: highway 90 of 100, city 120 of 300. Driver B: highway 340 of 400, city 20 of 60. Compare the drivers by route type and overall.

SOLUTION

Highway: A 90%, B 85%. City: A 40%, B about 33.3%. Driver A has the higher on-time rate on both routes.
Overall: $A = 210/400 = 52.5\%$ and $B = 360/460 \approx 78.3\%$. Driver B looks better overall!
The lurking variable is route type. Driver A handled mostly city routes (300 of 400), where on-time rates are low for everyone, while Driver B handled mostly highway routes. The overall rates are weighted averages with different weights.
Conclusion: compare the drivers within each route type, since that controls the lurking variable.

Driver	Highway on time	City on time	Overall on time
A	90 of 100 (90%)	120 of 300 (40%)	210 of 400 (52.5%)
B	340 of 400 (85%)	20 of 60 (33.3%)	360 of 460 (78.3%)

Coach's Playbook

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

- 1 **Underline the condition.** In "of those who ___, what percent ___, " the phrase right after "of those who" gives the denominator.
- 2 **Write the fraction before the percent.** Write $\frac{\text{count}}{\text{group total}}$ first, then convert.
- 3 **Check totals.** Rows and columns must add to the same grand total. A mismatch means an arithmetic error.
- 4 **Compare percents, never counts.** Different group sizes make raw counts misleading.
- 5 **Hunt for lurking variables.** If a combined result looks surprising, split the data by a third variable.

Common traps

1) Dividing by the grand total when the question gives a condition. 2) Mixing up $P(A | B)$ with $P(B | A)$. 3) Averaging two percents instead of combining counts. 4) Concluding that an association shows cause and effect.

Mistake Clinic

Spot the error 1

Problem: Using the lunch table, what percent of 11th graders eat off campus? Wrong work: " $35/250 = 14\%$." What went wrong: The student divided by the grand total. The question names 11th graders as the condition. Right solution: $35/80 = 43.75\%$.

Spot the error 2

Problem: Using the lunch table, what percent of off-campus students are 11th graders? Wrong work: " $35/80 = 43.75\%$." What went wrong: That is the percent of 11th graders who eat off campus. The condition here is eating off campus, with a total of 65. Right solution: $35/65 \approx 53.8\%$.

Spot the error 3

Problem: Driver A is on time 90 of 100 on highways and 120 of 300 in the city. Find the overall on-time rate. Wrong work: " $(90\% + 40\%)/2 = 65\%$." What went wrong: Averaging percentages ignores the different group sizes. Right solution: $(90 + 120)/(100 + 300) = 210/400 = 52.5\%$.

Real-World Connection

A school wants to know whether a tutoring app helps. A table shows that 60% of app users improved their grade compared with about 47% of non-users. That is an association. Because students chose whether to use the app, motivated students may be more likely to choose it, which is a lurking variable. Only a randomized experiment, where students are assigned to the app by chance, could support a claim that the app caused the improvement.

Challenge Corner

Problem: A survey of 100 people has row totals 40 (Group X) and 60 (Group Y) and column totals 50 (Yes) and 50 (No). If answer and group are independent, fill in the table. Solution: Under independence each cell equals $\frac{(\text{row total})(\text{column total})}{\text{grand total}}$. Group X: $40(50)/100 = 20$ Yes and 20 No. Group Y: $60(50)/100 = 30$ Yes and 30 No. Check: each group has 50% Yes, so the conditional distributions are identical.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

Practice

PRACTICE Categorical Data and Association

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

1. A survey of 120 students gives the table below. What percent of all students are sophomores who prefer movies?

Class	Music	Movies	Total
Freshman	30	20	50
Sophomore	28	42	70
Total	58	62	120

- (A) 35% (B) 42% (C) 60% (D) 70%

2. Using the same table, what percent of the students who prefer music are freshmen?

Class	Music	Movies	Total
Freshman	30	20	50
Sophomore	28	42	70
Total	58	62	120

- (A) 25% (B) 42% (C) 52% (D) 60%

3. Using the same table, of the sophomores, what percent prefer movies?

Class	Music	Movies	Total
Freshman	30	20	50
Sophomore	28	42	70
Total	58	62	120

- (A) 35% (B) 68% (C) 70% (D) 60%

4. If the conditional distributions of a response variable are identical for every category of the explanatory variable, what do we conclude?

- (A) The variables are associated.
 (B) The sample is biased.
 (C) There is no association between the variables.
 (D) One variable causes the other.

5. Which best describes Simpson's paradox?

- (A) A trend in each subgroup reverses when the groups are combined.
 (B) The mean is larger than the median.
 (C) Large samples give biased results.
 (D) A correlation of zero implies independence.

6. The marginal distribution of a variable in a two-way table is found using

- (A) one row or column total divided by the grand total
 (B) one cell divided by the grand total
 (C) a cell divided by its column total
 (D) the sum of two cell percentages

7. In a two-way table, a row has total 80 and one of its two cells is 31. What is the other cell?

- (A) 31 (B) 39 (C) 49 (D) 111

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

8. Which statement describes a conditional percent?
- (A) 40% of all survey respondents are seniors.
 (B) There are 40 seniors in the survey.
 (C) Seniors make up 40 of 200 students on the roster.
 (D) Of the students who walk to school, 40% are seniors.
9. Team X has a higher win rate than Team Y against strong opponents and also against weak opponents, yet a lower overall win rate. Which explanation is most plausible?
- (A) There is an arithmetic mistake.
 (B) Team X played mostly strong opponents and Team Y mostly weak ones.
 (C) Team X played the same schedule as Team Y.
 (D) Win rates cannot be compared.
10. Using the lunch table from the lesson (Grade 9: 45, 35, 10; Grade 10: 30, 30, 20; Grade 11: 20, 25, 35 for school, packed, off campus), what percent of the students who bring a packed lunch are 10th graders?

Grade	School lunch	Packed lunch	Off campus	Total
9th	45	35	10	90
10th	30	30	20	80
11th	20	25	35	80
Total	95	90	65	250

11. A club has 150 members. The row totals are Juniors 90 and Seniors 60. The column totals are Attended 100 and Did not attend 50. Juniors who attended: 65. Complete the table by finding the other three cells.

	Attended	Did not attend	Total
Juniors	65	?	90
Seniors	?	?	60
Total	100	50	150

12. Using the lunch table, what percent of 9th graders eat school lunch? What percent of 11th graders eat school lunch? Is there an association between grade and eating school lunch?
13. Use the lunch table. Decide whether being an 11th grader and eating off campus are independent events. Compare $P(\text{off campus})$ with $P(\text{off campus} | 11\text{th})$.
14. Tutor P: morning sessions 8 of 10 passed; evening sessions 27 of 90 passed. Tutor Q: morning 70 of 90 passed; evening 2 of 10 passed. Compare the tutors in each session type and overall, and name the lurking variable.
15. In a trial, 150 people used an app and 90 improved. Of 150 people who did not use the app, 70 improved. Find the percent who improved in each group and the difference. Can you conclude the app caused the improvement?

Answers**Answers and Explanations**

Complete solutions for every practice problem, numbered to match. Read the explanation even when your answer is right.

31.1 Exploring One-Variable Data**PRACTICE · EXPLORING ONE-VARIABLE DATA**

- B. The distribution is skewed right, so the median and IQR are resistant to the long tail. Choices A and D use non-resistant measures.
- D. $Q_1 = (5 + 7)/2 = 6$ and $Q_3 = (13 + 15)/2 = 14$, so $IQR = 8$. Choice C (14) is the range.
- C. The standard deviation is a typical distance from the mean, not a limit on values. Choice A treats it as a hard boundary.
- C. $IQR = 12$, so fences are $34 - 18 = 16$ and $46 + 18 = 64$. 15 is below 16, so it is an outlier. 16 and 64 sit exactly on a fence and are not flagged.
- C. Adding a constant does not change spread, and multiplying by 3 multiplies it by 3: $3(4) = 12$. Choice B adds 2 and choice D multiplies the mean.
- B. In a left-skewed distribution the long tail drags the mean below the median, so mean 44 and median 52 fits. Choice C is the right-skew pattern.
- D. The median moves only from 14 to 15. The mean jumps from 14 to about 28.3, and the range and standard deviation grow a lot.
- A. Mean 4, squared deviations 4, 0, 4 sum to 8, divide by $n - 1 = 2$ to get 4, and $\sqrt{4} = 2$. Choice B divides by n ; choice D is the variance.
- B. Q_1 to Q_3 contains the middle 50% of the data, which is 10 of 20 values.
- Mean $75/6 = 12.5$; median $(9 + 9)/2 = 9$. The 31 pulls the mean up, so the median (9) is the better typical value.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

11. Five numbers: 14, 17, 21, 25, 48. IQR = 8, $1.5(8) = 12$, upper fence = 37. Since $48 > 37$, it is an outlier.
12. 3.08. Mean = 6; squared deviations 16, 1, 1, 4, 16 sum to 38; $38/4 = 9.5$; $\sqrt{9.5} \approx 3.08$.
13. Mean = $1.8(20) + 32 = 68$ degrees F. Standard deviation = $1.8(3) = 5.4$ degrees F (the shift of 32 does not change spread).
14. (a) 20 to 30. Cumulative counts are 4, 13, 28, so the 20th and 21st values fall in the third interval. (b) $\frac{4(5)+9(15)+15(25)+8(35)+4(45)}{40} = \frac{990}{40} = 24.75$.
15. 46. Total of 12 values is 240. We need $(240 + x)/13 = 22$, so $x = 286 - 240 = 46$. If $x = 33$, the new mean is $273/13 = 21$.

31.2 Graphs and Comparing Distributions

PRACTICE · GRAPHS AND COMPARING DISTRIBUTIONS

1. B. The box splits at the median; the median to Q_3 section holds one quarter of the data. Choice C confuses it with the whole box (50%).
2. C. A histogram groups continuous values into equal-width intervals and shows shape. A bar chart is for categories, and a stemplot gets unwieldy with 200 values.
3. B. The key says the stem is the whole number of pounds and the leaf is tenths, so the row gives 4.0, 4.5, 4.5. Choice A ignores the key.
4. D. A long tail toward high values (skewed right) pulls the mean above the median.
5. C. Q_1 of B is 45, which exceeds the median of A (40), and 75% of B is at or above Q_1 . Choice B fails because both IQRs equal 20. Choice A fails because B's minimum 35 is below A's median 40. Box plots do not show group size.
6. A. Modified box plots stop each whisker at the most extreme non-outlier, here 41. The fence at 45 is not a data value, and 52 is an outlier.
7. D. Relative frequency = $28/80 = 0.35$. Choice A divides by 100, and choice C is $80/28$.
8. C. Two peaks suggest two subgroups mixed together, such as two teams with different typical heights. A uniform distribution has no peaks.
9. A. The long left whisker and median near the upper end indicate a tail toward smaller values, so it is skewed left.
10. Median 24 (the 7th of 13 values); $Q_1 = (18 + 20)/2 = 19$; $Q_3 = (30 + 33)/2 = 31.5$; IQR = 12.5.
11. 70% and 36%. Values below 30 are $5 + 12 + 18 = 35$, and $35/50 = 0.70$. The 20 to 30 interval has $18/50 = 0.36$.
12. Fences are 6 and 30 ($Q_1 = 15$, $Q_3 = 21$, IQR = 6). The outlier is 40. The upper whisker ends at 22, the largest non-outlier.
13. Class Q's median (75) is higher than Class P's (68), but Q's scores are more variable (IQR 24 compared with 10, range 55 compared with 40).
14. Mean $24/13 \approx 1.85$ pets; median 1 pet. $Q_1 = 0.5$, $Q_3 = 2.5$, IQR = 2, upper fence = 5.5, so 7 is an outlier. The distribution is skewed right with an outlier.
15. (a) About 12 values, because one quarter of 48 is 12. (b) IQR = 30, upper fence = $60 + 45 = 105$. Since $100 < 105$, 100 is not an outlier.

31.3 Exploring Two-Variable Data

PRACTICE · EXPLORING TWO-VARIABLE DATA

1. A. The sign gives direction (negative) and $|r| = 0.9$ is close to 1, which means strong. Choice B mixes up strength.
2. D. Strength depends on $|r|$, so -0.85 beats 0.70. Choice A ignores that negative values can be strong.
3. D. A slope is a rate of change in the predicted value, on average. Choice B describes the intercept, and choice A treats it as exact.
4. C. Residual = $80 - 74.5 = 5.5$. Positive means the actual value is above the line, so the model underestimated. Choices A and D have the wrong sign for actual minus predicted.
5. B. r^2 is the proportion of variation in y accounted for by the linear relationship. Choice A confuses r^2 with the fraction of points on the line.
6. C. r has no units and is unchanged by a positive linear change of scale. Choice D would need a negative multiplier.
7. B. The line passes through $(\bar{x}, \bar{y})$: $5 + 2(4) = 13$. Choice A uses $2(4)$ only.
8. A. Predicting outside the range of the data (10 to 15) is extrapolation and is unreliable.
9. A. A systematic curve in the residuals shows the pattern is not linear. Choice B is wrong because a curved pattern can occur with any value of r .
10. $b = -0.6(15/3) = -3$ and $a = 80 - (-3)(12) = 116$. The line is $\hat{y} = 116 - 3x$, and at $x = 12$, $\hat{y} = 80 = \bar{y}$.
11. Predicted = $20.5 + 2.2(15) = 53.5$. Residual = $58 - 53.5 = 4.5$ (the actual value is above the line).
12. $r \approx 0.837$ and $\hat{y} = 1.5 + 1.4x$. With $\bar{x} = 2.5$, $\bar{y} = 5$: $\sum dx dy = 7$, $\sum dx^2 = 5$, $\sum dy^2 = 14$, $r = 7/\sqrt{70}$, $b = 7/5 = 1.4$, $a = 5 - 1.4(2.5) = 1.5$.
13. (a) $r^2 = 0.6084$, about 60.8%. (b) Squaring a negative number gives a positive number, so r^2 shows only the strength, not the direction.
14. Slope: for each additional hour of streaming, the predicted battery percentage drops by 7.5 points on average. Intercept: before streaming (0 hours) the predicted battery is 98%. At 10 hours the model gives 23%, but 10 is outside the observed 0 to 8 hours, so this is extrapolation and unreliable.
15. It is an influential point (high leverage), because it is far from the other x values and off the trend. A single extreme point in the x direction can pull the line toward it. The analyst should investigate it and report results both with and without the point.

31.4 Categorical Data and Association

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

PRACTICE · CATEGORICAL DATA AND ASSOCIATION

1. A. The joint percent is $42/120 = 0.35$. Choice C (60%) is a conditional percent among sophomores; choice B is the count 42.
2. C. The condition is "prefer music," total 58. $30/58 \approx 52\%$. Choice A divides 30 by the grand total.
3. D. The condition is "sophomores," total 70. $42/70 = 60\%$. Choice A divides by the grand total.
4. C. Identical conditional distributions mean the variables are not associated (they are independent).
5. A. Simpson's paradox is a reversal of a subgroup trend after combining, caused by a lurking variable.
6. A. A marginal distribution uses only totals: a row or column total over the grand total.
7. C. $80 - 31 = 49$. Choice B (39) subtracts the wrong number, and choice D adds.
8. D. A conditional percent divides by the size of the stated group (walkers), not the entire sample. The others are marginal or counts.
9. B. Team X's mix of opponents matters: a team that plays mostly strong opponents has a low combined win rate even if it beats rivals of each type. This is the lurking-variable explanation.
10. 33.3%. The condition is "packed lunch" with a total of 90. The 10th-grade packed count is 30, and $30/90 \approx 33.3\%$. Dividing by 250 instead would give 12%, which is the wrong denominator.
11. Juniors did not attend: $90 - 65 = 25$. Seniors attended: $100 - 65 = 35$. Seniors did not attend: $60 - 35 = 25$. Check: $25 + 25 = 50$. Final: 65, 25; 35, 25.
12. 9th: $45/90 = 50\%$. 11th: $20/80 = 25\%$. The conditional percents differ a lot, so there is an association between grade and eating school lunch (younger students are more likely to).
13. Not independent. $P(\text{off campus}) = 65/250 = 26\%$, but $P(\text{off campus} \mid 11\text{th}) = 35/80 = 43.75\%$. These would be equal if the events were independent.
14. Morning: P 80% versus Q 77.8%. Evening: P 30% versus Q 20%. Tutor P is better in both. Overall: $P = 35/100 = 35\%$ and $Q = 72/100 = 72\%$. Q looks better overall because Q did mostly easy morning sessions (90 of 100) and P mostly the hard evening sessions. The lurking variable is the session type.
15. App: $90/150 = 60\%$. No app: $70/150 \approx 46.7\%$. The difference is about 13.3 percentage points. This shows an association, but the groups were not randomly assigned, so you cannot conclude that the app caused the improvement.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →