STATISTICS · STUDY GUIDE

Exam 2 Study Guide

This guide covers chapters 3 and 4, where raw numbers become pictures and summaries. You will build frequency tables, choose the right graph for the kind of variable you have, read the shape of a distribution and catch a graph that lies. Then you will measure center and spread, draw a boxplot with its fences, standardize values with z-scores and use the empirical rule and Chebyshev's theorem. Work one chapter at a time, try each worked example with the steps covered, and finish with the ten self-test items. The cram sheet at the end holds every formula.

WHAT THIS GUIDE COVERS

- Chapter 3 · frequency, relative and cumulative tables, grouped classes and class width, bar charts, Pareto charts and pies, histograms, stemplots and dot plots, shape and skew, and misleading graphs with the lie factor.
- Chapter 4 · mean, median and mode, range, variance and standard deviation, Bessel's correction and the CV, the five-number summary, IQR and boxplot fences, z-scores, the empirical rule and Chebyshev's theorem.

CHAPTER 3

Displaying Data

Read it in full: **Chapter 3 in the textbook**. Practice it in the lab: **Build a Frequency Table** and **Name the Shape**.

What to know

A **frequency distribution** pairs each value or class with how often it occurs. The frequency column must sum to n. **Relative frequency** is f ÷ n; that column sums to 1 (100%) and lets you compare groups of different sizes. **Cumulative frequency** is a running total that answers "how many at most?" and its last entry must equal n. In the book's exercise survey of 40 adults, the cumulative relative frequency at 3 days is 0.775, so 1 − 0.775 = 0.225 exercised more than 3 days.

When values rarely repeat, build a **grouped frequency distribution** with 5 to 10 classes. Class width ≈ (maximum − minimum) ÷ number of classes, **always rounded up**, or the largest value will have nowhere to go. Classes must not overlap and must not leave gaps, so each value lands in exactly one class. The **midpoint** of a class is the average of its two limits. Plotting cumulative frequency against upper class limits gives an **ogive**.

Qualitative data get a **bar chart**: equal-width bars, separated by gaps because the axis lists names, with the frequency axis starting at zero. Sorting bars from tallest to shortest makes a **Pareto chart**. A **pie chart** shows parts of one whole, with slice angle = (f ÷ n) × 360°, and the angles must total 360°. Pies need mutually exclusive, exhaustive categories, so "check all that apply" data never go in a pie. Eyes compare lengths better than angles, and a tilted 3D pie misreports sizes.

Quantitative data get a **histogram**, a **stem-and-leaf plot** or a **dot plot**. Histogram bars touch, because the axis is a continuous scale and one class ends where the next begins; bars are drawn against the class boundaries (49.5 to 59.5 and so on). A stemplot shows the same shape and keeps every value, and it always needs a key such as "6 | 4 means 64." A dot plot stacks one dot per value and is best for small or discrete data and for spotting outliers. The look of a histogram depends on the bin width, so try more than one.

Describe any display by **shape, center and spread**, with shape first. Skew is named for the **tail**, not the pile: a long tail to the right is skewed right, as with incomes and hospital stays. As a rule of thumb, right skew pulls the mean above the median and left skew pulls it below. The book's ten salaries have a mean of 60.0 thousand and a median of 39.0 thousand. A **uniform** distribution is flat. **Bimodal** data (two peaks) usually mix two groups, so split them before summarizing.

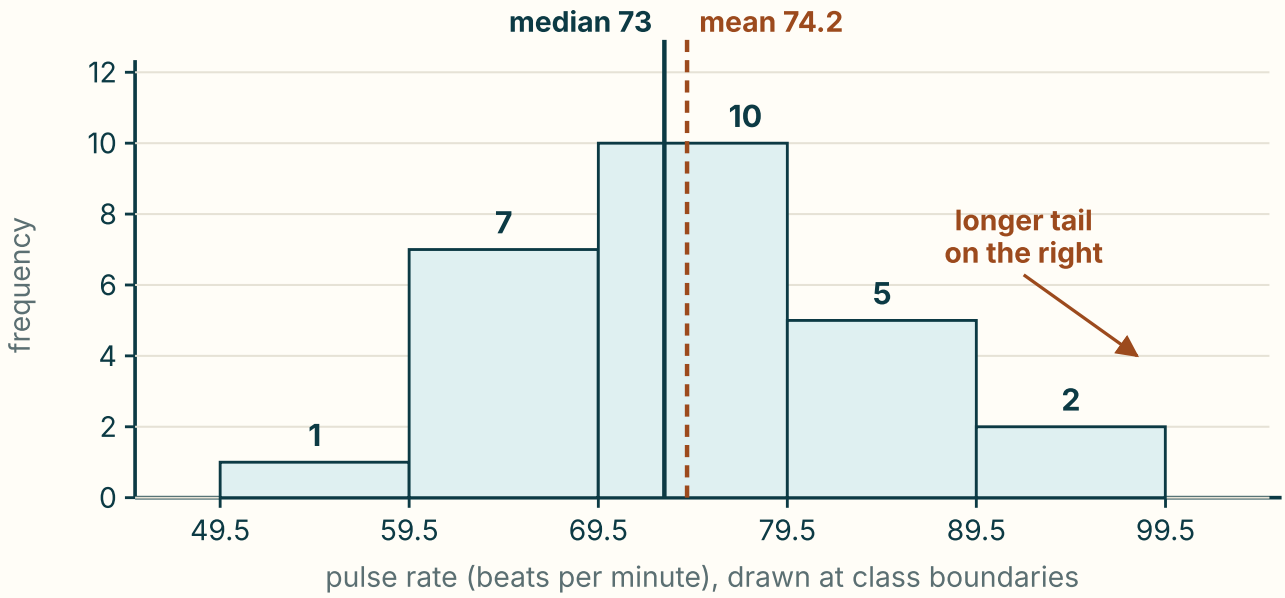
Graphs can mislead without changing a number. A **truncated axis** on a bar chart is the worst offender: with sales of 98, 100, 102 and 104 and an axis starting at 97, a real 6.1% rise looks like 600%. Tufte's **lie factor** = apparent effect ÷ actual effect, here about 98; an honest graph is near 1. Bar charts and histograms must start at zero. Line graphs may start elsewhere if the axis is clearly labeled.

Other distortions: resizing an icon in two dimensions turns a doubling into 4 times the area (8 times if it looks three-dimensional); unequal class widths inflate the widest bar; unevenly spaced

time axes fake acceleration; dual axes let the maker choose where lines cross; and an unlabeled log scale makes explosive growth look calm. Read the axes before the picture.

LOCK THESE IN

- Frequencies sum to n; relative frequencies (f ÷ n) sum to 1; cumulative frequency ends at n.
- Class width = (max − min) ÷ number of classes, rounded up. No overlaps, no gaps.
- Midpoint = (lower limit + upper limit) ÷ 2.
- Pie slice angle = (f ÷ n) × 360°; 1% of a pie is 3.6°.
- Bar chart bars have gaps (categories); histogram bars touch (continuous scale).
- Skew is named for the tail. Right skew: mean usually above median.
- Bimodal usually means two groups mixed together; split them.
- Lie factor = apparent effect ÷ actual effect. Bars must start at zero.



A histogram of 25 pulse rates. Classes of width 10 give counts 1, 7, 10, 5 and 2. The tail stretches right, so the mean (74.2) sits a little above the median (73).

Formulas and when to use them

Quantity	Formula	When to use it
Relative frequency	f ÷ n	Comparing groups of different sizes
Percent	(f ÷ n) × 100%	Reporting a share
Cumulative frequency	Running total of f	"How many at or below?"; ends at n
"More than" share	1 − cumulative relative frequency	The complement of "at most"
Class width	(max − min) ÷ classes, rounded up	Grouped table, 5 to 10 classes
Class midpoint	(lower + upper) ÷ 2	One value to stand for a class
Class boundary	Halfway between one upper limit and the next lower limit	Where histogram bars meet
Pie angle	(f ÷ n) × 360°	Exclusive, exhaustive categories only
Lie factor	apparent effect ÷ actual effect	Judging a graph; honest is about 1
Area inflation	scale² for area; scale³ for apparent volume	Resized icons

Displays and shapes

Item	What it is	Best use or tell
Bar chart	Equal bars with gaps, axis from zero	Categorical counts
Pareto chart	Bar chart sorted tallest first	"Which categories dominate?"
Pie chart	Wedges of one whole	A few exclusive categories
Histogram	Touching bars on a numeric axis	Shape of larger quantitative data
Stem-and-leaf	Stems and leaves, with a key	About 15 to 60 values; keeps every value
Dot plot	One dot per value on a line	Small or discrete data; outliers
Skewed right	Long tail toward high values	Mean usually above median
Skewed left	Long tail toward low values	Mean usually below median
Uniform	Flat	Every class about equally common
Bimodal	Two peaks	Two groups mixed; split them

Worked examples

WORKED EXAMPLE 1 · A GROUPED FREQUENCY TABLE

Twenty-two clinic waits in minutes, sorted: 4, 6, 9, 11, 12, 14, 15, 15, 17, 18, 19, 21, 22, 24, 25, 27, 29, 31, 34, 38, 41, 46. Build a grouped table with 5 classes and find the share who waited 30 minutes or less.

STEP 1 · RANGE AND WIDTH

Range = 46 − 4 = 42. 42 ÷ 5 = 8.4, rounded up to a width of **9**.

STEP 2 · CLASSES

Starting at the minimum: 4 to 12, 13 to 21, 22 to 30, 31 to 39, 40 to 48. The last class reaches 48, past the maximum of 46.

STEP 3 · COUNT

4 to 12: 5; 13 to 21: 7; 22 to 30: 5; 31 to 39: 3; 40 to 48: 2. Total 22 = n.

STEP 4 · MIDPOINTS

8, 17, 26, 35, 44.

STEP 5 · CUMULATIVE

Running totals: 5, 12, 17, 20, 22. At or below 30 minutes: 17 of 22 = **0.773**, about 77%.

WORKED EXAMPLE 2 · RELATIVE FREQUENCIES AND PIE ANGLES

200 households name their main way of watching TV: Streaming 96, Cable 54, Antenna 30, None 20. Find each relative frequency and pie angle, and check the totals.

STEP 1 · RELATIVE FREQUENCY

Streaming: 96 ÷ 200 = 0.48; Cable: 54 ÷ 200 = 0.27; Antenna: 30 ÷ 200 = 0.15; None: 20 ÷ 200 = 0.10.

STEP 2 · ANGLES

Streaming: 0.48 × 360° = 172.8°; Cable: 0.27 × 360° = 97.2°; Antenna: 0.15 × 360° = 54.0°; None: 0.10 × 360° = 36.0°.

STEP 3 · CHECK

Counts: 96 + 54 + 30 + 20 = 200. Shares: 1.00. Angles: 172.8 + 97.2 + 54.0 + 36.0 = **360.0°**.

IS A PIE ALLOWED?

Yes. Each household gave one main answer, so the categories are exclusive and together make one whole. Already sorted largest first, the same counts make a Pareto chart.

WORKED EXAMPLE 3 · MEAN, MEDIAN AND THE DIRECTION OF SKEW

Ten monthly rents at a building, in tens of dollars: 41, 43, 44, 46, 47, 49, 50, 52, 55, 213. Find the mean and median and name the shape.

STEP 1 · MEAN

$\Sigma x = 640$, so $\bar{x} = 640 \div 10 =$ **64.0**.

STEP 2 · MEDIAN

With n = 10, average the 5th and 6th values: (47 + 49) ÷ 2 = **48.0**.

STEP 3 · COMPARE

The mean is 16.0 above the median. Nine of the ten rents are below the mean.

STEP 4 · NAME THE SHAPE

The single 213 stretches a long tail to the right, so the data are **skewed right**. The median is the better "typical rent."

WORKED EXAMPLE 4 · THE LIE FACTOR OF A TRUNCATED AXIS

Monthly sign-ups are 50, 51, 52, 53 (thousands). A bar chart starts its axis at 49. Find the true change, the change as drawn and the lie factor.

STEP 1 · TRUE CHANGE

(53 − 50) ÷ 50 = 3 ÷ 50 = 0.0600, about **6.0%**.

STEP 2 · VISIBLE BARS

Subtract 49: 1, 2, 3, 4.

STEP 3 · APPARENT CHANGE

(4 − 1) ÷ 1 = 3.00, or **300%**.

STEP 4 · LIE FACTOR

3.00 ÷ 0.0600 ≈ **50**. The picture overstates growth about 50 times. Starting the axis at zero gives a lie factor of 1.

TRAPS PEOPLE FALL INTO

- **Rounding the class width down.** Then the classes stop short of the maximum. Always round up.
- **Overlapping classes such as 10 to 20, 20 to 30.** A value of 20 lands in two classes and the counts no longer sum to n.
- **Naming skew by the pile.** Most data sit on the left in a right-skewed distribution. Follow the tail.
- **Putting multiple-response data in a pie.** If people could pick more than one, the shares pass 100%. Use bars.
- **Gaps in a histogram or touching bars in a bar chart.** The gaps say "categories"; touching bars say "continuous scale." Swap them and the picture lies.
- **Trusting a bar chart that starts above zero.** Length means something only from zero. Read the axis first.
- **Resizing an icon to show a doubling.** Doubling width and height makes 4 times the area. Repeat icons instead.
- **Summarizing bimodal data with one mean.** The mean may fall in the valley where almost nobody is. Split the groups.

Chapter 3 self-test · ten items

1. In a sample of n = 50, one class has frequency 9. What is its relative frequency?

A. 0.09
B. 0.50
C. 1.80
D. 0.18

D. 0.18 9 ÷ 50 = 0.18, or 18%.

2. A cumulative relative frequency reaches 0.82 at "4 visits." What share made more than 4 visits?

A. 0.82
B. 0.41
C. 0.18
D. 0.04

C. 0.18 More than 4 is the complement of at most 4: 1 − 0.82 = 0.18.

3. Data run from 15 to 90 and you want 6 classes. What class width should you use?

A. 12
B. 13
C. 12.5
D. 15

B. 13 Range = 90 − 15 = 75, and 75 ÷ 6 = 12.5, which rounds up to 13. Width 12 would end the sixth class at 86, leaving 90 with nowhere to go.

4. A category holds 33 of 220 responses. What is its pie slice angle?

A. 54°
B. 33°
C. 15°
D. 73°

A. 54° 33 ÷ 220 = 0.15, and 0.15 × 360° = 54°.

5. Why do histogram bars touch?

A. Touching bars make the chart easier to print on one page
B. The axis is a continuous scale, so classes meet at boundaries
C. The categories are sorted from the largest to the smallest
D. Each bar shows a separate category that cannot overlap

B. The axis is a continuous scale, so classes meet at boundaries A histogram plots classes against boundaries on a numeric axis. A bar chart keeps gaps because its axis is a list of names.

6. Survey respondents could choose all the snacks they like. Which display is correct?

A. A pie chart with one slice per snack
B. A histogram of the snack names
C. A bar chart with percent of respondents
D. A stem-and-leaf plot of the counts

C. A bar chart with percent of respondents Multiple responses break the "one whole" rule of a pie. A bar chart labeled as percent of respondents is honest.

7. A distribution has most values on the left and a long thin tail to the right. Its shape is:

A. Skewed left
B. Uniform
C. Symmetric
D. Skewed right

D. Skewed right Skew is named for the direction of the tail. Tail right, skewed right.

8. In a set of home prices the mean is far above the median. What does that suggest?

A. A long right tail of high prices
B. A long left tail of low prices
C. A flat, uniform distribution
D. Two equal peaks of prices

A. A long right tail of high prices High values pull the mean up while the median stays near the middle, the usual sign of right skew.

9. A stemplot row reads "7 | 2 5 5 9" with key 7 | 2 means 72. Which values does it hold?

A. 7.2, 7.5, 7.5, 7.9
B. 72, 75, 75, 79
C. 72, 75, 79
D. 27, 57, 57, 97

B. 72, 75, 75, 79 Each leaf joins the stem: 72, 75, 75 and 79. The repeated 5 means 75 appears twice.

10. Revenue rose from 200 to 210, but the bar chart axis starts at 195. What is the lie factor?

A. 5

B. 2

C. 40

D. 0.05

C. 40 True change: 10 ÷ 200 = 0.05. Visible bars are 5 and 15, an apparent change of 10 ÷ 5 = 2.00. Lie factor: 2.00 ÷ 0.05 = 40.

CHAPTER 4

Center and Spread

Read it in full: **Chapter 4 in the textbook**. Practice it in the lab: **Standard Deviation** and **Boxplot Outlier Rule**.

What to know

There are three ideas of "middle." The **mean** $\bar{x} = \Sigma x \div n$ is the balance point. The **median** is the middle sorted value: position $(n + 1) \div 2$ when n is odd, and the average of the two middle values when n is even. The **mode** is the most frequent value, and the only center for categorical data. For 4, 8, 9, 9, 12, 13, 15 the mean is 10, the median 9 and the mode 9.

The median and mode are **resistant**: extreme values barely move them. The mean is not. Six salaries of 38, 42, 45, 48, 52 and 275 thousand have a mean of 83.3 but a median of 46.5, and five of six people earn below the "average." Mean above median suggests right skew; mean below median suggests left skew. Use the mean for roughly symmetric data and for later formulas, and the median for skewed or ordinal data. When they disagree, report both.

Spread separates data sets with the same center. The **range** (max − min) uses only the two most extreme values. Better measures use every value. Deviations $(x - \bar{x})$ always sum to zero, so we square them; their total is the **sum of squares** SS. Sample variance $s^2 = SS \div (n - 1)$; population

variance $\sigma^2 = SS \div N$. The **standard deviation** is the square root, back in the original units. Read it as roughly how far a typical value sits from the mean.

The $n - 1$ is **Bessel's correction**. Deviations from $\bar{x}$ are a bit too small, because $\bar{x}$ sits in the middle of its own sample, and dividing by $n - 1$ makes s^2 an unbiased estimate of σ^2. $n - 1$ is also the **degrees of freedom**: once the mean and $n - 1$ values are known, the last value is forced. The **coefficient of variation** $CV = (s \div \bar{x}) \times 100\%$ compares spread across different scales, and it needs ratio data with a positive mean.

The **five-number summary** is minimum, Q1, median, Q3, maximum. Find the median, then take the median of the lower half (Q1) and of the upper half (Q3); with odd n, the median itself is left out of both halves. **IQR** = Q3 − Q1, the width of the middle 50%, and it is resistant. Software may use slightly different quartile rules; neither answer is wrong.

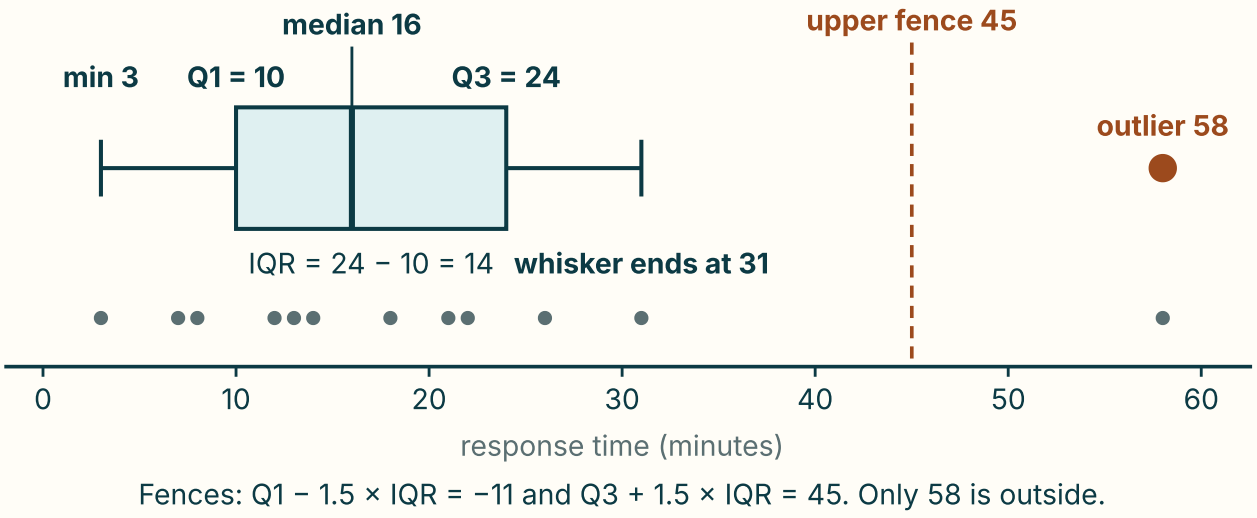
A **boxplot** draws a box from Q1 to Q3 with a line at the median. Fences sit at $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$. Values beyond a fence are flagged as outliers and drawn as dots, and whiskers reach the most extreme values still inside. For the book's response times, the fences are −11 and 45, so 58 is flagged and the whisker stops at 31. An outlier is a flag to investigate, never permission to delete.

A **z-score** $z = (x - \mu) \div \sigma$, or $(x - \bar{x}) \div s$, gives a value's distance from the mean in standard deviations. It has no units, so it compares values from different scales. Positive is above the mean, negative below. A whole standardized data set has mean 0 and SD 1. Backwards: $x = \mu + z\sigma$. Computing z needs no shape assumption; turning z into a percentile does.

For bell-shaped data the **empirical rule** says about 68%, 95% and 99.7% of values lie within 1, 2 and 3 SD of the mean (the exact 95% point is 1.96 SD). By symmetry about 16% lie above $\mu + \sigma$ and about 2.5% above $\mu + 2\sigma$. **Chebyshev's theorem** works for any shape: at least $1 - 1 \div k^2$ of the data lie within k SD, for $k > 1$. That is at least 75% for $k = 2$ and 88.9% for $k = 3$. Never apply 68, 95, 99.7 to skewed data.

LOCK THESE IN

- $\bar{x} = \Sigma x \div n$. Median: position $(n + 1) \div 2$, or the average of the two middle values.
- The median and IQR resist outliers; the mean, range and SD do not.
- $s^2 = \Sigma(x - \bar{x})^2 \div (n - 1)$; $\sigma^2 = \Sigma(x - \mu)^2 \div N$; SD = √variance.
- Deviations from the mean always sum to zero.
- IQR = Q3 − Q1. Fences: Q1 − 1.5 × IQR and Q3 + 1.5 × IQR.
- $z = (x - \mu) \div \sigma$, and $x = \mu + z\sigma$.
- Empirical rule (bell shapes only): 68%, 95%, 99.7% within 1, 2, 3 SD.
- Chebyshev (any shape): at least $1 - 1 \div k^2$; 75% at k = 2, 88.9% at k = 3.



The five-number summary drawn as a boxplot. The box spans the IQR from 10 to 24 with the median at 16. The whisker stops at 31, the last value inside the upper fence of 45, and 58 is flagged as an outlier.

Formulas and when to use them

Quantity	Formula	When to use it
Mean	$\bar{x} = \Sigma x \div n$	Symmetric quantitative data; later formulas
Median	Middle value; $(n + 1) \div 2$ position	Skewed or ordinal data; "typical"
Range	max − min	Quick look; fragile
Sample variance	$s^2 = \Sigma(x - \bar{x})^2 \div (n - 1)$	Spread of a sample
Population variance	$\sigma^2 = \Sigma(x - \mu)^2 \div N$	Every member measured
Standard deviation	$s = \sqrt{s^2}$; $\sigma = \sqrt{\sigma^2}$	Spread in original units
Coefficient of variation	$CV = (s \div \bar{x}) \times 100\%$	Comparing spread across scales; ratio data
IQR and fences	Q3 − Q1; Q1 − 1.5 IQR, Q3 + 1.5 IQR	Resistant spread; outlier flags
z-score	$z = (x - \mu) \div \sigma$	Comparing positions across scales
Back from z	$x = \mu + z\sigma$	Value at a given number of SDs
Chebyshev	at least $1 - 1 \div k^2$, $k > 1$	Any shape, unknown shape

The empirical rule and Chebyshev side by side

Interval	Bell-shaped data (about)	Any shape (at least)
μ ± 1σ	68% (68.27%)	No guarantee
μ ± 2σ	95% (95.45%)	75%
μ ± 3σ	99.7% (99.73%)	88.9%
μ ± 4σ	about 99.99%	93.75%
Above μ + 1σ	about 16%	Not set by the rule
Above μ + 2σ	about 2.5%	Not set by the rule
Exact 95% point	μ ± 1.96σ	Used in confidence intervals

Worked examples

WORKED EXAMPLE 1 · SAMPLE VARIANCE AND STANDARD DEVIATION BY HAND

Five patients wait 6, 9, 11, 14, 20 minutes. Find s^2 and s.

STEP 1 · MEAN

$\Sigma x = 60$, so $\bar{x} = 60 \div 5 = \mathbf{12}$.

STEP 2 · DEVIATIONS

$-6, -3, -1, 2, 8$. Check: they sum to 0.

STEP 3 · SQUARE AND ADD

$36 + 9 + 1 + 4 + 64 = SS = \mathbf{114}$.

STEP 4 · DIVIDE BY N – 1

$s^2 = 114 \div 4 = \mathbf{28.5}$ square minutes.

STEP 5 · SQUARE ROOT

$s = \sqrt{28.5} \approx \mathbf{5.34\ minutes}$. As a population the variance would be $114 \div 5 = 22.8$ and $\sigma \approx 4.77$, a little smaller.

WORKED EXAMPLE 2 · FIVE-NUMBER SUMMARY, FENCES AND OUTLIERS

Ten repair times in hours: 2, 5, 6, 8, 9, 11, 12, 14, 15, 33. Find the five-number summary, the fences and any outliers.

STEP 1 · MEDIAN

n = 10, so average the 5th and 6th values: (9 + 11) ÷ 2 = **10**.

STEP 2 · QUARTILES

Lower half 2, 5, 6, 8, 9: Q1 = **6**. Upper half 11, 12, 14, 15, 33: Q3 = **14**.

STEP 3 · IQR

14 − 6 = 8, and 1.5 × 8 = 12.0.

STEP 4 · FENCES

Lower: 6 − 12.0 = −6.0. Upper: 14 + 12.0 = 26.0.

STEP 5 · FLAG

Only 33 is above 26.0, so it is the one outlier; the right whisker stops at 15. Summary: min 2, Q1 6, median 10, Q3 14, max 33.

WORKED EXAMPLE 3 · WHICH SCORE IS MORE UNUSUAL?

Lena scores 78 on a quiz with mean 70 and SD 4, and 85 on a test with mean 76 and SD 6. Which performance is stronger relative to the class?

STEP 1 · QUIZ

z = (78 − 70) ÷ 4 = 8 ÷ 4 = **2.00**.

STEP 2 · TEST

z = (85 − 76) ÷ 6 = 9 ÷ 6 = **1.50**.

STEP 3 · COMPARE

2.00 SD beats 1.50 SD, so the quiz score of 78 is the stronger showing even though 85 is the bigger number.

STEP 4 · RUN IT BACKWARD

What test score would match the quiz? $x = \mu + z\sigma = 76 + 2.00(6) =$ **88**.

WORKED EXAMPLE 4 · EMPIRICAL RULE OR CHEBYSHEV?

Adult resting heart rates in a group are bell-shaped with μ = 72 and σ = 8. Clinic visit lengths are right-skewed with μ = 30 and σ = 10 minutes. What can you say about each within 2 SD?

STEP 1 · HEART RATE INTERVAL

72 ± 2(8) gives 56 to 88.

STEP 2 · APPLY THE EMPIRICAL RULE

The shape is bell, so about 95% fall between 56 and 88, and about 2.5% exceed 88.

STEP 3 · VISIT INTERVAL

30 ± 2(10) gives 10 to 50 minutes.

STEP 4 · APPLY CHEBYSHEV

The shape is skewed, so use $1 - 1 \div 2^2 = 0.75$: **at least 75%** of visits last 10 to 50 minutes. Claiming 95% here would be unjustified.

TRAPS PEOPLE FALL INTO

- **Calling the mean "typical" for skewed data.** Income, prices and wait times are skewed. The median is the typical value.
- **Dividing by n for a sample.** Sample variance divides by $n - 1$. Dividing by n underestimates σ^2.
- **Reading the variance as a distance.** It is in squared units. Take the square root to get the SD.
- **Including the median in both halves when n is odd.** The common textbook method leaves it out of both halves for Q1 and Q3.
- **Deleting a point because it is outside a fence.** The 1.5 × IQR rule flags values to investigate, not to throw away.
- **Treating z as a probability.** z is a location in SDs. A percentile needs a known shape.
- **Using 68, 95, 99.7 on skewed data.** If $\mu - 2\sigma$ is negative for a quantity that cannot be negative, the rule does not fit.
- **Applying Chebyshev with k = 1.** It gives 0%, which is true but useless. The theorem needs $k > 1$.

Chapter 4 self-test · ten items

1. For the data 5, 7, 7, 10, 16, which statement is true?

A. Mean 7, median 9, mode 7

B. Mean 9, median 7, mode 10

C. Mean 9, median 10, mode 7

D. Mean 9, median 7, mode 7

D. Mean 9, median 7, mode 7 Sum = 45, so the mean is 45 ÷ 5 = 9. The 3rd sorted value is 7, and 7 is the only repeat.

2. Which measure is resistant to outliers?

A. The standard deviation
B. The range
C. The mean
D. The interquartile range

D. The interquartile range The IQR uses only the middle half of the data, so extreme values do not move it.

3. Find the sample standard deviation of 2, 4, 6, 8.

A. 2.24
B. 6.67
C. 2.58
D. 5.00

C. 2.58 $\bar{x} = 5$. Squared deviations 9, 1, 1, 9 sum to 20. $s^2 = 20 \div 3 = 6.67$, so $s \approx 2.58$. Dividing by 4 instead gives 2.24.

4. Why do we square the deviations before averaging?

A. Squaring keeps the original units
B. Squaring removes the outliers
C. The deviations always sum to zero
D. The mean must be squared too

C. The deviations always sum to zero Positive and negative deviations cancel exactly. Squaring makes them all positive.

5. Two machines have means 50 and 500 with SDs 5 and 25. Which is more variable relative to its mean?

A. The second, CV 5% versus 10%
B. The first, CV 10% versus 5%
C. Both are equally variable
D. The second, since 25 is larger

B. The first, CV 10% versus 5% CV = s ÷ x̄: 5 ÷ 50 = 10% and 25 ÷ 500 = 5%. Raw SDs cannot be compared across different scales.

6. Q1 = 20 and Q3 = 32. Which value is an outlier by the 1.5 × IQR rule?

A. 51
B. 49
C. 3
D. 14

A. 51 IQR = 12 and 1.5 × 12 = 18. Fences: 20 − 18 = 2 and 32 + 18 = 50. Only 51 lies outside.

7. A value has z = −1.5 in a distribution with μ = 60 and σ = 6. What is the value?

A. 69
B. 51
C. 58.5
D. 54

B. 51 $x = \mu + z\sigma = 60 + (-1.5)(6) = 60 - 9 = 51$.

8. Scores are bell-shaped with μ = 100 and σ = 15. About what share exceed 115?

A. 34%
B. 32%
C. 2.5%
D. 16%

D. 16% About 68% lie between 85 and 115, so 32% lie outside, split evenly: about 16% above 115.

9. Wait times are strongly right-skewed. What share must lie within 3 SD of the mean?

A. At least 88.9%
B. About 99.7%
C. Exactly 95%
D. At least 75%

A. At least 88.9% With an unknown or skewed shape, use Chebyshev: $1 - 1 \div 3^2 = 1 - 1/9 \approx 0.889$.

10. A standardized data set always has which mean and SD?

A. Mean 0 and SD 1
B. Mean 1 and SD 0
C. Mean 0 and SD 0
D. The original mean and SD

A. Mean 0 and SD 1 Subtracting the mean centers the data at 0, and dividing by the SD makes the spread 1.

Cram sheet

Everything from chapters 3 and 4 in one place. Cover the right column and see if you can rebuild each row.

Formulas

Name	Formula	Note
Relative frequency	f ÷ n	Column sums to 1
Class width	(max − min) ÷ classes, round up	5 to 10 classes
Pie angle	(f ÷ n) × 360°	1% = 3.6°
Lie factor	apparent ÷ actual effect	Honest is about 1
Mean and median	Σx ÷ n; middle sorted value	Mean not resistant
Variance	$s^2 = SS \div (n - 1)$; $\sigma^2 = SS \div N$	$SS = \Sigma(x - \bar{x})^2$
CV	(s ÷ x̄) × 100%	Ratio data only
IQR and fences	Q3 − Q1; Q1 − 1.5 IQR; Q3 + 1.5 IQR	Beyond a fence = flag
z-score	$z = (x - \mu) \div \sigma$; $x = \mu + z\sigma$	Unitless
Chebyshev	$\geq 1 - 1 \div k^2$	75% at k = 2; 88.9% at k = 3

Rules that are not formulas

Topic	The rule
Checks	Frequencies sum to n; relative to 1; cumulative ends at n.
Classes	No overlaps, no gaps; round the width up.
Bars	Bar chart gaps for categories; histogram bars touch; both start at zero.
Pie	Only for exclusive, exhaustive categories.
Skew	Named for the tail; right skew pulls the mean above the median.
Bimodal	Two groups mixed; split before summarizing.
Choosing center	Mean for symmetric; median for skewed or ordinal; mode for categories.
Quartiles	Median of each half; leave the median out when n is odd.
Empirical rule	68, 95, 99.7 within 1, 2, 3 SD, bell shapes only.

Dr. Garg Learning · CC BY-SA 4.0