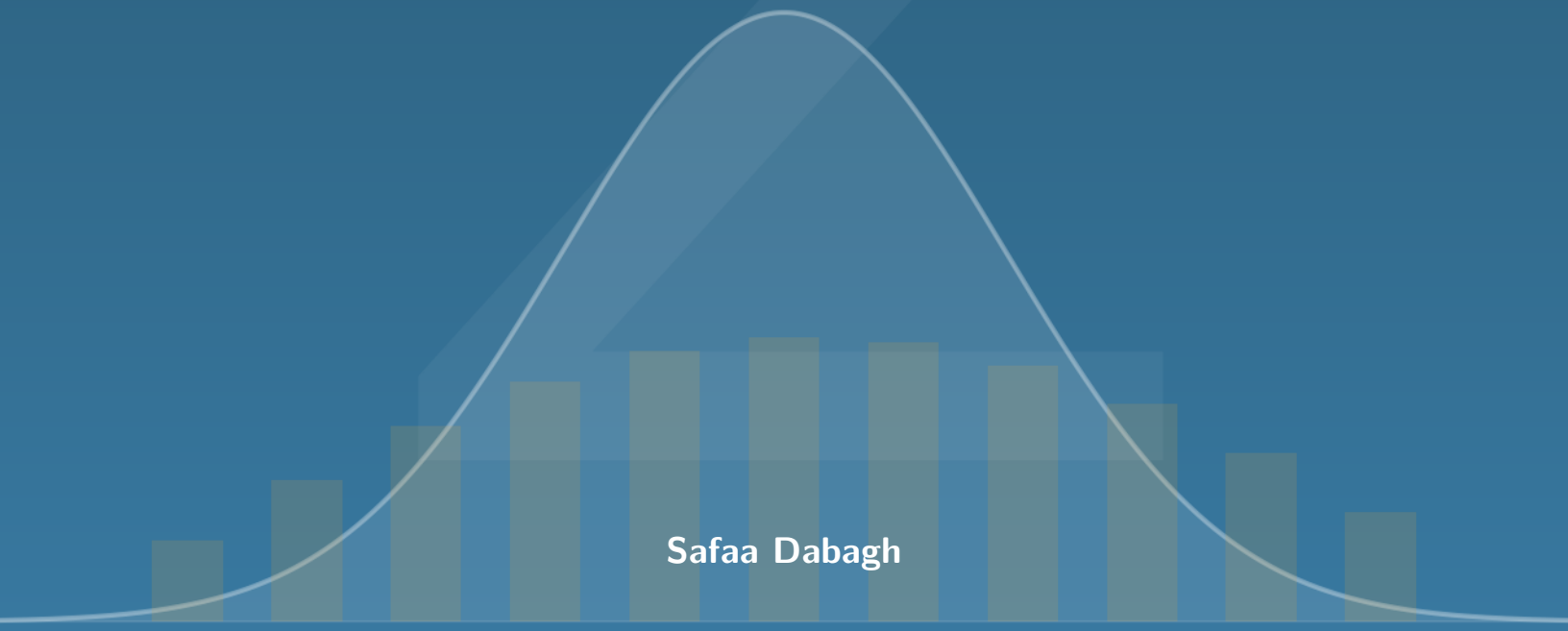
You're Smarter Than You Think STATISTICS

Introduction to Statistics ▪ Community College Edition

for Neurodivergent Learners

"You are not behind. You are not broken.
You are not 'bad at math.' That changes now."



Safaa Dabagh

You're Smarter Than You Think

Statistics

Introduction to Statistics ▪ Community College Edition

An Open Educational Resource for Neurodivergent Learners

Safaa Dabagh

Written for introductory statistics students at any college

"You are not behind. You are not broken.
You are not 'bad at math.' That changes now."

2026

A Note Before We Begin

A student once emailed me the night before our first exam. The whole message was:

> *"I don't think I'm a stats person. I'm probably going to fail. I just wanted you to know I tried."*

That student earned a B+ in the course.

The gap between what they *believed* about themselves and what they were actually capable of — that gap is exactly why I wrote this book. Statistics did not defeat that student. The story they had been told about themselves almost did.

This book grew out of years of teaching introductory statistics in community college classrooms. If you have ever felt your brain go quiet the moment numbers appear, or solved a problem correctly and assumed you got lucky, this book was written for *you*.

This book is for you if:

- You've ever said "I'm not a math person."
- Formulas make your brain slam shut before you've even read them.
- You understand it in class, then feel it vanish by the time you get to the parking lot.
- You've been told — by a teacher, a test, or your own inner voice — that you're behind.

You are not behind. You are not broken. You are not "bad at math." You are about to do statistics — and you are going to be good at it.

How This Book Works

Your brain learns better when it knows what's coming. So every chapter and every section follows the **same rhythm** — no surprises, no hidden steps.

Every chapter opens with a warm-up:

Math Skills You'll Need

The exact arithmetic and algebra this chapter uses — refreshed, with quick practice and answers — *before* any statistics. No more getting stuck on a fraction in the middle of a big idea.

Then every section follows these six steps:

1. **Read This First** — A real-life moment that *is* the concept, before we ever name it.
2. **Let's Talk About It** — A question to think about. There is no wrong answer here.
3. **Now We Name It** — We give the idea its official statistics name, and connect it back to what you already understood.
4. **Watch It Work** — A worked example, every step explained — including the steps most books skip.
5. **Your Turn** — You try it. When your brain says "I'm probably wrong," you keep going anyway. (Answers at the end of each chapter.)
6. **Check Your Voice** — A moment to notice what your inner voice said — and talk back to it.

Three more things you can count on:

Technology Companions (free, online)

You will never fight software in these pages. When you want to do a chapter's work on a computer or a calculator, four free step-by-step companions are waiting at `learnwithoutwalls.com`, one each for **R and RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. Each is keyed to these chapters and comes with ready-made datasets. A short pointer at the end of every chapter reminds you where to look. No paid subscriptions, ever.

Study Guide & Practice (after each Part)

After each group of chapters, a Study Guide pulls the big ideas together and gives you a set of mixed practice problems — with full answers — so you can rehearse before an exam exactly the way the exam will feel. (The first one reviews Chapters 1–4 together.)

Formula Sheet (after each Part)

Every formula from the Part, gathered on one page, with each symbol named in plain words and a note on *when* to use it — so you're never hunting through chapters mid-problem, and you have a ready-made reference for exams.

A note for instructors: this book is software-agnostic and uses everyday, generic data, so it drops into any introductory statistics course at any college. Use your own class data if you like — the examples are built to be swapped.

What You'll Be Able to Do

This book is built around what you'll be able to *do* by the end — the learning outcomes shared by introductory statistics courses at community colleges everywhere. Everything in these pages serves one of them.

STUDY GUIDE & PRACTICE

By the end of this book, you will be able to:

SLO 1. Describe data. Given a data set, analyze it and present the information using tables, graphs, and statistical calculations. *(Parts I–II)*

SLO 2. Estimate. Given sample data, choose and use appropriate estimation strategies to make inferences about a population's mean, proportion, and variation. *(Part IV)*

SLO 3. Test. Given sample data, choose and use an appropriate test to reach a conclusion about a hypothesis about a population parameter. *(Part V)*

Two outcomes run through every chapter: *use technology* to analyze data (through the free **Technology Companions**, one per tool, at learnwithoutwalls.com), and *think critically*, interpreting results in context and questioning statistical claims.

How the Parts build toward those outcomes:

Part	What it builds	Serves
I. Producing Data	where data comes from (foundation)	all SLOs
II. Describing Data	tables, graphs, numerical summaries	SLO 1
III. Probability & Distributions	the bridge to inference	SLO 2, 3
IV. Estimating with Confidence	confidence intervals (mean, proportion, variation)	SLO 2
V. Testing Claims	hypothesis tests: one, two, three+ parameters	SLO 3
VI. Categorical Data & Relationships	chi-square; regression & correlation inference	SLO 3

A Study Guide & Practice checkpoint and a one-page Formula Sheet follow each Part (Parts I and II share the first checkpoint, covering Chapters 1–4).

Part I

Producing Data — Where Numbers Come From

Chapter 1

What's the Story in the Data?

Statistics, Data, Variables, and the Difference Between Everyone and Some People

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll look at a spreadsheet, a news headline, or a survey and think: "Oh — I already think this way. Statistics just gives it precise names."

The Story. You collect, sort, and draw conclusions from data every single day. Statistics is simply doing that *on purpose*, and carefully.

The Skills You'll Have:

- Say what statistics actually is (and what it isn't)
- Tell the difference between a *population* and a *sample*
- Spot a *variable* and say whether it's words or numbers
- Read the vocabulary on a data table without panic

The Confidence Moment. When someone says "the sample mean," your brain will calmly translate: "the average of the slice of people we actually measured."

The Bridge. Chapter 2 asks *where data comes from* — because a beautiful analysis of badly collected data is still wrong. But first, we learn the language.

MATH SKILLS YOU'LL NEED

Statistics in this chapter mostly asks you to *read* numbers and turn counts into percentages. Here's the only math you need — let's warm it up.

Skill 1 — Turning a fraction into a percent. A percent is just "out of 100." Divide, then move the decimal two places right (that's $\times 100$).

$$\frac{18}{40} = 0.45 = 45\%$$

Skill 2 — Rounding. "Round to one decimal place" means keep one digit after the dot. Look at the *next* digit: 5 or more rounds up, 4 or less stays.

$$0.4538 \to 0.5 \qquad 12.43 \to 12.4$$

Skill 3 — Reading a table. A table is rows and columns. Find the row, then slide across to the column you want.

Practice (answers at the end of the chapter):

1. Write $\frac{27}{50}$ as a percent.
2. Round 63.857 to one decimal place.
3. Out of 200 students, 30 walk to campus. What percent walk?

1.1 You're Already a Data Collector

▪ READ THIS FIRST

Open your phone and look at your screen-time summary. It knows how many hours you spent on each app this week, which day was your highest, and whether you're up or down from last week. Nobody taught your phone "statistics." It just did three things: it *collected* numbers (minutes per app), it *organized* them (by app, by day), and it *drew a conclusion* ("up 12% from last week"). That's the whole field. That's statistics. You've been living inside it.

▪ LET'S TALK ABOUT IT

Think of one number that gets tracked about your life — steps, hours of sleep, money in an account, likes on a post. Who collects it? What decision does someone make because of it? (There's no wrong answer — just notice how often a number stands in for a story.)

▪ NOW WE NAME IT

Statistics is the science of collecting, organizing, analyzing, and interpreting data to answer a question or make a decision.

The raw material is **data**: the recorded facts or measurements themselves (the minutes, the steps, the answers on a survey). One piece of data is a *data value*; all of them together are a *data set*.

Notice the four verbs — collect, organize, analyze, interpret. Most of this book lives in "analyze" and "interpret," but a number only means something if you remember where it came from. Hold onto that; it's the heart of Chapter 2.

▪ WATCH IT WORK

Question: A coffee shop records how many drinks each customer buys one morning: 1, 2, 1, 3, 1, 2. What's the data, and what's one conclusion?

Step 1 — Identify the data. The data set is the six recorded values: $1, 2, 1, 3, 1, 2$. Each number is one customer's purchase.

Step 2 — Organize it. Count how often each value appears: one drink (3 customers), two drinks (2 customers), three drinks (1 customer).

Step 3 — Interpret it. Most customers bought a single drink that morning. A conclusion the manager might draw: "Solo-drink buyers are our typical morning customer."

That's collect $\rightarrow$ organize $\rightarrow$ interpret, on six numbers. Bigger data sets use the same three moves.

▪ YOUR TURN

A fitness app logs your steps for five days: 4,000, 7,500, 6,000, 9,000, 6,000.

1. What is the data set?
2. Which step-count shows up more than once?
3. Write one sentence interpreting your week.

Work space:

▪ CHECK YOUR VOICE

If a small voice just said "this is too easy, the real math is coming to get me" — notice that. Easy is allowed. Understanding the foundation *is* doing the work. Talk back: "I understood that because I'm capable, not because it was a trick."

1.2 Population vs. Sample (The Whole vs. The Slice)

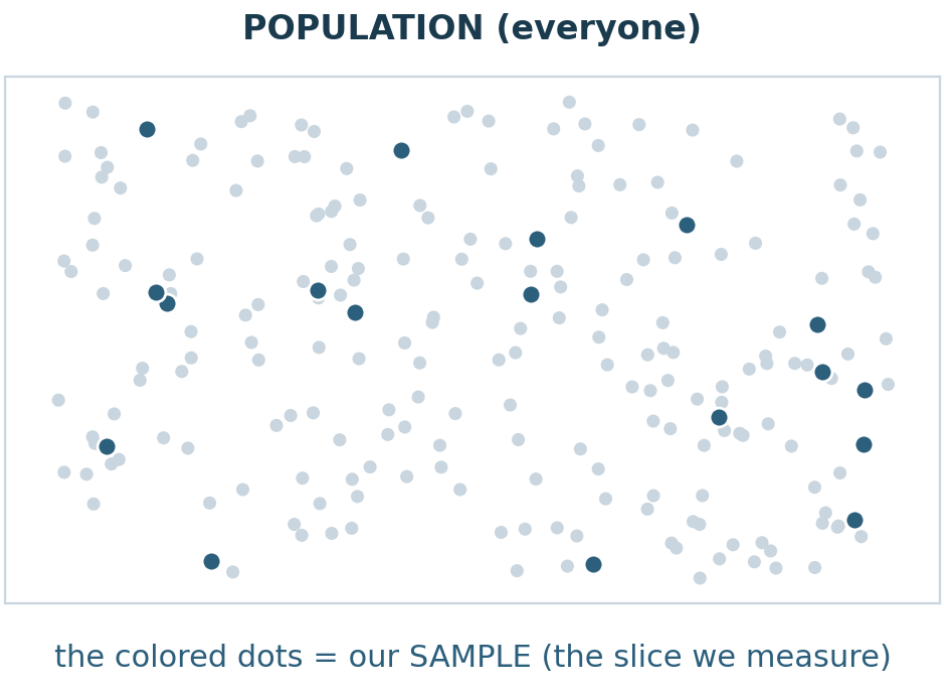


Figure 1.1. A sample is a slice of the population — the colored dots are the few we actually measure.

▪ READ THIS FIRST

You're about to make a big pot of soup for a party. Before you serve forty people, you taste *one spoonful*. From that single spoonful you decide whether the whole pot needs more salt.

You didn't taste all of it — that would leave nothing for the guests. You tasted a small, well-stirred *slice* and trusted it to represent the whole pot.

That spoonful-to-pot relationship is the single most important idea in this entire course.

▪ LET'S TALK ABOUT IT

Why did you stir the soup before tasting? What would go wrong if you only tasted the very top, or only the bottom? Keep that answer in mind — it's secretly the reason sampling can go wrong.

▪ NOW WE NAME IT

The **population** is the *entire* group you want to know about — the whole pot of soup, all 2,500 students at the college, every voter in California.

The **sample** is the smaller part you actually measure — the spoonful, the 80 students you surveyed, the 1,000 voters who were polled.

We study samples because measuring the whole population is usually impossible, too expensive, or too slow. The entire art of statistics is using a *slice* to make a trustworthy statement about the *whole* — and knowing how much to trust it.

▪ WATCH IT WORK

Question: A college has 2,500 students. A researcher surveys 80 of them about hours slept. Identify the population and the sample.

Step 1 — Ask: what group do we actually care about? All 2,500 students. That is the *population*.

Step 2 — Ask: who did we actually measure? The 80 students who were surveyed. That is the *sample*.

Step 3 — Say the relationship in plain words. "We'll use the 80 students' sleep (the slice) to estimate the typical sleep of all 2,500 students (the whole pot)."

The number describing the whole population (like the true average for all 2,500) is called a **parameter**. The number we calculate from our sample (the average of the 80) is a **statistic**.

Memory hook: **p**opulation → **p**arameter, **s**ample → **s**tatistic.

▪ YOUR TURN

A factory makes 50,000 phone chargers in a day. A quality inspector tests 200 of them and finds 6 defective.

1. What is the population?
2. What is the sample?
3. The "6 out of 200" defect rate — is that a parameter or a statistic? Why?

Work space:

▪ CHECK YOUR VOICE

Mixing up population and sample at first is not a sign you "don't get it." It's a sign you're learning a brand-new pair of words for an idea you already understood (the spoonful and the pot). If you felt unsure, reread the soup story — your intuition was already correct.

1.3 Types of Variables (Words vs. Numbers)

▪ READ THIS FIRST

Think about the last form you filled out — maybe signing up for an account or a doctor's visit. Some boxes you *checked* (Male/Female/Other, or Yes/No). Some boxes you *typed a number into* (your age, your weight, your zip code).

Your brain already knew, without being told, which boxes hold *labels* and which hold *amounts*. That instinct is the whole idea of this section.

▪ LET'S TALK ABOUT IT

Look at the contact card for a friend in your phone: name, phone number, birthday, "favorite" star, number of times you've called. Which of those are labels, and which are real amounts you could add or average? Is a phone number really a "number"?

▪ NOW WE NAME IT

A **variable** is any characteristic that can change from one individual to the next — age, major, eye color, hours of sleep.

Variables come in two types:

- A **categorical** (or qualitative) variable records a *label* or category: major, eye color, Yes/No, zip code.
- A **quantitative** variable records a *measured amount* you can do arithmetic on: age, hours of sleep, income.

The test that never fails: ask, "Does it make sense to *average* it?" Averaging ages (24.3 years) makes sense — quantitative. "Averaging" zip codes or eye colors is nonsense — categorical. A phone number is made of digits but you'd never average it, so it's *categorical*.

Quantitative variables split once more: **discrete** ones come from counting and land on whole numbers (number of pets, siblings, text messages), while **continuous** ones come from measuring and can take any value in a range (height, time, temperature).

▪ WATCH IT WORK

Question: Classify each variable as categorical or quantitative (and if quantitative, discrete or continuous): (a) T-shirt size, (b) number of siblings, (c) body temperature, (d) zip code.

(a) T-shirt size (S, M, L, XL). These are labels — you can't average "M." *Categorical.*

(b) Number of siblings. A count; averaging makes sense ("2.1 siblings on average"). *Quantitative, discrete* (you can't have 2.1 actual siblings — it comes from counting).

(c) Body temperature. A measured amount on a scale; averaging makes sense, and it can be 98.6° or anything between. *Quantitative, continuous.*

(d) Zip code. Looks like a number, but averaging zip codes is meaningless — it's a *label* for a place. *Categorical.* (This one fools almost everyone the first time. Now it won't fool you.)

▪ YOUR TURN

Classify each as categorical or quantitative (and if quantitative, discrete or continuous):

1. Eye color
2. Number of cups of coffee you drank today
3. Your weight
4. Jersey number on a sports team

Work space:

▪ CHECK YOUR VOICE

If the zip-code one (or the jersey number) tripped you up — good. That means you found the exact spot where "looks like a number" and "is an amount" part ways. Noticing the trap is how you stop falling into it. You didn't get it "wrong"; you just met the trick once so it can't surprise you on the exam.

1.4 Reading a Data Table Without Panic

▪ READ THIS FIRST

Picture a class sign-up sheet. Down the left side: one row per person. Across the top: the things we wrote about each person — name, email, t-shirt size, hours available.

A statistics "data set" is exactly that sign-up sheet. Rows are *who*, columns are *what we recorded*. You've read a hundred of these. A data table is not a new thing — it just has new names.

▪ LET'S TALK ABOUT IT

Think of any spreadsheet or table you've seen recently — a gradebook, a sports standings page, a menu. What did each *row* stand for? What did each *column* record? Once you see "rows = who, columns = what," tables stop being scary.

▪ NOW WE NAME IT

In a data table:

- Each **row** is one **case** (also called an individual or observation) — one person, one phone, one game.
- Each **column** is one **variable** — one characteristic recorded for every case.
- Each **cell** is one **data value** — a single case's value for a single variable.

The size of a data set is usually described as "n cases" — so "$n = 80$" means 80 rows, 80 individuals. When someone says "a data set with 80 observations and 5 variables," picture a table 80 rows tall and 5 columns wide.

▪ WATCH IT WORK

Question: Here is a tiny data set of four students.

Name	Major	Age	Works a Job?
Mia	Nursing	19	Yes
Leo	Business	27	No
Ava	Biology	22	Yes
Sam	Undeclared	20	Yes

Step 1 — How many cases? Four rows $\rightarrow n = 4$ students.

Step 2 — What are the variables? The columns: Name, Major, Age, Works a Job? — four variables recorded for each student.

Step 3 — Classify each variable. Name = categorical (a label). Major = categorical. Age = quantitative (you can average it). Works a Job? = categorical (Yes/No).

Step 4 — Read one cell. The cell in Ava's row under "Age" is the data value 22 — Ava's age. One case, one variable, one value.

▪ YOUR TURN

Using the same four-student table above:

1. How many *variables* are in the data set?
2. Is "Major" categorical or quantitative?
3. What is the data value in Leo's row under "Age"?
4. If we added a column "Hours of sleep last night," would it be categorical or quantitative?

Work space:

▪ CHECK YOUR VOICE

Notice you just read a data table and answered four questions about it — the same kind of table that, a few pages ago, might have made your eyes glaze over. That wasn't luck. You learned the three words (case, variable, value) and the whole table opened up. Say it: "I can read data."

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 1 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 1 — Your Turn Answers

Math Skills: (1) $\frac{27}{50} = 0.54 = 54\%$. (2) 63.9. (3) $\frac{30}{200} = 0.15 = 15\%$.

Section 1.1: (1) The data set is $\{4000, 7500, 6000, 9000, 6000\}$. (2) $6{,}000$ steps (it appears twice). (3) Example: "My most common day was about 6,000 steps, with one strong 9,000-step day."

Section 1.2: (1) Population: all 50,000 chargers made that day. (2) Sample: the 200 tested. (3) A *statistic* — it's calculated from the sample we measured, not from all 50,000.

Section 1.3: (1) Eye color = categorical. (2) Cups of coffee = quantitative, discrete (a count). (3) Weight = quantitative, continuous (a measurement). (4) Jersey number = categorical (a label; averaging jersey numbers is meaningless).

Section 1.4: (1) Four variables (Name, Major, Age, Works a Job?). (2) Categorical. (3) 27. (4) Quantitative (continuous) — hours of sleep is a measured amount you can average.

Chapter 2

Where Did This Data Come From?

Sampling, Bias, and the Difference Between Watching and Doing

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll see a headline like "9 out of 10 people love this!" and instantly ask the question that matters most: "Which 9 people? How were they picked?" You'll have a built-in radar for data you shouldn't trust.

The Story. A perfect calculation on a badly collected sample is still wrong. Where the data *came from* can quietly decide the answer before anyone does a single computation. This chapter teaches you to check the source.

The Skills You'll Have:

- Spot the four big ways a sample lies (and name each one)
- Explain why *random* selection is the cure for bias
- Tell an *observational study* apart from an *experiment*
- Catch a *confounding variable* and stop saying "causes" when you mean "goes with"
- Recognize the parts of a fair experiment: control, randomization, blinding

The Confidence Moment. When a friend says "ice cream must cause drowning — look, they rise together!", you'll calmly answer: "That's summer doing both. Correlation isn't causation."

The Bridge. Chapter 3 turns clean data into *graphs* you can read at a glance. But a graph of bad data is just a prettier lie — so first we make sure the data is honest.

MATH SKILLS YOU'LL NEED

This chapter leans on two simple skills: turning counts into percentages (to measure how much of your sample actually answered), and the idea of a *random number*. Let's warm them up.

Skill 1 — Response rate as a percent. If you contact people and only some reply, the *response rate* is how many replied out of how many you contacted, written as a percent.

$$\text{response rate} = \frac{\text{number who responded}}{\text{number contacted}} \times 100\%$$

For example, mail 500 surveys, get 90 back: $\frac{90}{500} = 0.18 = 18\%$. A low response rate is a warning sign — the people who *didn't* answer might be very different from the ones who did.

Skill 2 — A random number. A **random number** is one picked by pure chance, where every option in the range has an equal shot. Think of rolling a fair die: each face 1–6 has the same chance. We use random numbers to choose *who* ends up in a sample, so no human bias sneaks in.

Skill 3 — Random selection. To pick 3 people at random from a list of 10, you could number them 1–10, then draw 3 different numbers by chance (slips in a hat, or a computer). "At random" means no favorites — the chance doesn't depend on who's tall, friendly, or sitting in front.

Practice (answers at the end of the chapter):

1. A poll emails 800 people; 240 respond. What is the response rate as a percent?
2. You want one random student from a list of 25. Describe one fair way to pick.
3. Out of 1,500 contacted voters, 1,200 answered. What percent did *not* respond?

2.1 When the Slice Lies

▪ READ THIS FIRST

A morning TV show puts a question on screen: "Text YES or NO — do you think mornings should start later?" Thousands of people text in, and the host announces "78% say later!"

But who actually texted? People who felt strongly enough to grab their phone — mostly the ones annoyed about early mornings. The calm, fine-with-it majority just kept drinking their coffee.

The number is real. The conclusion is junk. The slice they measured was never the whole pot.

▪ LET'S TALK ABOUT IT

Think of the last online poll or "rate us" pop-up you saw. Who tends to actually fill those out — people who feel nothing, or people who feel something (great or terrible)? If only the people with strong feelings answer, what happens to the result?

▪ NOW WE NAME IT

When the way you collect a sample makes it systematically miss part of the population, you have **sampling bias**: the sample leans in a direction that doesn't match the whole. Here are the four big culprits.

- **Voluntary-response bias**: you let people *choose* to include themselves (call-in polls, online reviews). The people with strong opinions volunteer; everyone else stays silent, so the sample over-represents the loud ends.
- **Nonresponse bias**: you reach out to a fair group, but a chunk *doesn't answer* — and the non-answerers differ from the answerers. A survey on free time that only busy people ignore will look like everyone has lots of free time.
- **Undercoverage**: part of the population has little or no chance of being chosen at all. Surveying only people with landline phones *undercovers* younger people who use only cell phones.
- **Convenience sampling**: you grab whoever is easiest to reach (the first 20 people you pass). Easy to do, but the "easy" group is rarely a fair picture of everyone.

The thread connecting all four: the bias is baked in by the *method*, not fixed by collecting more responses. A bigger biased sample is just a bigger wrong answer.

▪ WATCH IT WORK

Question: A reporter stands outside a busy gym at 6 a.m. and asks people leaving, "How many days a week do you exercise?" She concludes, "The average adult exercises 5 days a week." What's wrong?

Step 1 — Name the population she claims to describe. "The average adult" — meaning all adults.

Step 2 — Name who she actually sampled. People *leaving a gym at 6 a.m.* — a group that exercises far more than typical adults.

Step 3 — Identify the bias type. The sample systematically misses non-gym-goers, who had essentially no chance of being asked. That's *undercoverage*, and because she grabbed whoever was handy, it's also *convenience sampling*.

Step 4 — State the effect. The estimate is biased *upward*: it will overstate how much the average adult exercises, because gym-leavers are not a fair slice of all adults.

The fix is not "ask more gym-goers." It's "ask a fair cross-section of all adults" — which is exactly where Section 2.2 is headed.

▪ YOUR TURN

For each situation, name the most clearly present type of sampling bias and say which direction the result is likely skewed.

1. A website asks visitors to "rate your experience"; only 4% click, mostly people who were furious or delighted.
2. A city mails 2,000 surveys about satisfaction with bus service; the 300 who reply are mostly daily riders.
3. A study on phone use surveys only people who answer their home landline.
4. A student polls the three friends sitting next to her about how hard the exam was.

Work space:

▪ CHECK YOUR VOICE

If a voice said "I should already know the official name for each kind of bias" — you're learning them right now, and that's the point. Honestly, most adults can *feel* that a call-in poll is sketchy without ever naming the reason. You're just adding the precise labels to instincts you already have. That's an upgrade, not a test you failed.

2.2 Random Is the Goal

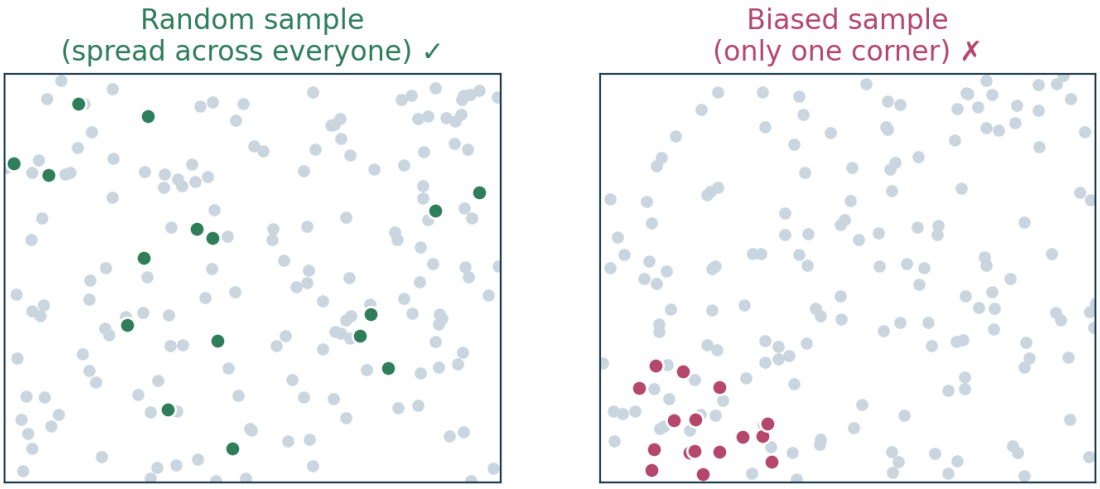


Figure 2.1. A random sample spreads across the whole population; a biased sample over-represents one corner.

▪ READ THIS FIRST

Back to the soup from Chapter 1. Why did you *stir* before tasting? Because an unstirred pot has salt pooled at the bottom — a spoonful from the top would fool you.

Stirring mixes everything so your single spoonful has a fair shot at every part of the pot. In sampling, the stirring is called *randomness*. It's how you give every person an equal chance to land in your spoonful.

▪ LET'S TALK ABOUT IT

Imagine you have to pick 5 students from your class of 30 to represent everyone. If you just pick the 5 you know best, what goes wrong? What could you do instead so that the quiet kid in the back has exactly the same chance as your best friend?

▪ NOW WE NAME IT

A **simple random sample** (SRS) of size n is a sample chosen so that *every individual in the population has an equal chance of being selected* — and, more precisely, every possible group of size n is equally likely.

Why does this defeat bias? Because the selection no longer depends on who's loud, convenient, or easy to reach. Chance has no opinion. Over many draws, a random method has *no systematic tendency* to over- or under-pick any group, so the sample tends to mirror the population.

A few other random (probability) methods you'll hear about:

- **Stratified sampling**: split the population into groups ("strata") that share a trait — say, class year — then take a random sample *within each group*. Good when you want every group represented in the right proportion.
- **Cluster sampling**: split the population into many similar groups ("clusters") — say, sections of a course — randomly pick a few whole clusters, and measure everyone in them. Cheaper when people are already grouped.
- **Systematic sampling**: order the population and pick every kth individual (every 10th name on a list) after a random start.

The line that matters: *convenience sampling is not random.* Grabbing whoever's nearest gives some people a much higher chance than others, so it carries bias. "Random" is a specific promise about *equal chance*, not a synonym for "haphazard."

▪ WATCH IT WORK

Question: A club has 50 members, numbered 1 to 50. You want a simple random sample of 5. How do you do it, and why is it an SRS?

Step 1 — Number every individual. Each member already has a number 1–50. Every member has exactly one number, and every number belongs to one member.

Step 2 — Generate random numbers. Use a computer (or slips in a hat) to draw 5 *different* numbers between 1 and 50 — say $7, 23, 31, 42, 48$.

Step 3 — Match numbers to people. The members holding those five numbers are your sample.

Step 4 — Check the SRS condition. Before drawing, every member had the same chance ($\frac{5}{50} = \frac{1}{10}$) of being chosen, and no group of 5 was favored over any other. Equal chance for all $\Rightarrow$ this is a simple random sample.

Notice what we did *not* do: pick the 5 members who showed up early, or the 5 we know best. Those would tilt the sample. The random draw is the stirring.

▪ YOUR TURN

1. In your own words, what makes a sample a *simple random sample*?
2. A school divides students by grade level (9, 10, 11, 12) and randomly samples 25 students from *each* grade. Which sampling method is this?
3. A factory tests every 100th item coming off the line, starting from a random item in the first 100. Which method is this?
4. True or false: "I surveyed the first 40 people who walked into the cafeteria" is a random sample. Explain.

Work space:

▪ CHECK YOUR VOICE

If the four method names (simple, stratified, cluster, systematic) feel like a lot to hold at once — you don't have to memorize them as a list. Anchor on one idea: *random means equal chance*. Every method here is just a different way of keeping that promise. If you understand *why* we stir the soup, you already understand the whole section. The vocabulary will settle in with practice.

2.3 Watching vs. Doing

▪ READ THIS FIRST

Two scientists want to know if a new study app raises test scores.

The first *watches*: she finds students who already use the app, compares their scores to students who don't, and notes a difference. The second *does*: she takes a group of students, randomly assigns half to use the app and half not, then compares.

Same question, two very different setups — and only one of them can actually claim the app *caused* anything.

▪ LET'S TALK ABOUT IT

Students who *choose* to use a study app on their own — how might they already be different from students who don't, before the app does a single thing? List two or three differences. (This is the crack that the whole section pours through.)

▪ NOW WE NAME IT

An **observational study** measures individuals without trying to influence them — you *watch* and record what's already happening. An **experiment** deliberately *imposes a treatment* on individuals and then measures the response — you *do* something and see what changes.

To talk about cause, name the two roles a variable can play:

- The **explanatory variable** is the one you think might cause or explain a change (using the study app).
- The **response variable** is the outcome you measure (the test score).

Here's the trap. In an observational study, a third variable can be quietly driving both. A **confounding variable** (or lurking variable) is one tied to both the explanatory and the response variable, so you can't tell which one is really doing the work. Maybe the app-users are also the more motivated students — motivation is confounded with app use, and motivation alone could raise the scores.

That's why **correlation is not causation**: two things rising and falling together (correlated) does *not* prove one causes the other. A confounder, or pure coincidence, can produce the same pattern. Only a well-designed *experiment* — where you assign the treatment yourself — can support a real cause-and-effect claim.

▪ WATCH IT WORK

Question: Across a year, a town finds that as monthly ice cream sales go up, drownings also go up. The two are strongly correlated. Does eating ice cream cause drowning?

Step 1 — Name the variables. Explanatory (claimed cause) = ice cream sales. Response (outcome) = number of drownings.

Step 2 — Look for a lurking third variable. What makes *both* go up at the same time? **Summer / hot weather.** In hot months people buy more ice cream *and* more people swim, so more drownings occur.

Step 3 — Name what's happening. Temperature is a *confounding variable*: it's linked to both ice cream sales and drownings. Ice cream and drowning move together only because summer moves both.

Step 4 — State the correct conclusion. Ice cream sales and drownings are *correlated*, but ice cream does *not* cause drowning. Correlation is not causation. This was an observational study, so it cannot establish cause — and the obvious confounder shows exactly why.

▪ YOUR TURN

1. A researcher records the diets of 1,000 people and their health, changing nothing. Is this an observational study or an experiment?
2. In "Does more sleep improve reaction time?", name the explanatory variable and the response variable.
3. Children with bigger feet tend to read better. Eating ice cream isn't the culprit here — what *lurking variable* explains both bigger feet and better reading?
4. True or false: a strong correlation between two variables proves one causes the other. Explain.

Work space:

▪ CHECK YOUR VOICE

If you read "correlation is not causation" and thought "okay but it really *seems* like one causes the other" — that instinct is human, and it's exactly the instinct good statistics protects you from. Your brain is built to see causes everywhere; spotting the lurking variable is you out-thinking that reflex. The fact that the ice cream example felt tricky for a second means you were taking it seriously. That's the work.

2.4 Building a Fair Experiment

▪ READ THIS FIRST

Say you want to test whether a new vitamin boosts energy. You give it to 100 people, and a month later most report feeling more energetic. Proof?

Not yet. Maybe they'd feel better anyway (it's spring now). Maybe just *believing* a pill helps made them feel better. Without something to compare against, you can't tell the vitamin's effect from everything else going on.

A fair experiment is built precisely to rule out those other explanations.

▪ LET'S TALK ABOUT IT

If everyone in your study gets the new vitamin, what would you compare their energy to? Why might it matter that the people taking the pill *know* they're taking a special new vitamin?

▪ NOW WE NAME IT

A trustworthy experiment is built from a few key pieces:

- A **control group** gets no treatment (or a standard one), giving you a baseline to compare the treatment group against. Without it, you have nothing to measure the effect *against*.
- **Random assignment** means you use chance to decide who goes in the treatment group and who goes in the control group. This balances out other differences (motivation, age, health) between the groups, so the treatment is the main thing that differs — which is what lets you claim cause.
- **Replication** means using enough subjects (and ideally repeating the study) so the result isn't a fluke of a few people. Two people improving could be luck; two hundred is a signal.

Now the mind's role. The **placebo effect** is when people improve simply because they *believe* they're getting a treatment. To control for it, the control group gets a **placebo** — a fake treatment (a sugar pill) that looks identical to the real one — so belief is the same in both groups.
Blinding means the subjects don't know whether they got the real treatment or the placebo. A **double-blind** study goes further: *neither* the subjects *nor* the people measuring the response know who got what, so no one's expectations can tilt the results.
Note the difference from Section 2.2. *Random selection* (sampling) is about who gets *into* the study — it makes the sample fair. *Random assignment* is about which group subjects land in *once they're in* — it makes the comparison fair. Different jobs, both powered by chance.

▪ WATCH IT WORK

Question: Design a fair experiment to test whether a new vitamin boosts energy in 200 volunteers. What are the parts?

Step 1 — Form a control group. Randomly split the 200 volunteers into two groups of 100. One group gets the real vitamin (treatment group); the other gets a placebo — an identical-looking sugar pill (control group).

Step 2 — Use random assignment. Decide who goes in which group *by chance* (e.g., number them 1–200 and randomly draw 100 for treatment). This balances out other differences so the groups start out comparable.

Step 3 — Make it double-blind. Neither the volunteers nor the staff measuring energy levels know who got the real vitamin. This blocks both the placebo effect and any unconscious bias in how energy is rated.

Step 4 — Build in replication. With 100 people per group (not 5), a real difference can show through the noise, and the result is far less likely to be a fluke.

Step 5 — Compare. After a month, compare average energy in the treatment group to the control group. If the treatment group is meaningfully higher, you have evidence the *vitamin* — not belief, not chance, not group differences — caused it.

▪ YOUR TURN

1. Why does an experiment need a *control group*? What does it give you?
2. What is the job of *random assignment* in an experiment?
3. A study gives the control group a sugar pill that looks just like the real medicine. What is this called, and what effect is it controlling for?
4. In a *double-blind* study, who is kept "in the dark" about who got the real treatment?

Work space:

▪ CHECK YOUR VOICE

Five new terms in one section — control, random assignment, replication, placebo, blinding — and if your brain feels full, that's not a flaw, that's a full brain doing its job. You don't need to recite definitions cold. You need to recognize them in a story, and you just did, in the vitamin example. Reread that worked example once and the words attach to a picture. "I can design a fair test" is a sentence you've now earned.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 2 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 2 — Your Turn Answers

Math Skills: (1) $\frac{240}{800} = 0.30 = 30\%$. (2) Number the students 1–25, then draw one number by chance (a computer's random pick, or 25 slips in a hat) — each student has a $\frac{1}{25}$ chance. (3) $1{,}500 - 1{,}200 = 300$ did not respond; $\frac{300}{1500} = 0.20 = 20\%$.

Section 2.1: (1) Voluntary-response bias; skewed toward extreme opinions (the result looks more polarized than the typical visitor). (2) Nonresponse bias; skewed toward the views of daily riders, overstating satisfaction with the service. (3) Undercoverage (cell-phone-only people have no chance of being reached); results miss them entirely. (4) Convenience sampling; the three nearby friends aren't a fair picture of the whole class.

Section 2.2: (1) A sample in which every individual has an equal chance of being chosen (and every possible group of that size is equally likely). (2) Stratified sampling (random samples taken within each grade). (3) Systematic sampling (every 100th item after a random start). (4) False — the first 40 to walk in aren't chosen by equal chance (early arrivers differ from latecomers); it's a convenience sample, not random.

Section 2.3: (1) Observational study (the researcher only records, changing nothing). (2) Explanatory = amount of sleep; response = reaction time. (3) Age (older children have both bigger feet and stronger reading) — a lurking variable, not a cause. (4) False — correlation alone doesn't prove causation; a

confounding variable or coincidence can create the same pattern. Only a controlled experiment can establish cause.

Section 2.4: (1) It gives a baseline to compare the treatment group against, so you can tell the treatment's effect from everything else. (2) It uses chance to balance other differences between the groups, so the treatment is the main thing that differs — which is what lets you claim cause. (3) A placebo; it controls for the placebo effect (improvement from merely *believing* you're being treated). (4) Both the subjects *and* the people measuring the response.

Part II

Describing Data

Chapter 3

Show Me a Picture

Turning Columns of Numbers Into Graphs You Can Actually Read

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll glance at a chart in the news, on a slide, or in a report and think two things at once: "I can read exactly what this is saying" — and "wait, is this graph trying to trick me?"

The Story. A long column of numbers makes almost everyone's eyes slide off the page. A picture of those same numbers makes the pattern jump out instantly. Graphing isn't decoration — it's the fastest way the human brain has to *see* what the data is doing.

The Skills You'll Have:

- Build a frequency table and turn counts into relative frequencies (percents)
- Pick the right picture: bar chart and pie chart for categories; dotplot, stem-and-leaf, and histogram for numbers
- Describe any distribution by its *shape*, *center*, and *spread* — and say which way it's skewed without second-guessing
- Catch a misleading graph in the wild and explain exactly how it cheats

The Confidence Moment. When someone shows you a histogram with a long tail stretching to the right, you'll say "that's skewed right" — calmly, correctly, the first time — because you'll know the trick: *the skew is the direction the tail points.*

The Bridge. A picture shows you the shape. Chapter 4 puts numbers on it: *where is the center* (mean, median) and *how spread out is it* (range, standard deviation). The graph tells the story; the next chapter measures it.

MATH SKILLS YOU'LL NEED

This chapter is about turning numbers into pictures, so the math is mostly counting, dividing, and reading a scale. Let's warm up the three moves you'll lean on.

Skill 1 — Relative frequency: count → fraction → percent. A *count* is how many. A *relative frequency* is that count divided by the total — a fraction, which you can also write as a percent. If 12 of 40 people chose tea:

$$\frac{12}{40} = 0.30 = 30\%$$

The relative frequency tells you the *share*, not just the raw number — and shares are what you compare when groups are different sizes.

Skill 2 — Reading a number line and its scale. A number line (or a graph's axis) is a ruler. Before you read a single bar or dot, find the scale: how much is one tick worth? If the marks read $0, 5, 10, 15$, each tick is 5 units, so a bar reaching halfway between 10 and 15 sits at about 12.5. Always read the scale first.

Skill 3 — Grouping numbers into equal-width bins. To picture a pile of numbers, we sort them into intervals ("bins") that are all the *same width* and don't overlap. For ages 0–39 you might use bins of width 10: $[0, 10)$, $[10, 20)$, $[20, 30)$, $[30, 40)$. The bracket "[" means included, the parenthesis ")" means not included, so 20 falls in $[20, 30)$, not $[10, 20)$. Equal width is what keeps the picture honest.

Practice (answers at the end of the chapter):

1. In a class of 25 students, 10 ride the bus. Write that as a relative frequency and a percent.
2. A scale is marked $0, 20, 40, 60$. A bar reaches the mark halfway between 40 and 60. What value is that?
3. Using bins of width 10 starting at 0, which bin does the value 30 fall into: $[20, 30)$ or $[30, 40)$?

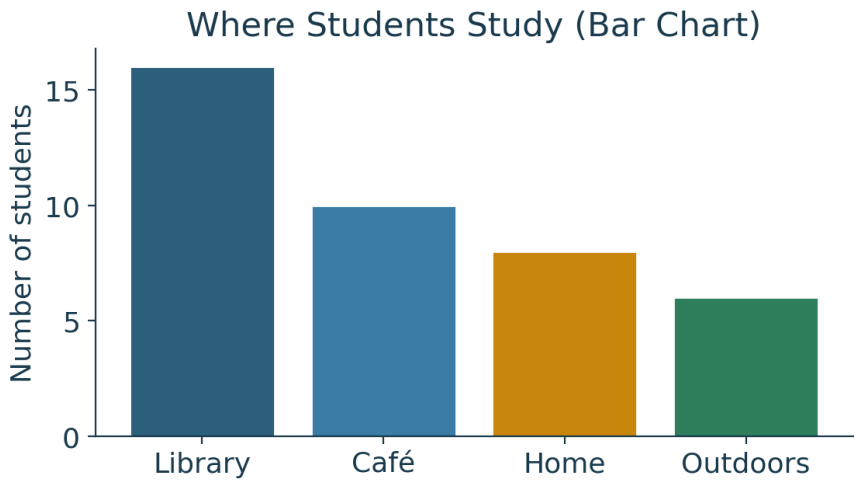


Figure 3.1. A bar chart shows the count for each category.

3.1 Picturing Categories

▪ READ THIS FIRST

Imagine you asked 30 friends one question: "What's your go-to drink at a coffee shop?" You write down 30 answers — coffee, tea, coffee, hot chocolate, coffee, tea, and so on down the page. Staring at that list of 30 words tells you almost nothing. But if you sort the answers into piles — all the coffees here, all the teas there — and then look at how *tall* each pile is, the story appears in two seconds: "coffee wins, tea is second, hot chocolate is a small group."

That sorting-into-piles instinct is the entire idea behind graphing categories.

▪ LET'S TALK ABOUT IT

Think of a question with word-answers, not number-answers — favorite season, phone brand, how you get to school. If you asked 20 people, what "piles" would you expect? Which pile would be tallest? You're already predicting a bar chart in your head.

▪ NOW WE NAME IT

When data is *categorical* (labels, not amounts — remember Chapter 1), we summarize it by counting.

A **frequency table** lists each category and its **frequency** — the count of how many cases fall in it. The **relative frequency** is that count divided by the total, written as a proportion or percent — the category's *share* of the whole.

Two pictures show categorical data:

- A **bar chart** draws one bar per category; the bar's *height* is the frequency (or relative frequency). Bars sit apart with gaps, because the categories are separate groups, not a continuous scale.
- A **pie chart** draws one slice per category; each slice's *share of the circle* is its relative frequency. A pie chart only makes sense when the categories are *parts of one whole* that add to 100%.

Rule of thumb: use a bar chart to *compare* categories (it's easier to judge heights than slice angles); use a pie chart only to show *parts of a whole*, and only with a few categories.

▪ WATCH IT WORK

Question: Here are 20 students' favorite study spot: Library, Home, Cafe, Home, Library, Home, Cafe, Library, Home, Library, Home, Cafe, Home, Library, Home, Cafe, Home, Library, Home, Cafe. Build a frequency table with relative frequencies, and describe the bar chart.

Step 1 — Count each category. Tally as you go: Home $= 9$, Library $= 6$, Cafe $= 5$. Check the total: $9 + 6 + 5 = 20$. ✓

Step 2 — Compute relative frequencies. Divide each count by 20:

Study spot	Frequency	Relative frequency
Home	9	$9/20 = 0.45 = 45\%$
Library	6	$6/20 = 0.30 = 30\%$
Cafe	5	$5/20 = 0.25 = 25\%$
Total	20	$1.00 = 100\%$

Step 3 — Picture it as a bar chart. Three bars, one per spot, with gaps between them. The Home bar is tallest (height 9), Library next (6), Cafe shortest (5). The vertical axis is labeled "Number of students" and runs $0, 2, 4, 6, 8, 10$.

Step 4 — Read the story. "Home is the most popular study spot — almost half the class (45%) — with the library and cafe splitting the rest fairly evenly." That's the whole point of the picture: one sentence, instantly.

▪ YOUR TURN

A survey of 40 commuters asked how they get to work: Car (22), Bus (10), Bike (5), Walk (3).

1. Make a frequency table and add a relative-frequency (percent) column. Check that the percents add to 100%.
2. Which mode of transport has the tallest bar in a bar chart?
3. Would a pie chart be appropriate here? Why or why not?
4. Write one sentence describing the data.

Work space:

▪ CHECK YOUR VOICE

If counting and dividing feels almost too plain to be "real statistics" — pause on that feeling. A frequency table *is* the foundation every fancier graph is built on. You're not skipping the hard part; you're standing on the part everything else needs. Say it: "Simple and correct beats fancy and confused."

3.2 Picturing Numbers

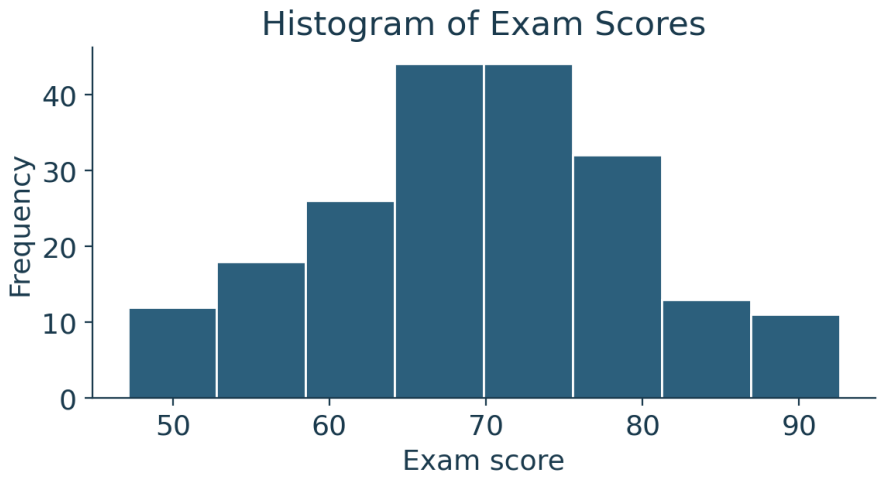


Figure 3.2. A histogram groups numerical data into equal-width intervals.

▪ READ THIS FIRST

Now imagine the question has *number* answers instead of word answers: "How many hours did you sleep last night?" You collect 15 numbers: $6, 7, 8, 7, 5, 9, 7, 6, 8, 7, 6, 10, 7, 8, 6$.

You can't make "piles" the same way — there are too many possible numbers and they're on a real scale, where 5 is genuinely less than 9. So we picture numbers differently: we line them up *along their own scale* and see where they cluster.

Same goal as before — see the pattern fast — but now the picture respects that these are amounts, not labels.

▪ LET'S TALK ABOUT IT

Grab a number you could collect from 15 people — ages, shoe sizes, minutes on the bus. Picture laying those numbers out on a ruler from smallest to largest. Where do you think most of them would bunch up? That bunch is what every graph in this section is trying to reveal.

▪ NOW WE NAME IT

For *quantitative* data we have three go-to pictures, all built on the same idea — place values along a number scale:

- A **dotplot** puts the number line along the bottom and stacks one dot above its value for each case. Great for small data sets; you can still see every individual value.
- A **stem-and-leaf plot** splits each number into a *stem* (the leading digit) and a *leaf* (the last digit). All leaves for a stem sit in one row. It groups the data *and* keeps every original number readable.
- A **histogram** groups the numbers into equal-width *bins* (Math Skill 3) and draws a bar for each bin, where the bar's height is how many values fall in that bin. Bars touch (no gaps), because the scale underneath is continuous. This is the workhorse for larger data sets.

Bars touch or don't? Bar chart (categories) = gaps. Histogram (numbers) = no gaps. The gap is the visual signal of "separate labels vs. a continuous scale."
Choosing bins: too few bins hides the shape (everything in two giant bars); too many bins makes it jagged and noisy (lots of bars of height 1). Aim for roughly 5 to 12 equal-width bins and let the shape show.

▪ WATCH IT WORK

Question: Build a histogram from these 16 quiz scores (out of 20): $11, 14, 15, 15, 16, 12, 18, 15, 13, 17, 14, 16, 19, 15, 14, 16$.

Step 1 — Find the range. Smallest $= 11$, largest $= 19$. So everything lives between 11 and 19.

Step 2 — Choose equal-width bins. Width 2 works nicely: $[10, 12)$, $[12, 14)$, $[14, 16)$, $[16, 18)$, $[18, 20)$. Each value goes in exactly one bin (left edge included, right edge not).

Step 3 — Tally each bin.

Bin (score)	Tally	Frequency
$[10, 12)$	11	1
$[12, 14)$	12, 13	2
$[14, 16)$	14, 15, 15, 14, 15, 14, 15	7
$[16, 18)$	16, 17, 16, 16	4
$[18, 20)$	18, 19	2
Total		16

Step 4 — Draw it. Five touching bars along a score axis. Heights: $1, 2, 7, 4, 2$. The $[14, 16)$ bar towers over the rest.

Step 5 — Read the story. "Most quiz scores cluster in the 14–15 range, with a few stragglers low and a few high scorers near 19." The tallest bar shows where the data *piles up* — that's the center of the action, which we'll measure precisely in Chapter 4.

▪ YOUR TURN

Here are 12 ages of people at a coffee shop: $19, 21, 20, 22, 35, 20, 23, 21, 24, 20, 22, 21$.

1. Make a dotplot in words: which age has the tallest stack of dots?
2. Build a stem-and-leaf plot using the tens digit as the stem (stems 1, 2, 3).
3. Using bins of width 5 starting at 15 ($[15, 20)$, $[20, 25)$, $[25, 30)$, $[30, 35)$, $[35, 40)$), give the frequency in each bin.
4. Which single age looks far from the others?

Work space:

▪ CHECK YOUR VOICE

If choosing bins felt fuzzy — like there's a "right answer" you might miss — here's the relief: there isn't one perfect bin width. Reasonable people pick slightly different bins and tell the same story. You're not solving for a hidden correct number; you're choosing a lens that lets the shape show. That freedom is allowed.

3.3 Reading a Distribution's Shape

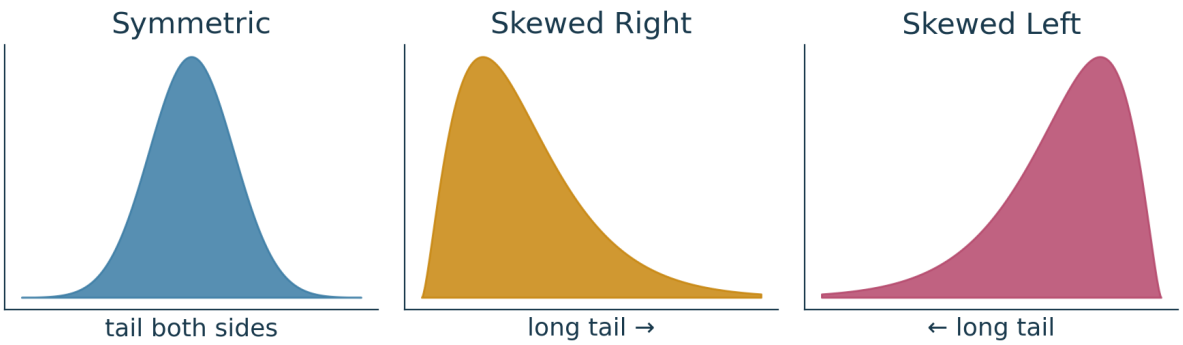


Figure 3.3. Three common shapes. Skew is named for the direction the long tail points.

▪ READ THIS FIRST

Picture the histogram you just built sitting on the table. Now squint at its silhouette — just the outline of the bars. Is it a hill in the middle with sides falling away evenly? A hill shoved to one side with a long ramp stretching off the other way? Two hills? A flat plateau?

That silhouette is the **distribution**'s *shape*, and describing it is one of the most valuable things you can do with a graph — before computing a single average. The shape tells you what kind of data you're dealing with.

▪ LET'S TALK ABOUT IT

Think about incomes in a town, or the number of likes on posts. A few people earn enormous amounts; most earn modest amounts. If you pictured that as a hill, which side would have the long stretched-out tail — the low side or the high side? Hold that mental picture; it's the most-confused idea in this chapter, and you're about to nail it.

▪ NOW WE NAME IT

We describe a distribution with three things, in this order:

- **Shape** — the overall silhouette (covered here).
- **Center** — where the typical value sits (mean, median — Chapter 4).
- **Spread** — how stretched out the values are (range, standard deviation — Chapter 4).
- And always: any **outliers** — values sitting far away from the rest.

Common shapes:

- **Symmetric**: the left and right halves roughly mirror each other (a balanced hill, like a bell).
- **Skewed right**: a long tail stretches to the *right* (toward high values); most data bunches on the left.
- **Skewed left**: a long tail stretches to the *left* (toward low values); most data bunches on the right.
- **Uniform**: bars are all about the same height (a flat plateau — no clear peak).
- **Bimodal**: two distinct peaks (often a sign of two groups mixed together).

The memory hook that ends the confusion: *the skew direction is the direction the long TAIL points.* Right tail → skewed right. Left tail → skewed left. Ignore where the tall bars are; follow the tail.

One more consequence worth knowing now: in a right-skewed distribution, those few large values *pull the mean upward*, so the **mean lands above the median**. In a left-skewed distribution it flips — the mean is pulled *below* the median. (We'll prove this with numbers in Chapter 4; for now, just file it: the mean chases the tail.)

▪ WATCH IT WORK

Question: A histogram of 100 households' incomes (in thousands) looks like this — describe its shape, and say where the mean sits relative to the median.

Income bin ($000s)	Frequency
$[0, 40)$	38
$[40, 80)$	31
$[80, 120)$	18
$[120, 160)$	8
$[160, 200)$	3
$[200, 240)$	2

Step 1 — Sketch the silhouette. The tall bars are on the left (38, 31), and heights shrink steadily as income rises: $18, 8, 3, 2$. The bars trail off into a long, low tail on the *right*.

Step 2 — Follow the tail. The long tail points to the right (toward high incomes). By the memory hook, this is **skewed right**.

Step 3 — Place the mean vs. the median. The handful of high-income households (in the \$160k–\$240k bins) drag the average upward. So the **mean income is higher than the median income** — the mean chased the tail to the right.

Step 4 — Say it in plain words. "Household income is right-skewed: most households earn under \$80k, but a small number of very high earners stretch the upper tail, pulling the average above the typical (middle) household." That single sentence is exactly why news reports usually quote the *median* income, not the mean — the median isn't yanked around by the tail.

▪ YOUR TURN

For each described distribution, name the shape (symmetric, skewed right, skewed left, uniform, or bimodal):

1. Exam scores where most students did very well, with a thin tail of low scores trailing toward the left.
2. Heights of adults, forming a balanced bell with the peak in the middle.
3. Rolls of a fair six-sided die recorded 600 times — each face roughly equally common.
4. Ages at a community event that drew both young children and grandparents but few people in between (two peaks).
5. For the exam scores in (1), is the mean above or below the median?

Work space:

▪ CHECK YOUR VOICE

If you've ever mixed up "skewed right" and "skewed left" — you are in the overwhelming majority, including plenty of people who've taken stats before. The fix isn't memorizing harder; it's the tail rule: *point at the long thin tail, and read off its direction.* You now have a method, not just a memory. That's the difference that holds up under exam pressure.

3.4 When Graphs Lie

▪ READ THIS FIRST

You've probably felt it: a chart on a slide makes one bar look *wildly* bigger than another, you lean in, and the actual numbers are 52 versus 49. The picture screamed "huge difference!" while the data whispered "barely any difference."

The graph didn't show false numbers. It showed true numbers in a *distorted frame*. That's the most common way graphs mislead — and once you know the handful of tricks, you'll catch them everywhere, automatically.

▪ LET'S TALK ABOUT IT

Next time you see a chart trying to convince you of something — an ad, a political post, a company's earnings slide — ask one question: "If I covered the picture and just read the numbers, would I still be impressed?" Often the answer is no, and that gap is where the manipulation lives.

▪ NOW WE NAME IT

A graph is *misleading* when its visual impression doesn't match the underlying numbers. The usual suspects:

- **Truncated axis** (the big one): the vertical axis doesn't start at zero, so a tiny real difference looks enormous. A bar from 49 to 52 can be drawn to look three times taller than its neighbor if the axis starts at 48.
- **Inconsistent or missing scale**: tick marks are uneven, or there's no scale at all, so you can't tell what the heights actually mean.
- **Distorted area or 3-D**: using a wider/taller *picture* (or a 3-D effect) to represent a value doubles both width and height, so the area grows like the *square* — a value that's twice as big looks four times as big.
- **Cherry-picked bins or time windows**: choosing bin widths or a start/end date that flatter one conclusion and hide the rest of the story.

The defense is always the same: **read the axis before you read the bars.** Check where the vertical axis starts (is it zero?), whether the ticks are evenly spaced, and whether the time window or bins were chosen to make a point.

▪ WATCH IT WORK

Question: Two products sold this quarter: Product A sold 1,020 units, Product B sold 1,000 units. A slide draws a bar chart where A's bar looks twice as tall as B's. How is the chart lying, and how would you fix it?

Step 1 — Read the actual numbers. $1{,}020$ vs. $1{,}000$. The real difference is 20 units — about a 2% edge for A. Tiny.

Step 2 — Look at the axis. The vertical axis starts at $1{,}000$, not 0, and runs $1{,}000, 1{,}005, 1{,}010, 1{,}015, 1{,}020$. So B's bar has *height zero* on this chart and A's bar is the full height — making A look infinitely (or at least doubly) bigger.

Step 3 — Name the trick. This is a *truncated axis*. By cutting off the bottom $1{,}000$ units, the chart converts a 2% difference into a dramatic visual gap.

Step 4 — Fix it. Redraw with the vertical axis starting at **0**, running $0, 250, 500, 750, 1000$. Now both bars are nearly the same towering height, with A's just a hair taller — which is the honest picture: the products sold almost the same amount.

The takeaway: the data never changed. Only the *frame* did. Always ask where the axis starts.

▪ YOUR TURN

For each scenario, name the misleading technique and say what an honest version would do:

1. A chart shows website visits "doubling" — but the y-axis runs from 9,800 to 10,200.
2. An infographic shows "twice the savings" using a money-bag icon drawn twice as tall *and* twice as wide as the other.
3. A company posts a line graph of profits going up sharply, starting the timeline right after its worst-ever month.
4. A bar chart of survey results has no numbers on the vertical axis at all.

Work space:

▪ CHECK YOUR VOICE

Notice what just happened: you didn't need advanced math to catch a lying graph — you needed to *slow down and read the axis*. That skepticism is not negativity; it's literacy. The voice that says "wait, that looks off" is your statistical instinct waking up. Trust it. "I can be convinced by data — but only honest data."

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 3 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 3 — Your Turn Answers

Math Skills: (1) $\frac{10}{25} = 0.40 = 40\%$. (2) Halfway between 40 and 60 is 50. (3) 30 falls in $[30, 40)$ (the left edge is included, so 30 belongs to the bin that *starts* at 30).

Section 3.1: (1) Car $22/40 = 55\%$, Bus $10/40 = 25\%$, Bike $5/40 = 12.5\%$, Walk $3/40 = 7.5\%$; they add to 100%. (2) Car (frequency 22) has the tallest bar. (3) Yes — the four modes are parts of one whole (all 40 commuters) and there are only a few categories, so a pie chart works; a bar chart is still easier to compare. (4) Example: "Most commuters drive (55%), with the bus a distant second and biking and walking rare."

Section 3.2: (1) Age 20 and age 21 each appear three times — they tie for the tallest stack (3 dots each). (2) Stem-and-leaf (stem = tens digit): $1 \mid 9 \quad 2 \mid 0\,0\,0\,1\,1\,1\,2\,2\,3\,4 \quad 3 \mid 5$. (3) $[15, 20)$: 1; $[20, 25)$: 10; $[25, 30)$: 0; $[30, 35)$: 0; $[35, 40)$: 1. (4) The age 35 sits far from the cluster of early-twenties ages — a likely outlier.

Section 3.3: (1) Skewed left (the long tail points left, toward low scores). (2) Symmetric. (3) Uniform. (4) Bimodal (two peaks). (5) For the left-skewed exam scores, the mean is *below* the median — the low tail pulls the mean down.

Section 3.4: (1) Truncated (non-zero) y-axis; an honest version starts the axis at 0, showing the visits barely changed. (2) Distorted area — doubling both width and height makes the icon look about four times bigger; an honest version scales only one dimension (or uses bars). (3) Cherry-picked time window (start date); an honest version shows the full timeline including the bad month. (4) Missing scale; an honest version labels the vertical axis with numbers so the bar heights mean something.

Chapter 4

The Typical and the Spread

Measuring the Center of Data — and How Far It Wanders

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. In Chapter 3 you learned to *look* at a distribution and describe its shape. Now we put numbers on two of the things your eye already noticed: *where is the middle*, and *how spread out is the pile*. By the end you'll calculate a mean, a median, and a standard deviation — and, more importantly, say in plain words what each one means.

The Story. Every time you say "a normal grocery bill is about \$80" or "my commute is usually 20 minutes, give or take 5," you are doing this chapter in your head. "About \$80" is a center. "Give or take 5" is a spread. Statistics just gives those two instincts exact names and formulas.

The Skills You'll Have:

- Find the *mean*, *median*, and *mode*, and pick the right one for the situation
- Explain why an outlier drags the mean but leaves the median calm
- Measure spread with the *range*, the *variance*, and the *standard deviation*
- Use a *z-score* to say how far above or below average a single value sits
- Build a *five-number summary* and use the $1.5\times$IQR rule to flag outliers

The Confidence Moment. When someone says "the standard deviation is about 5," you'll calmly translate: "a typical value sits about 5 away from the average." No panic, no formula-fright — just a sentence you understand.

The Bridge. Once you can describe *one* data set by its center and spread, Part III asks the next question: *how do random outcomes behave?* The mean and standard deviation you learn here become the two dials on every probability distribution in the chapters ahead.

MATH SKILLS YOU'LL NEED

This chapter finally asks you to *calculate*, not just read. Good news: it's all arithmetic you already own. Let's warm up the four moves we'll lean on.

Skill 1 — Order of operations (PEMDAS). Do what's inside **P**arentheses first, then **E**xponents (squaring), then **M**ultiply/**D**ivide, then **A**dd/**S**ubtract. So in $(7-4)^2$ you subtract *first* ($7-4=3$), then square ($3^2=9$) — never the other way.

Skill 2 — Squaring and square roots. "Squaring" means multiplying a number by itself: $4^2 = 4\times 4 = 16$. A square root undoes that: $\sqrt{16}=4$, because $4\times 4=16$. Squaring always gives a number that is zero or positive — even $(-3)^2=9$. (That fact is the secret reason the standard-deviation formula works.)

Skill 3 — Adding a list, and the $\sum$ symbol. When we add up a whole list of numbers, we write the Greek letter sigma, $\sum$, which just means "add these up." So $\sum x$ is shorthand for "add all the x values." For the list $4, 8, 6$, we have $\sum x = 4+8+6=18$. That's it — a fancy symbol for plain addition.

Skill 4 — Dividing. Splitting a total into equal parts. To average, we add the list and then divide by how many numbers there are. $\dfrac{18}{3}=6$.

Practice (answers at the end of the chapter):

1. Compute $(10-6)^2$ using the correct order of operations.
2. Find $\sqrt{49}$.
3. For the list $5, 9, 7, 3$, find $\sum x$ and then divide by 4.

4.1 Finding the Center

▪ READ THIS FIRST

Someone asks, "How long is your morning commute?" You don't recite all twenty commutes from last month. You give them *one number* — "about 25 minutes" — that stands in for the whole pile.

That single stand-in number is what statisticians call a *measure of center*. You've been computing them your whole life: a "typical" grade, a "usual" grocery bill, a "normal" bedtime. There are three official ways to find the center, and you already use all three without naming them.

▪ LET'S TALK ABOUT IT

Think of the prices of five things you bought this week. If a friend asked "what's a typical price?", would you add them all and divide — or would you just point to the one in the middle? Notice that your gut already picks a method depending on whether one giant price would "mess up" the answer.

▪ NOW WE NAME IT

There are three measures of center.

The **mean** is the ordinary average: add all the values and divide by how many there are. In symbols, $\bar{x} = \dfrac{\sum x}{n}$, where $\bar{x}$ ("x-bar") is the sample mean and n is the number of values.

The **median** is the middle value when the data are lined up in order. With an odd count, it's the single middle number; with an even count, it's the average of the two middle numbers. Half the data sit below it, half above.

The **mode** is the value that appears most often. A data set can have one mode, several, or none.

Here's the idea that matters most: the mean uses the actual *size* of every value, so one extreme value (an **outlier**) can yank it. The median only cares about *position* in the lineup, so it shrugs off extremes. Tie this back to Chapter 3: in a *right-skewed* distribution, the few large values in the long right tail drag the **mean above the median**. The mean chases the tail; the median holds its ground.

▪ WATCH IT WORK

Question: Six friends report their monthly streaming spending (in dollars): $10, 12, 15, 12, 11, 50$. Find the mean, median, and mode, and say which best describes a "typical" friend.

Step 1 — Mean. Add them: $10 + 12 + 15 + 12 + 11 + 50 = 110$. There are $n = 6$ values, so $\bar{x} = \dfrac{110}{6} \approx 18.3$ dollars.

Step 2 — Median. Line them up: $10, 11, 12, 12, 15, 50$. With six values (even), average the two middle ones (the 3rd and 4th): $\dfrac{12 + 12}{2} = 12$ dollars.

Step 3 — Mode. The value 12 appears twice; everything else once. The mode is 12 dollars.

Step 4 — Interpret. That one big \$50 spender pulled the *mean* up to \$18.30 — higher than what anybody typical actually pays. The *median* of \$12 describes the real "typical" friend far better. This little data set is right-skewed (one long tail toward \$50), and sure enough the mean (18.3) landed *above* the median (12) — exactly the Chapter 3 rule.

▪ YOUR TURN

A small café records the ages of seven customers one afternoon: $19, 21, 20, 22, 19, 24, 70$.

1. Find the mean age.
2. Find the median age.
3. Find the mode.
4. Which measure best describes the "typical" customer, and why?

Work space:

▪ CHECK YOUR VOICE

If the word "median" ever made you freeze, notice what you just did: you lined numbers up and pointed at the middle. That's all it is. The inner critic that says "real math people use formulas" is wrong — choosing the median over the mean *because* of an outlier is sophisticated statistical thinking. You did it with your gut before you had the words.

4.2 Measuring the Spread

▪ READ THIS FIRST

Two coffee shops both have an average wait of 5 minutes. At the first, every customer waits almost exactly 5 minutes. At the second, some wait 1 minute and some wait 9. Same center — completely different experience.

The center alone never tells the whole story. You also need to know *how far the values wander* from that center. That "wandering" is called *spread*, and measuring it is the other half of describing data.

▪ LET'S TALK ABOUT IT

Think about your last five paychecks, or your last five test scores. Were they all clustered close together, or all over the place? If you had to put a single number on "how much they bounce around," what would you even measure — the biggest gap? the typical gap? Hold that thought; you're about to meet both.

▪ NOW WE NAME IT

The simplest measure of spread is the **range**: the largest value minus the smallest. It's quick, but it only uses two numbers and one outlier can blow it up.

The serious measure of spread is the **standard deviation**. In plain words, the standard deviation is **the typical distance of a value from the mean**. If the mean is the meeting point, the standard deviation says how far, on average, everybody is standing from it.

To get there we first compute the **variance**, which is the average of the *squared* distances from the mean. We square the distances for two reasons: it makes the below-average gaps (which are negative) stop canceling the above-average gaps, and it keeps the math well-behaved. The standard deviation is just the square root of the variance, which brings us back to the original units (dollars, minutes, points).

For a *sample*, the standard deviation is

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

Read it inside-out: $(x - \bar{x})$ is each value's distance from the mean (a **deviation**); we square each one, add them up with $\sum$, divide by $n - 1$, and take the square root. We divide by $n - 1$ (not n) whenever the data are a *sample* — it's a small correction that keeps us from underestimating the true spread. Use $n - 1$ for samples throughout this book.

▪ WATCH IT WORK

Question: Five days of daily high temperatures (in °F) are recorded: $4, 8, 6, 5, 7$. Find the range and the sample standard deviation.

Step 1 — Range. Largest minus smallest: $8 - 4 = 4$ degrees.

Step 2 — Find the mean. $\bar{x} = \dfrac{4+8+6+5+7}{5} = \dfrac{30}{5} = 6$ degrees.

Step 3 — Deviations from the mean $(x - \bar{x})$:

Value x	Deviation $(x - \bar{x})$	Squared $(x - \bar{x})^2$
4	$4 - 6 = -2$	4
8	$8 - 6 = 2$	4
6	$6 - 6 = 0$	0
5	$5 - 6 = -1$	1
7	$7 - 6 = 1$	1

Step 4 — Sum the squared deviations. $\sum(x - \bar{x})^2 = 4 + 4 + 0 + 1 + 1 = 10$.

Step 5 — Divide by $n - 1$. Here $n = 5$, so $n - 1 = 4$: $\dfrac{10}{4} = 2.5$. (That's the variance.)

Step 6 — Take the square root. $s = \sqrt{2.5} \approx 1.58$ degrees.

Interpret: a typical day's high temperature sits about $1.58°$ away from the average of $6°$. Small number $\rightarrow$ the temperatures are clustered tightly.

▪ YOUR TURN

A student's last five quiz scores (out of 10) are: $6, 9, 7, 8, 10$.

1. Find the range.
2. Find the mean.
3. Find the sample standard deviation s (show the deviations, square them, sum, divide by $n-1$, then take the square root).
4. In one sentence, say what your s means in plain words.

Work space:

▪ CHECK YOUR VOICE

Standard deviation is the spot where almost everyone's inner critic shows up shouting "you'll never remember that formula." So let's name it and shrink it. You don't memorize the formula — you follow a recipe: distance, square, add, divide, root. And the whole point of the scary-looking s is one calm sentence: "the typical distance from the average." If you understood the two-coffee-shops story, you already understand standard deviation. The formula is just the bookkeeping.

4.3 Where Do I Stand? Z-Scores and Percentiles

▪ READ THIS FIRST

You scored an 85 on a test. Is that good? You genuinely cannot tell — not yet. If the class average was 70, you crushed it. If the average was 95, you struggled. A raw number means nothing until you know *where it sits relative to the group*.

There's a clean way to answer "where do I stand?" in a single number, and it uses exactly the two things you just learned: the mean and the standard deviation.

▪ LET'S TALK ABOUT IT

Think of a time you got a score back — a grade, a 5K time, a credit score — and your very first question was "wait, is that good?" What did you compare it to? Notice that you instinctively wanted to know the average, and how far you were from it. That instinct *is* the z-score.

▪ NOW WE NAME IT

A **z-score** measures how many standard deviations a value sits above or below the mean:

$$z = \frac{x - \bar{x}}{s}.$$

The top, $x - \bar{x}$, is your distance from the mean. Dividing by s rescales that distance into "number of standard deviations." A **positive** z-score means you're above average; a **negative** one means below; a z-score of 0 means you're exactly average. A z-score of $+2$ means "two standard deviations above the mean" — noticeably high.

A close cousin is the **percentile**: the percent of the data that falls at or below your value. If your score is at the 90th percentile, you scored at or above 90% of the group. Percentiles are why a doctor says a child is "in the 60th percentile for height," and why a test report says you're "in the 75th percentile." Both the z-score and the percentile answer the same human question — *where do I stand?* — just in different units.

▪ WATCH IT WORK

Question: On a test, the class mean is $\bar{x} = 70$ and the standard deviation is $s = 8$. You scored 86. How many standard deviations above average is that?

Step 1 — Find your distance from the mean. $x - \bar{x} = 86 - 70 = 16$ points above average.

Step 2 — Divide by the standard deviation. $z = \dfrac{16}{8} = 2$.

Step 3 — Interpret. Your z-score is $+2$: you scored *two full standard deviations above the class average.* That's a strong, well-above-average result.

One more: a classmate scored 66. Then $z = \dfrac{66 - 70}{8} = \dfrac{-4}{8} = -0.5$ — half a standard deviation *below* average. The negative sign instantly tells you "below the mean," and the 0.5 tells you "not by much."

▪ YOUR TURN

A 5K running club has a mean finish time of $\bar{x} = 30$ minutes with a standard deviation of $s = 4$ minutes.

1. Maria finishes in 22 minutes. Find her z-score.
2. Is Maria faster or slower than average? How do you know from the sign?
3. Tom finishes in 34 minutes. Find his z-score.
4. In plain words, what does Tom's z-score say about his time?

Work space:

▪ CHECK YOUR VOICE

A z-score has a fearsome reputation and a friendly job. All it does is answer "how far from average am I, measured in standard-deviation steps?" If you've ever said "I was way above average" or "I was a little below," you've spoken in z-scores — "way" and "a little" were your units. Now you have a number for them. That's not new math; it's a new name for something you already say.

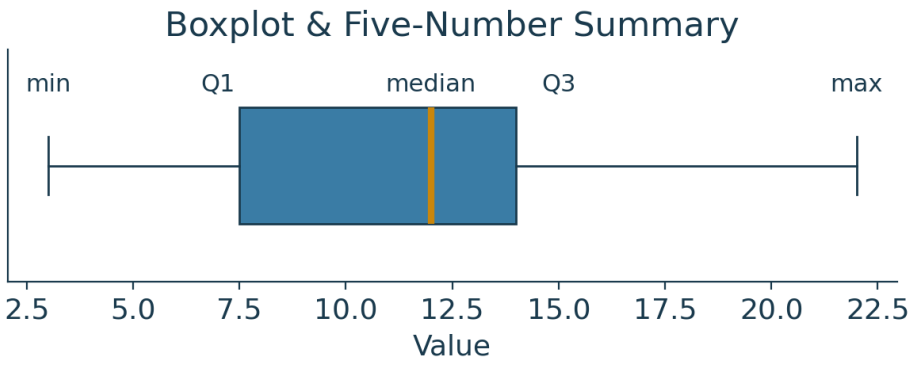


Figure 4.1. A boxplot displays the five-number summary at a glance.

4.4 The Five-Number Summary and the Boxplot

▪ READ THIS FIRST

Imagine lining up all the data, smallest to largest, and then marking five fence-posts: the very start, the value a quarter of the way in, the middle, the value three-quarters of the way in, and the very end. Those five posts tell you almost everything about how the data are spread out — without listing every value.

That set of five landmarks has a name (the five-number summary), and the little picture drawn from it (the boxplot) is one of the most useful graphs in all of statistics.

▪ LET'S TALK ABOUT IT

Think of a class where exam scores ran from a low of 40 to a high of 100. Knowing just the lowest and highest, can you tell whether most people did well? Probably not. What if you also knew the middle score, and the scores a quarter and three-quarters of the way up? Notice how those extra landmarks suddenly let you "see" the whole class.

▪ NOW WE NAME IT

The **five-number summary** is, in order: the **minimum**, the **first quartile** (Q_1), the **median** (Q_2), the **third quartile** (Q_3), and the **maximum**. Q_1 is the median of the lower half of the data (25% sit below it); Q_3 is the median of the upper half (75% sit below it).

From the quartiles we get the **interquartile range**, the spread of the middle half of the data:

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

The IQR is a spread measure that, like the median, ignores extremes — it only describes the middle 50%.

We can also use it to flag outliers with the **1.5 × IQR rule**: a value is an *outlier* if it falls

$$\text{below } Q_1 - 1.5 \cdot \text{IQR} \qquad \text{or} \qquad \text{above } Q_3 + 1.5 \cdot \text{IQR}.$$

Those two cutoffs are sometimes called the lower and upper "fences." Anything outside the fences is unusually far from the pack.

A **boxplot** (or box-and-whisker plot) draws this summary. Picture a number line. A *box* is drawn from Q_1 to Q_3 (so the box spans the IQR), with a line inside it at the median. Then a "whisker" extends out from each end of the box to the smallest and largest values that are still inside the fences. Any point beyond a fence is drawn as a separate dot — a visible outlier. At a glance, a long box or a long whisker means more spread, and a median line sitting off-center hints at skew.

▪ WATCH IT WORK

Question: Seven houses on a block sold for these prices (in thousands of dollars): $3, 7, 8, 12, 13, 15, 22$ (already in order). Find the five-number summary, the IQR, and check whether \$22 (thousand) is an outlier by the $1.5\times$IQR rule.

Step 1 — Min and max. Minimum $= 3$, maximum $= 22$.

Step 2 — Median (Q_2). With $n = 7$ values, the middle is the 4th value: 12.

Step 3 — Q_1. The lower half (the values below the median) is $3, 7, 8$. Its median is the middle value, $Q_1 = 7$.

Step 4 — Q_3. The upper half is $13, 15, 22$. Its median is $Q_3 = 15$.

Five-number summary: $3,\ 7,\ 12,\ 15,\ 22$.

Step 5 — IQR. IQR $= Q_3 - Q_1 = 15 - 7 = 8$.

Step 6 — Outlier check. $1.5 \times$ IQR $= 1.5 \times 8 = 12$. Lower fence: $Q_1 - 12 = 7 - 12 = -5$. Upper fence: $Q_3 + 12 = 15 + 12 = 27$. Since 22 falls *between* -5 and 27, it is **not** an outlier — it's high, but not far enough out to clear the fence.

▪ YOUR TURN

Nine days of daily customer counts at a food truck (in order): $12, 15, 18, 20, 22, 25, 28, 30, 60$.

1. Find the median (Q_2).
2. Find Q_1 and Q_3. (Hint: with $n = 9$, the median is the 5th value; the lower half is the four values below it, the upper half the four above it. For a four-value half, the quartile is the average of its two middle numbers.)
3. Find the IQR.
4. Using the $1.5\times$IQR rule, is the 60-customer day an outlier?

Work space:

▪ CHECK YOUR VOICE

If five numbers and a fence rule felt like a lot, look at what each piece actually *does*: the five-number summary is just "start, quarter, middle, three-quarter, end," and the IQR is the width of the middle. The outlier rule isn't a law from on high — it's a reasonable line in the sand for "unusually far out." You're not memorizing trivia; you're learning to describe a pile of numbers the way a careful person naturally would.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 4 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 4 — Your Turn Answers

Math Skills: (1) $(10-6)^2 = 4^2 = 16$ (subtract inside the parentheses first, then square). (2) $\sqrt{49} = 7$ (because $7 \times 7 = 49$). (3) $\sum x = 5+9+7+3 = 24$, and $\dfrac{24}{4} = 6$.

Section 4.1: Ages $19, 21, 20, 22, 19, 24, 70$. (1) Mean $= \dfrac{19+21+20+22+19+24+70}{7} = \dfrac{195}{7} \approx 27.9$ years. (2) In order: $19, 19, 20, 21, 22, 24, 70$; with $n = 7$ the median is the 4th value, 21 years. (3) Mode $= 19$ (it appears twice). (4) The *median* (21) best describes the typical customer: the single 70-year-old is an outlier that pulls the mean up to 27.9, higher than anyone typical. (The data are right-skewed, so the mean lands above the median — the Chapter 3 rule again.)

Section 4.2: Quiz scores $6, 9, 7, 8, 10$. (1) Range $= 10 - 6 = 4$. (2) Mean $= \dfrac{6+9+7+8+10}{5} = \dfrac{40}{5} = 8$. (3) Deviations and squares: $(6-8) = -2 \to 4$; $(9-8) = 1 \to 1$; $(7-8) = -1 \to 1$; $(8-8) = 0 \to 0$; $(10-8) = 2 \to 4$. Sum of squares $= 4+1+1+0+4 = 10$. Divide by $n-1 = 4$: $\dfrac{10}{4} = 2.5$ (the variance). Square root: $s = \sqrt{2.5} \approx 1.58$. (4) A typical quiz score sits about 1.58 points away from the average of 8.

Section 4.3: Mean 30 min, SD 4 min. (1) Maria: $z = \dfrac{22-30}{4} = \dfrac{-8}{4} = -2$. (2) Faster than average — the *negative* sign means she's below the mean finish time, and a lower time is faster. (3) Tom: $z = \dfrac{34-30}{4} = \dfrac{4}{4} = 1$. (4) Tom's z-score of $+1$ means his time is one standard deviation *above* (slower

than) the average finish time.

Section 4.4: Customer counts $12, 15, 18, 20, 22, 25, 28, 30, 60$. (1) With $n = 9$, the median is the 5th value: $Q_2 = 22$. (2) Lower half (below the median) $= 12, 15, 18, 20$; its quartile is the average of the two middle values, $Q_1 = \dfrac{15+18}{2} = 16.5$. Upper half $= 25, 28, 30, 60$; $Q_3 = \dfrac{28+30}{2} = 29$. (3) IQR $= Q_3 - Q_1 = 29 - 16.5 = 12.5$. (4) $1.5 \times$ IQR $= 1.5 \times 12.5 = 18.75$. Upper fence $= Q_3 + 18.75 = 29 + 18.75 = 47.75$. Since $60 > 47.75$, the 60-customer day **is** an outlier.

Parts I–II Checkpoint — Study Guide & Practice (Chapters 1–4)

This Study Guide pulls together everything from Chapters 1–4 and lets you rehearse before an exam the way the exam will actually feel.

STUDY GUIDE & PRACTICE

You've built a lot in three chapters. This isn't a test — it's a dress rehearsal. We're going to gather the big ideas in one place, check the vocabulary, and then run a set of mixed problems across Chapters 1–3 plus a focused Chapter 4 set, exactly the way a real exam does. When your inner voice says "I forgot all of this," keep going anyway — the answers are right here, fully worked, waiting for you.

The Big Ideas, in One Place

Chapter 1 — The Language of Data.

- **Statistics** is the science of collecting, organizing, analyzing, and interpreting **data** to answer a question. One recorded fact is a *data value*; all of them together are a *data set*.
- A **population** is the *whole* group you want to know about (the whole pot of soup). A **sample** is the *slice* you actually measure (the spoonful).
- A number describing the *population* is a **parameter**; a number calculated from a *sample* is a **statistic**. Memory hook: **p**opulation → **p**arameter, **s**ample → **s**tatistic.
- A **variable** is a characteristic that changes from case to case. **Categorical** (qualitative) variables record a *label* (major, eye color, Yes/No, zip code). **Quantitative** variables record a *measured amount* you can average (age, income, hours of sleep). The test that never fails: "Does it make sense to average it?"
- Quantitative splits into **discrete** (from counting — whole numbers, like siblings) and **continuous** (from measuring — any value in a range, like height).

- A data table is just a sign-up sheet: each **row** is a **case** (individual), each **column** is a **variable**, each **cell** is a **data value**. "n" is the number of cases (rows).

Chapter 2 — Where Data Comes From.

- A beautiful analysis of badly collected data is still wrong. *How* you get the sample decides whether you can trust it.
- **Bias** happens when a sampling method systematically misses or distorts part of the population. Common types: *selection/undercoverage bias* (some groups can't be chosen), *voluntary-response bias* (only people who care strongly reply), *nonresponse bias* (the people you picked don't answer), and *response bias* (the question or setting pushes a certain answer).
- A **simple random sample (SRS)** gives every group of the chosen size an equal chance — the gold standard. Other honest methods: *stratified* (split into groups, sample within each), *cluster* (sample whole groups), *systematic* (every kth case), and *convenience* (whoever's easiest — usually biased).
- An **observational study** watches and records without changing anything. An **experiment** deliberately *applies a treatment* and compares outcomes.
- The **explanatory variable** is the suspected cause; the **response variable** is the outcome you measure.
- A **confounding variable** is a hidden third variable tied to both the explanatory variable and the response, so you can't tell which one caused the change. Observational studies can show *association* but not *causation* — correlation $\neq$ causation.
- Good experiments use *control* (a comparison/placebo group), *randomization* (assign treatments by chance to balance out confounders), and *replication* (enough subjects so results aren't a fluke). Only a well-designed experiment lets you claim cause and effect.

Chapter 3 — Seeing the Data.

- A **frequency** is the count of how often a value occurs. A **relative frequency** is that count divided by the total — a proportion or percent. Relative frequencies of all categories add to 1 (or 100%).
- **Match the graph to the variable type.** *Categorical* data $\rightarrow$ **bar chart** or **pie chart**. *Quantitative* data $\rightarrow$ **dotplot**, **stem-and-leaf plot**, or **histogram**.

- Describe a distribution by **shape**, **center**, and **spread** (and any unusual points, called *outliers*).
- Shapes: *symmetric* (a mirror image, like a bell), *skewed right* (tail stretches toward the high/right side), *skewed left* (tail stretches toward the low/left side). **The skew is named for the direction of the tail, not the hump.**
- Watch for **misleading graphs**: a y-axis that doesn't start at 0 (exaggerates differences), uneven scales or bar widths, cut or stretched axes, and pictures sized by height *and* width so areas mislead.

Chapter 4 — Describing Center and Spread.

- **Center**: the **mean** $\bar{x} = \frac{\sum x}{n}$ (the average — best for symmetric data); the **median** (middle of the ordered list — best when skewed or with outliers); the **mode** (most frequent value).
- **Spread**: the **range** $(\max - \min)$ and, the everyday measure, the **standard deviation** $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ — read it as "the typical distance of a value from the mean." Divide by $n-1$ for a sample.
- A **z-score** $z = \frac{x-\bar{x}}{s}$ says how many standard deviations a value sits above $(+)$ or below $(-)$ the mean.
- The **five-number summary** (min, Q_1, median, Q_3, max) and the **IQR**$= Q_3 - Q_1$ describe spread resistant to outliers; a **boxplot** pictures them. A value is an **outlier** if it falls below $Q_1 - 1.5\,(IQR)$ or above $Q_3 + 1.5\,(IQR)$.
- Connecting to shape (Chapter 3): in a right-skewed distribution the few large values pull the *mean above the median*; in a left-skewed one the mean sits *below* the median.

Vocabulary Check

Can you say each of these in one plain sentence? If one feels fuzzy, that's exactly the one to reread — not a sign you're behind.

- population, sample, parameter, statistic
- categorical vs. quantitative; discrete vs. continuous
- case, variable, data value, n
- bias (selection, voluntary-response, nonresponse, response)

- simple random sample; stratified, cluster, systematic, convenience sampling
- observational study vs. experiment
- explanatory variable, response variable, confounding variable
- control, randomization, replication
- frequency, relative frequency
- bar chart, pie chart, dotplot, stem-and-leaf, histogram
- symmetric, skewed right, skewed left, outlier
- misleading graph
- mean, median, mode; range, standard deviation; z-score
- quartile, IQR, five-number summary, boxplot, $1.5 \times IQR$ outlier rule

Mixed Practice

These twelve problems jump around on purpose — just like an exam. Try each one before peeking. The full worked answers come right after.

1. Classify each variable as categorical or quantitative (and if quantitative, discrete or continuous): (a) favorite pizza topping, (b) number of text messages sent today, (c) height in centimeters, (d) area code of a phone number.

2. A streaming service has $4{,}000{,}000$ subscribers. Analysts study $5{,}000$ of them and find those $5{,}000$ watch 3.2 hours per day on average. Identify the population, the sample, and state whether "3.2 hours" is a parameter or a statistic.

3. A radio host asks listeners to call in and vote on whether the city should build a new stadium. Most callers say "yes." Name the type of bias most likely at work and explain why the result can't be trusted as the city's true opinion.

4. A researcher wants to estimate the average GPA of all $8{,}000$ students at a college. She stands by the library exit one afternoon and asks the first 50 students who walk out. (a) What sampling method is this? (b) Why might it be biased? (c) Briefly describe how she could take a simple random sample instead.

5. For each study, say whether it is an observational study or an experiment, and identify the explanatory and response variables: (a) Researchers track 500 adults for a year and record their coffee intake and their reported stress levels. (b) Researchers randomly assign 200 volunteers to either a new sleep app or no app, then measure hours slept after a month.

6. A town finds that months with higher ice-cream sales also have more drowning incidents. A reporter writes, "Ice cream causes drowning." Explain why this conclusion is wrong, and name a confounding variable that better explains the link.

7. A class of 40 students chose their favorite study spot: Library 16, Café 10, Home 8, Outdoors 6. Build a relative-frequency table (as percents) and confirm the relative frequencies sum correctly.

8. You surveyed people about their blood type (A, B, AB, O). Which graph type is appropriate to display these results — a histogram or a bar chart — and why?

9. A dotplot of exam scores has most scores bunched up high (in the 80s and 90s) with a few low scores trailing down toward 40. Is this distribution skewed left or skewed right? Explain using the tail.

10. A company's bar chart shows "Sales doubled!" Brand X's bar looks twice as tall as Brand Y's, but when you read the y-axis it runs from 90 to 100, and the values are 92 and 96. Explain what makes this graph misleading and what the true comparison is.

11. In a data set of 120 patients with 7 recorded characteristics each, what is n, what does each row represent, and what does the cell in row 5, column "Age" contain?

12. A nutrition company runs an experiment: 300 volunteers are randomly split into a group that takes a new vitamin and a group that takes a look-alike sugar pill, and neither group is told which they got. Identify the use of (a) control, (b) randomization, and (c) replication in this design.

Answers

Worked out in full — read these even for the ones you got right, because the *reasoning* is what the exam rewards.

1. (a) Favorite pizza topping $=$ *categorical* (a label; you can't average it). (b) Number of text messages $=$ *quantitative, discrete* (a count of whole things). (c) Height in centimeters $=$

quantitative, continuous (a measurement that can land anywhere in a range). (d) Area code $=$ *categorical* — it's made of digits, but averaging area codes is meaningless, so it's a label.

2. **Population:** all $4{,}000{,}000$ subscribers. **Sample:** the $5{,}000$ studied. The "3.2 hours" is a *statistic* because it was calculated from the sample we measured, not from the whole population. (Population $\rightarrow$ parameter, sample $\rightarrow$ statistic.)

3. This is *voluntary-response bias*. The sample is made only of people who felt strongly enough to call in, so it over-represents intense opinions and is not a random, representative slice of the city. The true citywide opinion would need a properly drawn sample, not self-selected callers.

4. (a) This is a *convenience sample* (whoever was easiest to reach). (b) It's biased through *undercoverage/selection*: students who don't visit the library that afternoon — including those who never use it — have no chance of being chosen, and library-goers may differ systematically from other students (e.g., study habits, GPA). (c) For an SRS, number all $8{,}000$ students 1 to $8{,}000$, then use a random number generator to pick, say, 50 distinct numbers; survey exactly those students. Every student then has an equal chance of selection.

5. (a) *Observational study* — the researchers only watch and record; no treatment is imposed. Explanatory variable: coffee intake; response variable: stress level. (Because it's observational, it can show association but not cause.) (b) *Experiment* — a treatment (sleep app vs. no app) is deliberately assigned, and at random. Explanatory variable: whether you get the app; response variable: hours slept.

6. This confuses correlation with causation — two things rising together does not mean one causes the other. A *confounding variable* is *hot weather (temperature/season)*: hot months drive both more ice-cream sales *and* more swimming (hence more drownings). Temperature is tied to both, so it, not ice cream, explains the link.

7. Divide each count by 40:

Study spot	Frequency	Relative frequency
Library	16	$16/40 = 0.40 = 40\%$
Café	10	$10/40 = 0.25 = 25\%$
Home	8	$8/40 = 0.20 = 20\%$
Outdoors	6	$6/40 = 0.15 = 15\%$
Total	40	$1.00 = 100\%$

Check: $40\% + 25\% + 20\% + 15\% = 100\%$. The relative frequencies add to 1, as they always must.

8. A *bar chart*. Blood type is a *categorical* variable (labels with no numeric order), and bar charts (or pie charts) display categories. A histogram is for *quantitative* data grouped into numeric intervals, so it doesn't fit here.

9. *Skewed left.* The bulk of the data sits high (80s–90s) and the *tail* stretches toward the low end (down to 40). Skew is named for the direction of the tail, and the tail points left (toward the smaller values), so the distribution is skewed left (also called negatively skewed).

10. The graph is misleading because the y-axis doesn't start at 0 — it starts at 90. Cutting the axis turns a small real gap into a bar that looks twice as tall. The true difference is 96 versus 92, only about a 4% increase, nowhere near "doubled." An honest version starts the y-axis at 0, which would make the two bars look nearly the same height.

11. $n = 120$ (the number of cases). Each *row* represents one patient (one individual/case). The cell in row 5, column "Age" contains a single *data value*: the age of the 5th patient — one case, one variable, one value.

12. (a) *Control*: the sugar-pill (placebo) group is the comparison group, so any effect can be measured against a baseline. (b) *Randomization*: volunteers are randomly split into the two groups, which balances out hidden differences (confounders) between them. (c) *Replication*: using 300 volunteers (many subjects, not one or two) means the result is less likely to be a fluke. Together these let the company claim cause and effect. (Bonus: "neither group is told which they got" makes it a *blind* study, guarding against response bias.)

You just rehearsed the big ideas from Chapters 1–3 — classifying variables, telling populations from samples, naming bias, spotting confounders, building a relative-frequency table, matching graphs to data, and reading skew (the Chapter 4 set below adds center and spread). If some were shaky, now you know exactly where to look. That's not failing the rehearsal; that's what the rehearsal is *for*. You're ready.

STUDY GUIDE & PRACTICE

Chapter 4 Quick Practice — Center, Spread, and Position (answers below)

1. For the data set $4, 7, 7, 10, 12$: find the mean, the median, and the mode.

2. A data set has mean $\bar{x} = 50$ and standard deviation $s = 10$. Find the z-score of the value 65 and say what it means in words.

3. A data set has five-number summary min $= 12$, $Q_1 = 20$, median $= 25$, $Q_3 = 32$, max $= 70$. (a) Find the IQR. (b) Use the $1.5 \times IQR$ rule to decide whether 70 is an outlier.

4. Class A has mean 75 with $s = 5$; Class B has mean 75 with $s = 15$. Both have the same average — what does the difference in standard deviation tell you?

Answers:

1. Mean $= \frac{4+7+7+10+12}{5} = \frac{40}{5} = 8$; median $= 7$ (the middle value of the ordered list); mode $= 7$ (it appears twice).

2. $z = \frac{65-50}{10} = 1.5$. The value 65 sits 1.5 standard deviations *above* the mean.

3. (a) $IQR = Q_3 - Q_1 = 32 - 20 = 12$. (b) Upper fence $= Q_3 + 1.5(IQR) = 32 + 1.5(12) = 32 + 18 = 50$. Since $70 > 50$, the value 70 *is* an outlier.

4. Same typical score, very different *consistency*: Class A's scores cluster tightly near 75 (small spread), while Class B's scores are much more spread out (some far above and below 75). The mean alone hides this — the standard deviation reveals it.

Part II — Formula Sheet: Describing Data

Every formula from Part II on one page. Each symbol is named in plain words, with a note on when *you'd reach for it. Bring this to the exam in your head — or on paper, if your instructor allows a sheet.*

FORMULA SHEET

The symbols, in plain words:

- x = one data value n = how many values are in the sample $\sum$ = "add up all of the..."
- $\bar{x}$ = the sample mean (average) s = the sample standard deviation s^2 = the sample variance
- Q_1, Q_3 = the first and third quartiles (25th and 75th percentiles) IQR = interquartile range

Picturing data (Chapter 3)

$$\text{relative frequency} = \frac{\text{count in the category}}{\text{total count}} \quad (\times 100 \text{ for a percent})$$

When: summarizing categorical data and building bar or pie charts.

Center — the "typical" value (Chapter 4)

$$\bar{x} = \frac{\sum x}{n} \qquad \text{median} = \text{middle value of the } \textit{ordered} \text{ list } \left(\text{position } \tfrac{n+1}{2}\right)$$

When: use the **mean** for roughly symmetric data; use the **median** when data are skewed or have outliers (the median isn't dragged by the tail). The **mode** (most frequent value) works even for categories.

Spread — how far the data wander (Chapter 4)

$$\text{range} = \max - \min \qquad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \qquad s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

When: the standard deviation s is your everyday measure of spread — read it as "the typical distance of a value from the mean." Divide by $n-1$ because this is a *sample*.

Position — where one value stands (Chapter 4)

$$z = \frac{x - \bar{x}}{s}$$

When: to say how many standard deviations a value sits above $(+)$ or below $(-)$ the mean — and to compare values measured on different scales.

The five-number summary and outliers (Chapter 4)

$$\text{min},\ Q_1,\ \text{median},\ Q_3,\ \text{max} \qquad\qquad IQR = Q_3 - Q_1$$

a value is an **outlier** if it is below $Q_1 - 1.5\,(IQR)$ or above $Q_3 + 1.5\,(IQR)$

When: to summarize a distribution's spread resistant to outliers, to build a boxplot, and to flag unusual values.

Part III

Probability & Distributions — The Bridge to Inference

Chapter 5

What Are the Chances?

The Logic of Probability — How Likely, and Why It Matters for Inference

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll hear "there's a 30% chance" or "the odds of that are basically zero" and think: "I know exactly what that number means, where it lives, and how to combine it with another one."

The Story. Probability is just the math of *how likely*. You already reason about likelihood constantly — "it'll probably rain," "no way I draw the ace." This chapter gives that instinct precise rules so it stops being a feeling and becomes something you can compute.

The Skills You'll Have:

- Name the parts of a chance situation: experiment, outcome, sample space, event
- Use the *complement* to find "the chance it *doesn't* happen"
- Combine chances with the *addition rule* (either/or) and the *multiplication rule* (this and that)
- Tell whether two events are *independent*, and read probabilities straight off a two-way table

The Confidence Moment. When someone says "the probability is 0.05, so this would almost never happen by chance," your brain will calmly translate: "out of a hundred tries, you'd see this about five times — rare, but not impossible." That sentence is the seed of every hypothesis test you'll meet later.

The Bridge. Chapter 6 takes these rules and builds *probability distributions* — a full map of every outcome and its chance at once. But you can't map outcomes until you can name and combine them. So first, the logic of chance.

MATH SKILLS YOU'LL NEED

Probability lives almost entirely in fractions, decimals, and percents — and the good news is those three are the *same number wearing different outfits*. Let's warm them up.

Skill 1 — Three faces of one number. A fraction, a decimal, and a percent are three ways to write a single amount. To go fraction $\rightarrow$ decimal, divide. To go decimal $\rightarrow$ percent, multiply by 100 (move the dot two places right).

$$\frac{3}{4} = 0.75 = 75\% \qquad \frac{1}{2} = 0.5 = 50\%$$

Skill 2 — Multiplying fractions and decimals. To multiply fractions, multiply across the top and across the bottom. With decimals, just multiply as usual.

$$\frac{1}{2} \times \frac{1}{3} = \frac{1 \times 1}{2 \times 3} = \frac{1}{6} \qquad 0.5 \times 0.2 = 0.10$$

Skill 3 — The complement (one minus a probability). If something happens with probability p, the chance it *doesn't* happen is whatever's left over to reach 1 (a whole, 100%).

$$1 - 0.30 = 0.70 \qquad 1 - \frac{1}{4} = \frac{3}{4}$$

Practice (answers at the end of the chapter):

1. Write $\frac{2}{5}$ as a decimal and as a percent.
2. Multiply $\frac{1}{4} \times \frac{2}{3}$.
3. The chance of rain today is 0.65. What is the chance of no rain?

5.1 The Language of Chance

▪ READ THIS FIRST

You flip a coin to decide who picks the movie. Before it lands, you genuinely don't know what's coming — heads or tails. But you *do* know something: there are exactly two ways it can land, and over a long evening of coin-flip arguments, it comes up heads about half the time.

You didn't take a course to know that. You've watched enough coins, dice, and card draws to feel the difference between "could happen" and "basically never." Probability just gives that feeling a number between 0 and 1.

▪ LET'S TALK ABOUT IT

Think of something with a genuinely uncertain result you'll see today — whether a bus is on time, whether a text gets answered, which song shuffles up next. How many ways could it turn out? Would you say each way is equally likely, or is one way more likely than another? (No wrong answer — you're just noticing that you already sort outcomes by how likely they feel.)

▪ NOW WE NAME IT

An **experiment** is any process whose result is uncertain before it happens: flipping a coin, rolling a die, drawing a card.

An **outcome** is one possible result of the experiment — "heads," or "the 4," or "the queen of hearts."

The **sample space** is the set of *all* possible outcomes. For one coin flip the sample space is {heads, tails}. For one die roll it's $\{1, 2, 3, 4, 5, 6\}$.

An **event** is any collection of outcomes you care about — "rolling an even number" is the event $\{2, 4, 6\}$, a slice of the sample space.

The **probability** of an event is a number from 0 to 1 measuring how likely it is. We write it $P(A)$. Read it as long-run *relative frequency*: if you repeated the experiment a huge number of times, $P(A)$ is the fraction of those times event A happens. Flip a fair coin ten thousand times and "heads" shows up close to half the flips, so $P(\text{heads}) = 0.5$.

Three rules are baked in, and they'll never change:

- Every probability is between 0 and 1: $0 \leq P(A) \leq 1$. (0 means impossible, 1 means certain.)
- The probabilities of all the outcomes in the sample space add up to exactly 1. (Something has to happen.)
- The **complement rule**: the chance an event does *not* happen is $P(\text{not } A) = 1 - P(A)$.

▪ WATCH IT WORK

Question: You roll one fair six-sided die. (a) What is the sample space? (b) What is the probability of rolling a 4? (c) What is the probability of rolling an even number? (d) What is the probability of *not* rolling a 4?

Step 1 — Write the sample space. A fair die has six equally likely outcomes: $\{1, 2, 3, 4, 5, 6\}$. Each single outcome has probability $\frac{1}{6}$, and the six of them add to $6 \times \frac{1}{6} = 1$. Good — something has to happen.

Step 2 — Probability of a 4. Just one outcome out of six: $P(4) = \dfrac{1}{6} \approx 0.167$.

Step 3 — Probability of an even number. The event "even" is $\{2, 4, 6\}$ — three outcomes out of six: $P(\text{even}) = \dfrac{3}{6} = \dfrac{1}{2} = 0.5$.

Step 4 — Probability of *not* a 4. Use the complement rule instead of recounting: $P(\text{not } 4) = 1 - P(4) = 1 - \dfrac{1}{6} = \dfrac{5}{6} \approx 0.833$.

Notice every answer landed between 0 and 1, and the complement saved us work in Step 4.

▪ YOUR TURN

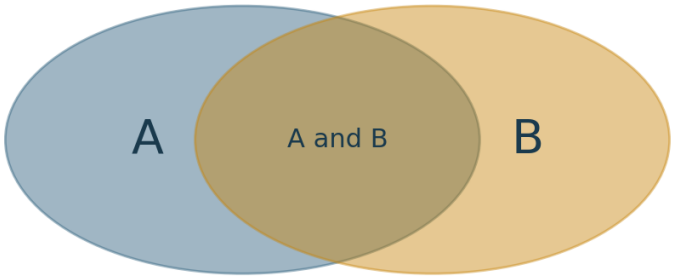
You draw one card at random from a standard 52-card deck.

1. How many outcomes are in the sample space?
2. What is the probability of drawing the ace of spades?
3. There are 13 hearts. What is the probability of drawing a heart?
4. Using the complement rule, what is the probability of *not* drawing a heart?

Work space:

▪ CHECK YOUR VOICE

If a voice just whispered "probability is where I always get lost" — hear it, then answer it. So far you've only done one thing: count the ways something can happen and divide. You did that as a kid splitting candy. The notation $P(A)$ is new; the thinking is not. Say it back: "I can count outcomes, and that's most of this."



P(A or B) = P(A) + P(B) − P(A and B)

Figure 5.1. The overlap is counted in both circles, so the addition rule subtracts it once.

5.2 Either/Or: The Addition Rule

▪ READ THIS FIRST

Someone asks, "What's the chance the card you draw is a heart *or* a face card?" Your gut wants to add: 13 hearts plus 12 face cards. But wait — the jack, queen, and king of hearts are *both* hearts *and* face cards. If you just add, you count those three twice.

You already know this from real life: if 18 people like pizza and 12 like tacos, you can't say 30 people total like one of them — some people like both, and you'd be double-counting them. The addition rule is just "add the two groups, then subtract the overlap so nobody gets counted twice."

▪ LET'S TALK ABOUT IT

Picture two overlapping circles — one for "students who play a sport," one for "students who have a job." Where do the people who do *both* sit? If you wanted the total number of students who do at least one of the two, why can't you simply add the two circle counts? (You're rediscovering the addition rule before we even name it.)

▪ NOW WE NAME IT

The **addition rule** finds the probability that event A happens *or* event B happens (or both):

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

You add the two chances, then subtract the overlap — the chance of *both* — because that overlap got counted once in $P(A)$ and again in $P(B)$.

Two events are **mutually exclusive** (also called **disjoint**) when they cannot both happen at the same time — there is no overlap. "Rolling a 2" and "rolling a 5" on one die are mutually exclusive: one roll can't be both. For mutually exclusive events the overlap is zero, $P(A \text{ and } B) = 0$, so the rule simplifies to

$$P(A \text{ or } B) = P(A) + P(B).$$

The lesson: *always* ask whether the events can happen together. If they can, subtract the overlap. If they can't, the subtraction is just subtracting zero.

▪ WATCH IT WORK

Question: Draw one card from a 52-card deck. (a) What is the probability it's a heart or a face card (jack, queen, king)? (b) What is the probability it's a heart or a spade?

Part (a) — events that overlap.

Step 1 — Find each piece. There are 13 hearts, so $P(\text{heart}) = \dfrac{13}{52}$. There are 12 face cards (3 in each of 4 suits), so $P(\text{face}) = \dfrac{12}{52}$.

Step 2 — Find the overlap. Cards that are *both* a heart and a face card: the jack, queen, and king of hearts — 3 cards. So $P(\text{heart and face}) = \dfrac{3}{52}$.

Step 3 — Apply the addition rule.

$$P(\text{heart or face}) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26} \approx 0.423.$$

Without the $-\frac{3}{52}$ we'd have gotten $\frac{25}{52}$ — wrong, because the three red face cards would be counted twice.

Part (b) — mutually exclusive events.

Step 1 — Can a card be both? A single card cannot be a heart *and* a spade at the same time. So these events are mutually exclusive and $P(\text{heart and spade}) = 0$.

Step 2 — Add (overlap is zero).

$$P(\text{heart or spade}) = \frac{13}{52} + \frac{13}{52} = \frac{26}{52} = \frac{1}{2} = 0.5.$$

▪ YOUR TURN

Roll one fair six-sided die.

1. Are "rolling an even number" and "rolling a 3" mutually exclusive? Why?
2. What is $P(\text{even or } 3)$?
3. Now consider "rolling an even number" and "rolling a number greater than 3" ($\{4, 5, 6\}$). What is the overlap event, and what is its probability?
4. Use the addition rule to find $P(\text{even or greater than } 3)$.

Work space:

▪ CHECK YOUR VOICE

If you forgot to subtract the overlap on your first try — that is the single most common slip in this entire topic, and noticing it is exactly how you stop making it. You didn't "fail at probability." You met the double-counting trap once, on purpose, in a safe place. Now it's yours.

5.3 This AND That: Conditional Probability and Independence

▪ READ THIS FIRST

You're guessing whether a stranger likes a certain band. Knowing nothing, you'd shrug. But now I tell you they're wearing that band's tour shirt — suddenly your guess shoots up. The *new information* changed the chance.

That's conditional probability: the chance of something *given that* you already know something else. You do this all day — "odds they're home? higher now that I see their car in the driveway." We're just going to write it down precisely.

▪ LET'S TALK ABOUT IT

Think of a guess that changes once you learn one extra fact: the chance it's raining (then you hear thunder), the chance a friend is free (then you learn it's their day off). Does every extra fact change your guess? Can you think of a fact that would *not* change it at all? (Hold that "doesn't change it" idea — it's secretly the definition of independence.)

▪ NOW WE NAME IT

The **conditional probability** of A *given* B is the chance A happens once you already know B happened. We write $P(A \mid B)$ and compute it by zooming in on just the B world:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

You restrict attention to the times B happens (the denominator), then ask what fraction of *those* also have A.

Rearranging that formula gives the **multiplication rule** for the chance of A *and* B both happening:

$$P(A \text{ and } B) = P(A)\,P(B \mid A).$$

In words: the chance both happen is the chance of the first, times the chance of the second *given* the first already did.

Two events are **independent** when knowing one happened doesn't change the chance of the other — that "fact that doesn't move your guess" from the Talk box. Formally, A and B are independent when

$$P(A \mid B) = P(A).$$

When events are independent, the multiplication rule loses its condition and becomes a clean product:

$$P(A \text{ and } B) = P(A)\,P(B).$$

Coin flips are the classic example: the coin has no memory, so the second flip's chance of heads is $\frac{1}{2}$ no matter what the first did.

▪ WATCH IT WORK

Question: (a) Flip a fair coin twice. What is the probability of heads both times? (b) Draw two cards from a 52-card deck *without replacing* the first. What is the probability both are hearts?

Part (a) — independent events.

Step 1 — Check independence. The coin has no memory: the second flip's chance of heads is $\frac{1}{2}$ regardless of the first. Independent.

Step 2 — Multiply.

$$P(\text{heads and heads}) = P(\text{heads})\,P(\text{heads}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0.25.$$

Part (b) — dependent events (the deck changes).

Step 1 — First card. There are 13 hearts in 52 cards: $P(\text{first heart}) = \dfrac{13}{52} = \dfrac{1}{4}$.

Step 2 — Second card, given the first was a heart. Now one heart is gone, so 12 hearts remain among 51 cards: $P(\text{second heart} \mid \text{first heart}) = \dfrac{12}{51}$.

Step 3 — Multiply with the conditional.

$$P(\text{both hearts}) = \frac{13}{52} \times \frac{12}{51} = \frac{1}{4} \times \frac{12}{51} = \frac{12}{204} = \frac{1}{17} \approx 0.059.$$

This is *smaller* than if we'd wrongly treated the draws as independent ($\frac{1}{4} \times \frac{1}{4} = \frac{1}{16} \approx 0.063$), because removing a heart made the second heart a bit less likely. The condition matters.

▪ YOUR TURN

1. You roll a fair die and flip a fair coin. Are these independent? What is $P(\text{roll a 6 and flip heads})$?
2. A bag has 5 red and 3 blue marbles (8 total). You draw two *without replacement*. What is the probability the first is red?
3. Given the first draw was red, what is the probability the second is also red? (Remember: one red is gone.)
4. Use the multiplication rule to find the probability both marbles are red.

Work space:

▪ CHECK YOUR VOICE

If part (b) of the worked example made your stomach tighten — "wait, the number on the bottom changed?" — that tightening is your brain doing the real work, not failing. Dependence is genuinely the harder idea, and you just watched exactly why the denominator drops from 52 to 51. That's not confusion; that's understanding arriving.

5.4 Reading Probabilities from a Two-Way Table

▪ READ THIS FIRST

You've seen a table like this a hundred times without calling it statistics: a grid where the rows are one trait (say, "has a job / no job") and the columns are another ("lives on campus / off campus"), and each box holds a count of how many people fall into that combination.

Every probability we've learned — single events, *and*, *or*, conditional — can be read straight off such a grid by counting the right boxes. It's the most practical probability skill you'll use, and it's just careful reading.

▪ LET'S TALK ABOUT IT

Imagine surveying your class: do you commute, and do you work a job? You'd get four kinds of people (commute+job, commute+no job, on-campus+job, on-campus+no job). Before any formula — if you wanted "the chance a person works a job, *among commuters only*," which boxes would you look at, and which total would you divide by? (You're inventing conditional probability from a table.)

NOW WE NAME IT

A **two-way table** (also called a **contingency table**) cross-classifies every case by two categorical variables, with a count in each cell and *totals* along the margins.
Three kinds of probability come straight off the table:

- A **marginal probability** uses a margin total: one variable only, ignoring the other (e.g. $P(\text{works a job}) =$ the job row total over the grand total).
- A **joint probability** is $P(A \text{ and } B)$: a single inside cell over the grand total (e.g. $P(\text{commuter and works})$).
- A **conditional probability** $P(A \mid B)$ zooms into one row or column: a cell divided by *that row or column's* total, not the grand total.

The whole trick is the denominator. Marginal and joint probabilities divide by the *grand total* (everybody). A conditional probability divides by a *margin* (just the group you're given). Get the denominator right and the rest is reading.

▪ WATCH IT WORK

Question: A class of 100 students was surveyed on whether they commute and whether they work a job. Here are the counts.

	Works a Job	No Job	Total
Commuter	30	20	50
Lives on Campus	15	35	50
Total	45	55	100

Step 1 — Marginal: P(works a job). Use the column total over the grand total: $\frac{45}{100} = 0.45$.

Step 2 — Joint: P(commuter and works a job). Use the single inside cell (commuter *and* works) over the grand total: $\frac{30}{100} = 0.30$.

Step 3 — Conditional: P(works a job | commuter). We're given the student is a commuter, so zoom into the commuter row (50 students) and ask how many of *those* work: $\frac{30}{50} = 0.60$. Notice the denominator is 50, not 100 — we only look at commuters.

Step 4 — Check independence. Are "works a job" and "commuter" independent? Compare $P(\text{works}) = 0.45$ with $P(\text{works} \mid \text{commuter}) = 0.60$. They're different, so the events are *not* independent — being a commuter raises the chance of working.

Step 5 — Verify with the formula. The conditional formula should agree with Step 3: $P(\text{works} \mid \text{commuter}) = \frac{P(\text{works and commuter})}{P(\text{commuter})} = \frac{0.30}{0.50} = 0.60$. It matches.

▪ YOUR TURN

Use the same 100-student table above.

1. What is P(lives on campus)? (marginal)
2. What is P(lives on campus and no job)? (joint)
3. What is P(no job | lives on campus)? (conditional — which total is your denominator?)
4. What is P(commuter or works a job)? (Hint: use the addition rule and subtract the overlap.)

Work space:

▪ CHECK YOUR VOICE

Look at what just happened: you pulled marginal, joint, and conditional probabilities — plus an independence check — out of one little grid of numbers. A page ago that grid was just "a table." The thing that scares people about this topic was never the table; it was not knowing which number to divide by. You know now. Say it: "I pick the right denominator."

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 5 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 5 — Your Turn Answers

Math Skills: (1) $\frac{2}{5} = 0.4 = 40\%$. (2) $\frac{1}{4} \times \frac{2}{3} = \frac{2}{12} = \frac{1}{6}$. (3) $1 - 0.65 = 0.35$ (a 35% chance of no rain).

Section 5.1: (1) 52 outcomes (one per card). (2) $P(\text{ace of spades}) = \frac{1}{52} \approx 0.019$. (3) $P(\text{heart}) = \frac{13}{52} = \frac{1}{4} = 0.25$. (4) $P(\text{not heart}) = 1 - \frac{1}{4} = \frac{3}{4} = 0.75$.

Section 5.2: (1) Yes, mutually exclusive — 3 is odd, so a roll can't be both even and a 3 at once (overlap is 0). (2) $P(\text{even or } 3) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \approx 0.667$. (3) The overlap of "even" $\{2, 4, 6\}$ and "greater than 3" $\{4, 5, 6\}$ is $\{4, 6\}$, so $P(\text{even and } >3) = \frac{2}{6} = \frac{1}{3}$. (4) $P(\text{even or } >3) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3} \approx 0.667$ (the event is $\{2, 4, 5, 6\}$ — four outcomes, which checks out).

Section 5.3: (1) Independent (the die doesn't affect the coin): $P(6 \text{ and heads}) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12} \approx 0.083$. (2) $P(\text{first red}) = \frac{5}{8} = 0.625$. (3) Given one red is gone, 4 reds remain of 7 marbles: $P(\text{second red} \mid$

first red) $= \frac{4}{7}$. (4) $P(\text{both red}) = \frac{5}{8} \times \frac{4}{7} = \frac{20}{56} = \frac{5}{14} \approx 0.357$.

Section 5.4: (1) $P(\text{on campus}) = \frac{50}{100} = 0.50$. (2) $P(\text{on campus and no job}) = \frac{35}{100} = 0.35$. (3) Denominator is the on-campus total, 50: $P(\text{no job} \mid \text{on campus}) = \frac{35}{50} = 0.70$. (4) Using the addition rule with $P(\text{commuter}) = \frac{50}{100}$, $P(\text{works}) = \frac{45}{100}$, and overlap $P(\text{commuter and works}) = \frac{30}{100}$: $P(\text{commuter or works}) = \frac{50}{100} + \frac{45}{100} - \frac{30}{100} = \frac{65}{100} = 0.65$.

Chapter 6

Predictable Randomness

Probability Distributions, the Binomial, and the Famous Bell Curve

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll see a random thing — a coin flip, a quiz guess, a person's height — and realize the randomness has a *shape*. Once you know the shape, you can predict it. That's the quiet magic of this whole chapter: random does not mean unknowable.

The Story. You already expect patterns inside chaos. You know roughly half of coin flips come up heads, that most people are "average height" and very tall or very short people are rare. Probability distributions just write those instincts down as math you can compute with.

The Skills You'll Have:

- Read a probability distribution and find its *expected value* (mean) and *standard deviation*
- Recognize a *binomial* situation and compute the probability of exactly k successes
- Use the *Empirical Rule* (68–95–99.7) to describe a bell curve in your head
- Turn any value into a z*-score* and find a probability from it — and go backward from a percentile

The Confidence Moment. When you see the binomial formula $\binom{n}{k}p^k(1-p)^{n-k}$, your stomach won't drop. You'll think: "ways to win, times the wins, times the losses. I can plug in numbers."

The Bridge. Chapter 7 takes the bell curve and does something almost unbelievable with it: it shows that *averages* of samples form a bell curve too, even when the original data doesn't. That's the Central Limit Theorem — and it's why everything after this works. But first, we meet the distributions themselves.

MATH SKILLS YOU'LL NEED

This chapter asks you to plug numbers into a couple of formulas carefully. None of it is hard — it's bookkeeping. Let's warm up the exact moves you'll use.

Skill 1 — Exponents (repeated multiplication). A small raised number just means "multiply that many times."

$$0.5^3 = 0.5 \times 0.5 \times 0.5 = 0.125 \qquad 0.8^2 = 0.8 \times 0.8 = 0.64$$

A handy fact you'll need: anything to the power 0 equals 1. So $0.5^0 = 1$.

Skill 2 — Factorials and combinations. A *factorial*, written $n!$, multiplies every whole number from n down to 1: $4! = 4 \times 3 \times 2 \times 1 = 24$. (And $0! = 1$ by definition — don't fight it, just use it.)

A *combination* counts how many ways you can *choose* k items out of n when order doesn't matter:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Read $\binom{5}{2}$ as "5 choose 2." It equals $\dfrac{5!}{2!\,3!} = \dfrac{120}{2 \times 6} = \dfrac{120}{12} = 10$. There are 10 ways to pick 2 things out of 5.

Skill 3 — Substituting carefully. The whole game is: write the formula, replace each letter with its number, then simplify one piece at a time. Slow is smooth; smooth is correct.

Practice (answers at the end of the chapter):

1. Compute 0.2^2.
2. Compute $3!$ (three factorial).
3. Compute $\binom{4}{1}$ ("4 choose 1").

6.1 Random Variables and Their Distributions

▪ READ THIS FIRST

Imagine you roll one ordinary die. Before it lands, you can't say which face shows up — but you *can* list every possibility (1 through 6) and you know each is equally likely. You don't know the outcome, yet you know the menu.

That's the heart of it. A random thing has a fixed list of possible values, each with its own chance of happening. Write that list down with its chances, and you've captured the randomness on paper.

▪ LET'S TALK ABOUT IT

Think of something in your day with a few possible outcomes and known-ish chances — the number of texts you get in an hour, how many green lights you hit on your drive. Could you list the possible values? Could you guess a rough chance for each? You're already thinking in distributions.

▪ NOW WE NAME IT

A **random variable** is a quantity whose value comes from a random process — the result of a die roll, the number of customers in the next hour. We usually call it X.

A **discrete random variable** can only take separate, countable values (like $0, 1, 2, 3$). Its **probability distribution** is a table listing each possible value x next to its probability $P(x)$. Two rules always hold: every $P(x)$ is between 0 and 1, and *all the probabilities add up to exactly* 1 (something must happen).

Two numbers summarize a distribution. The **expected value** (also called the mean) is what you'd average if you repeated the process forever:

$$\mu = \sum x\,P(x)$$

That's "multiply each value by its chance, then add up." The **standard deviation** measures how spread out the outcomes are around that mean:

$$\sigma = \sqrt{\sum (x-\mu)^2\,P(x)}$$

Don't panic at the symbols — $\sum$ just means "add these up," and we'll do it one row at a time.

▪ WATCH IT WORK

Question: A coffee cart sells either 0, 1, or 2 pastries to each customer. Records show $P(0) = 0.5$, $P(1) = 0.3$, $P(2) = 0.2$. Find the mean and standard deviation.

Step 0 — Sanity check. Do the probabilities add to 1? $0.5 + 0.3 + 0.2 = 1.0$. Yes. Good, it's a valid distribution.

Step 1 — Find the mean $\mu = \sum x\,P(x)$. Multiply each value by its probability, then add:

$$\mu = (0)(0.5) + (1)(0.3) + (2)(0.2) = 0 + 0.3 + 0.4 = 0.7$$

So on average, a customer buys 0.7 pastries. (Nobody buys 0.7 pastries — it's a long-run average, like "2.1 kids per family.")

Step 2 — Find each $(x - \mu)^2 P(x)$. Subtract the mean, square it, multiply by the probability:

- $x = 0$: $(0 - 0.7)^2(0.5) = (0.49)(0.5) = 0.245$
- $x = 1$: $(1 - 0.7)^2(0.3) = (0.09)(0.3) = 0.027$
- $x = 2$: $(2 - 0.7)^2(0.2) = (1.69)(0.2) = 0.338$

Step 3 — Add them and take the square root. The sum is $0.245 + 0.027 + 0.338 = 0.61$. Then

$$\sigma = \sqrt{0.61} \approx 0.78 \text{ pastries.}$$

That's it. Mean 0.7, standard deviation about 0.78. You just summarized a random process with two numbers.

▪ YOUR TURN

A small raffle pays out per ticket: \$0 with probability 0.6, \$5 with probability 0.3, and \$20 with probability 0.1. Let X be the payout.

1. Check that the probabilities add to 1.
2. Find the expected payout μ.
3. Find the standard deviation σ.

Work space:

▪ CHECK YOUR VOICE

If the standard-deviation formula made your eyes widen — that's fair, it has a lot of pieces. But notice what you actually did: subtract, square, multiply, add, square-root. Five small steps you already know. The formula looks scary only when you stare at the whole thing at once. One row at a time, it's just arithmetic. Say it: "I can do this in pieces."

6.2 The Binomial Distribution

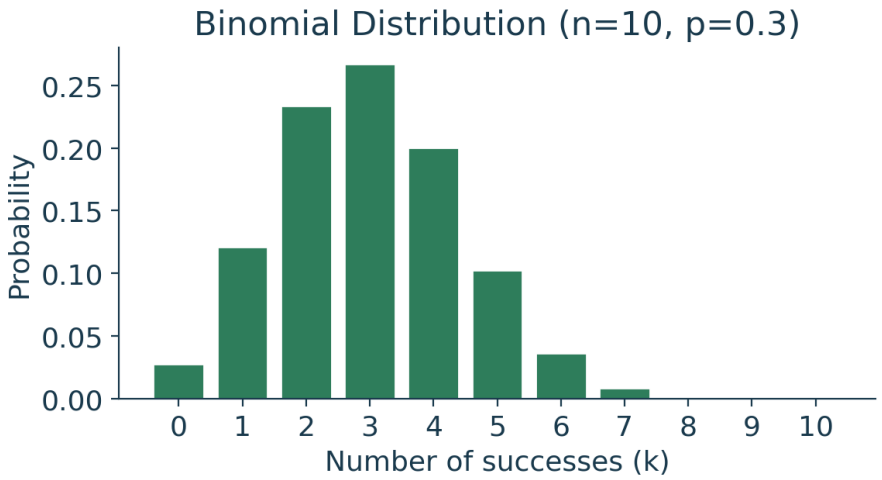


Figure 6.1. A binomial distribution ($n = 10$, $p = 0.3$): the probability of each number of successes.

▪ READ THIS FIRST

Picture guessing on a 5-question true/false quiz. Each question is its own little coin flip — right or wrong, with the same chance every time. You can't know exactly how many you'll get right, but "how many out of 5" is a question with a real, computable answer.

That repeated yes/no, same-chance-each-time setup shows up everywhere: free throws made out of 10, defective parts out of 20, people who say "yes" out of 100. It has a name and a formula.

▪ LET'S TALK ABOUT IT

Think of a situation in your life that's a fixed number of independent yes/no tries with the same chance each time — making the bus on time each morning this week, sinking free throws, a recipe turning out. What counts as a "success"? Roughly what's the chance each try?

▪ NOW WE NAME IT

A **binomial distribution** counts the number of successes in a fixed number of independent yes/no trials. It applies only when *all four conditions* hold (memory hook: **B-I-N-S**):

- **B**inary: each trial has two outcomes — "success" or "failure."
- **I**ndependent: trials don't affect each other.
- **N**umber fixed: there's a set number of trials, n.
- **S**ame probability: the success probability p is the same every trial.

Let X be the number of successes. The probability of getting *exactly* k successes is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Read it left to right as a story: $\binom{n}{k}$ is the *number of ways* to arrange k successes among n trials; p^k is the chance of those k successes; $(1-p)^{n-k}$ is the chance of the remaining failures. Ways $\times$ wins $\times$ losses.

Two shortcuts you'll use constantly — the mean and standard deviation of a binomial:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

▪ WATCH IT WORK

Question: You guess on a 5-question true/false quiz, so $n = 5$ and $p = 0.5$ (a 50% chance per question). What's the probability of getting exactly 3 right?

Step 1 — Confirm it's binomial. Five fixed questions (N), each right/wrong (B), each independent (I), each with the same $p = 0.5$ (S). All four hold. Here $k = 3$.

Step 2 — Write the formula with numbers in.

$$P(X = 3) = \binom{5}{3} (0.5)^3 (0.5)^{5-3} = \binom{5}{3} (0.5)^3 (0.5)^2$$

Step 3 — Compute the combination. $\binom{5}{3} = \dfrac{5!}{3!\,2!} = \dfrac{120}{6 \times 2} = \dfrac{120}{12} = 10$. There are 10 ways to be right on 3 of the 5 questions.

Step 4 — Compute the powers. $(0.5)^3 = 0.125$ and $(0.5)^2 = 0.25$.

Step 5 — Multiply it all together.

$$P(X = 3) = 10 \times 0.125 \times 0.25 = 10 \times 0.03125 = 0.3125$$

So there's about a 31.25% chance of getting exactly 3 right by pure guessing.

Bonus — mean and SD. $\mu = np = 5(0.5) = 2.5$ questions right on average, and $\sigma = \sqrt{np(1-p)} = \sqrt{5(0.5)(0.5)} = \sqrt{1.25} \approx 1.12$.

▪ YOUR TURN

A free-throw shooter makes each shot with probability $p = 0.8$, independently, and takes $n = 4$ shots.

1. Confirm the four binomial conditions are reasonable here.
2. Find the probability she makes exactly 3 of the 4. (You'll need $\binom{4}{3}$, 0.8^3, and 0.2^1.)
3. Find the mean μ and standard deviation σ for the number she makes.

Work space:

▪ CHECK YOUR VOICE

The binomial formula is where a lot of people decide "I'm just not a math person." Catch that thought. You didn't fail — you met a formula with three parts and computed each part separately, exactly like following a recipe. If you got a probability between 0 and 1, you did it right. The formula is long, not hard. Talk back: "Long isn't the same as hard."

6.3 The Normal Distribution (the Bell Curve)

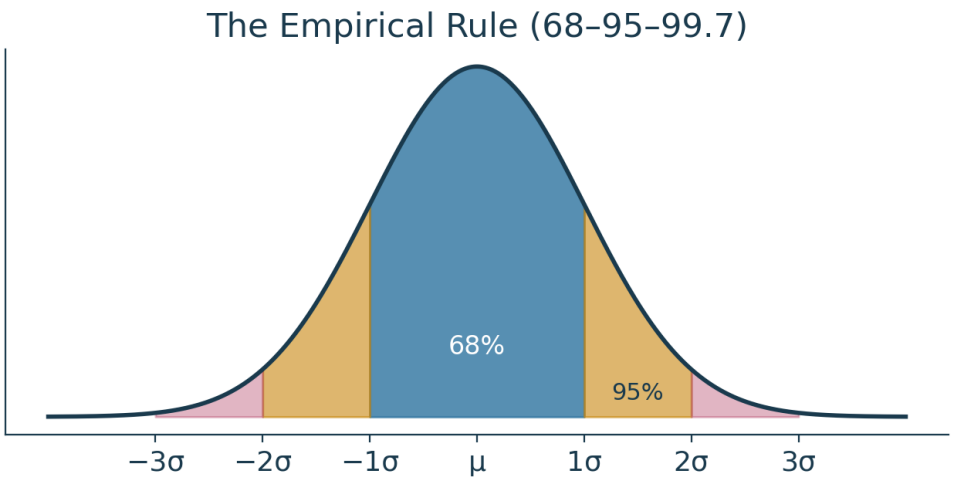


Figure 6.2. The Empirical Rule: about 68%, 95%, and 99.7% of data fall within 1, 2, and 3 standard deviations of the mean.

▪ READ THIS FIRST

Picture the heights of all the adults you know. A few are quite short, a few are quite tall, but most cluster in the middle, and the further you get from average in either direction, the rarer people become. Sketch that with a pen and you draw a smooth hill — high in the middle, tapering symmetrically on both sides.

That hill is the most famous shape in statistics: the bell curve. Tons of real-world measurements — heights, test scores, measurement errors — pile up in exactly this shape.

▪ LET'S TALK ABOUT IT

Think of a measurement where most people land near the middle and extremes are rare: shoe size, resting heart rate, how long your commute takes. If you imagined a graph of "how common is each value," would it look like a centered hill? That hill is the normal curve.

▪ NOW WE NAME IT

A **normal distribution** is a continuous distribution whose graph is a symmetric, bell-shaped curve. Picture a smooth mound: tallest at the center, sloping down evenly on both sides, never quite touching the floor. It is completely described by two numbers — its mean μ (where the peak sits, the center) and its standard deviation σ (how wide and spread out the hill is). A big σ makes a short, wide hill; a small σ makes a tall, narrow one.

Because the shape is so regular, we can say exactly how much of the data falls near the center. That's the **Empirical Rule** (the 68–95–99.7 rule), and it's worth memorizing:

- About **68%** of the data falls within 1σ of the mean (between $\mu - \sigma$ and $\mu + \sigma$).
- About **95%** falls within 2σ of the mean (between $\mu - 2\sigma$ and $\mu + 2\sigma$).
- About **99.7%** (almost everything) falls within 3σ of the mean.

Because the curve is symmetric, each of those percentages splits evenly into the two halves — half above the mean, half below.

▪ WATCH IT WORK

Question: Adult resting heart rates are roughly normal with mean $\mu = 70$ beats per minute and standard deviation $\sigma = 8$. Use the Empirical Rule to describe the data.

Step 1 — Mark off one, two, and three standard deviations. Add and subtract $\sigma = 8$ from the mean repeatedly:

- 1σ: from $70 - 8 = 62$ to $70 + 8 = 78$
- 2σ: from $70 - 16 = 54$ to $70 + 16 = 86$
- 3σ: from $70 - 24 = 46$ to $70 + 24 = 94$

Step 2 — Attach the percentages. About 68% of people have a resting heart rate between 62 and 78. About 95% fall between 54 and 86. About 99.7% fall between 46 and 94.

Step 3 — Use symmetry for a one-sided question. "What fraction have a rate above 78?" Since 68% sit inside 62–78, that leaves 32% outside, split evenly between the two tails. So about 16% are above 78 (and 16% below 62).

▪ YOUR TURN

Scores on a placement test are roughly normal with mean $\mu = 500$ and standard deviation $\sigma = 100$.

1. Between what two scores do about 95% of students fall?
2. About what percent of students score between 400 and 600?
3. About what percent score *above* 700? (Hint: 700 is 2σ above the mean; use symmetry.)

Work space:

▪ CHECK YOUR VOICE

If memorizing 68–95–99.7 feels like one more thing to cram — here's a gift: you only have to remember three numbers and the picture of a hill. Everything else (the cutoffs, the percentages in the tails) comes from adding, subtracting, and using symmetry. You're not memorizing a hundred facts; you're memorizing three and reasoning out the rest. That's the opposite of cramming.

6.4 Standard Scores and Normal Probabilities

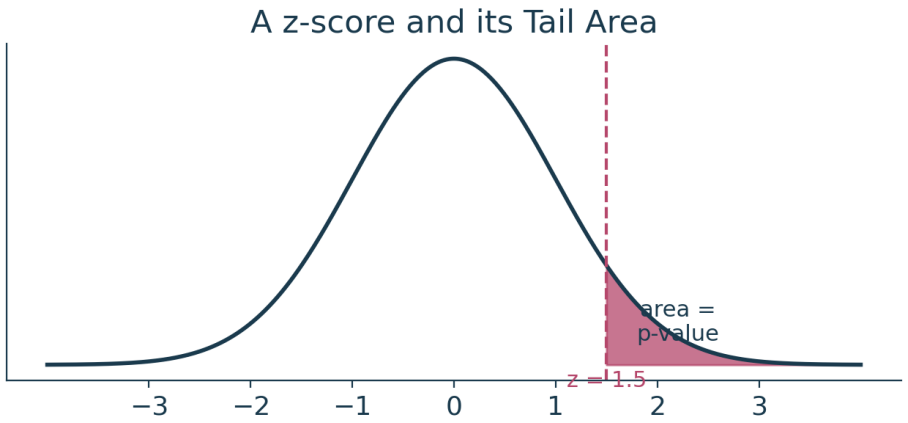


Figure 6.3. A z-score marks a location on the standard normal curve; the shaded tail area is a probability.

▪ READ THIS FIRST

Suppose you scored 85 on one exam and 90 on another. Which was the better performance? You can't tell yet — it depends on how everyone else did and how spread out the scores were. A raw score alone doesn't tell you where you stand.

What you really want is: "How many standard deviations above or below average am I?" That single number lets you compare scores from totally different tests on one fair scale. Statisticians call it a z-score.

▪ LET'S TALK ABOUT IT

Have you ever heard a result described as "two standard deviations above average" — maybe a baby's height percentile, or a standardized test? That phrasing *is* a z-score. Where else have you seen "percentile" used to say where someone stands in a crowd?

▪ NOW WE NAME IT

The **standard normal distribution** is just the normal curve with mean 0 and standard deviation 1 — the "reference" bell curve we measure everything against. Any normal value can be placed on it by converting to a **z-score** (or standard score):

$$z = \frac{x - \mu}{\sigma}$$

In words: subtract the mean (how far above or below average you are), then divide by the standard deviation (how big a "step" is). A z of $+1$ means "one standard deviation above the mean"; a z of -2 means "two below." Positive is above average, negative is below, and 0 is exactly average. Once you have a z-score, you can look up the **probability** (the area under the curve) to its left — that area is the fraction of data below your value, i.e. your *percentile*. You can find this area with technology (a spreadsheet or jamovi) or a printed z-table. And you can run it *backward*: start from a desired percentile, find the matching z, then unstandardize with $x = \mu + z\sigma$ to get the actual value.

▪ WATCH IT WORK

Question: Exam scores are normal with $\mu = 75$ and $\sigma = 10$. (a) What is the z-score for a student who scored 88? (b) What percentile is that, roughly? (c) What score sits at the 90th percentile?

Part (a) — Standardize. Plug into the formula:

$$z = \frac{x - \mu}{\sigma} = \frac{88 - 75}{10} = \frac{13}{10} = 1.3$$

The student scored 1.3 standard deviations above the mean.

Part (b) — Find the area to the left. Using technology or a z-table, the area to the left of $z = 1.3$ is about 0.9032. So this student scored higher than about 90% of the class — roughly the 90th percentile.

Part (c) — Go backward from a percentile. We want the score with 90% below it. First find the z for the 90th percentile: from a table or tool, that's about $z = 1.28$. Now unstandardize:

$$x = \mu + z\sigma = 75 + (1.28)(10) = 75 + 12.8 = 87.8$$

A score of about 88 marks the 90th percentile — which matches part (b), as it should.

▪ YOUR TURN

Heights of adult women are roughly normal with $\mu = 64$ inches and $\sigma = 2.5$ inches.

1. Find the z-score for a woman who is 69 inches tall.
2. Is that above or below average, and by how many standard deviations?
3. The 25th percentile corresponds to about $z = -0.67$. What height is that? (Use $x = \mu + z\sigma$.)

Work space:

▪ CHECK YOUR VOICE

The z-score formula is short, but the *idea* can feel slippery the first time, especially going backward from a percentile. That's normal — you're learning to move in two directions at once. If forward (value $\to z$) makes sense and backward ($z \to$ value) still feels wobbly, you haven't failed; you've just practiced one direction more than the other. Do one more backward problem and watch it click. Say it: "Wobbly now means solid soon."

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 6 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 6 — Your Turn Answers

Math Skills: (1) $0.2^2 = 0.2 \times 0.2 = 0.04$. (2) $3! = 3 \times 2 \times 1 = 6$. (3) $\binom{4}{1} = \dfrac{4!}{1!\,3!} = \dfrac{24}{1 \times 6} = 4$.

Section 6.1: (1) $0.6 + 0.3 + 0.1 = 1.0$, valid. (2) $\mu = (0)(0.6) + (5)(0.3) + (20)(0.1) = 0 + 1.5 + 2.0 = \$3.50$. (3) Compute each $(x - \mu)^2 P(x)$ with $\mu = 3.5$: $x = 0$: $(0 - 3.5)^2(0.6) = (12.25)(0.6) = 7.35$; $x = 5$: $(5 - 3.5)^2(0.3) = (2.25)(0.3) = 0.675$; $x = 20$: $(20 - 3.5)^2(0.1) = (272.25)(0.1) = 27.225$. Sum $= 7.35 + 0.675 + 27.225 = 35.25$, so $\sigma = \sqrt{35.25} \approx \$5.94$.

Section 6.2: (1) Four fixed shots (N), each made/missed (B), assumed independent (I), each with the same $p = 0.8$ (S) — reasonable. (2) $P(X = 3) = \binom{4}{3}(0.8)^3(0.2)^1$. Here $\binom{4}{3} = 4$, $0.8^3 = 0.512$, $0.2^1 = 0.2$, so $P(X = 3) = 4 \times 0.512 \times 0.2 = 0.4096 \approx 41\%$. (3) $\mu = np = 4(0.8) = 3.2$ shots made; $\sigma = \sqrt{np(1-p)} = \sqrt{4(0.8)(0.2)} = \sqrt{0.64} = 0.8$.

Section 6.3: (1) 95% fall within 2σ: from $500 - 200 = 300$ to $500 + 200 = 700$. (2) 400 to 600 is $\mu \pm 1\sigma$, so about 68%. (3) 700 is 2σ above the mean. Since 95% fall within $\pm 2\sigma$, the remaining 5% splits between the two tails, so about 2.5% score above 700.

Section 6.4: (1) $z = \dfrac{69 - 64}{2.5} = \dfrac{5}{2.5} = 2.0$. (2) Above average, by 2 standard deviations. (3) $x = \mu + z\sigma = 64 + (-0.67)(2.5) = 64 - 1.675 \approx 62.3$ inches.

Chapter 7

From Sample to Truth

Sampling Distributions and the Central Limit Theorem — the Engine Behind Inference

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll understand why a single sample can tell you something trustworthy about a whole population — and exactly how much to trust it. The scary-sounding "Central Limit Theorem" will turn out to be a promise that works *in your favor*.

The Story. Back in Chapter 1 you tasted one spoonful of soup to judge the whole pot. But here's the thing nobody mentioned: if you took a *second* spoonful, it would taste a little different. And a third, different again. This chapter is about that wobble — and the beautiful, predictable pattern hiding inside it.

The Skills You'll Have:

- Explain what a *sampling distribution* is in plain words
- Compute the *standard error* of a sample mean, $\frac{\sigma}{\sqrt{n}}$
- State the Central Limit Theorem and say why it's a gift
- Find a probability about a sample mean using a z-score
- Do the same for a sample proportion

The Confidence Moment. When someone says "the standard error of the mean," your brain will calmly translate: "how much the sample average naturally bounces around the real truth."

The Bridge. This chapter is the *engine*. Everything in Part IV (confidence intervals, Chapter 8) and Part V (hypothesis tests) runs on the one idea you'll learn here: a sample statistic varies in a predictable, measurable way. Once you can measure the wobble, you can make honest statements about the truth.

MATH SKILLS YOU'LL NEED

This chapter asks you to do one small computation over and over: divide a number by a square root. Let's warm up the exact moves.

Skill 1 — Square roots. A square root undoes squaring. $\sqrt{36} = 6$ because $6 \times 6 = 36$. For numbers that aren't perfect squares, use a calculator: $\sqrt{30} \approx 5.477$. You only ever need the decimal.

Skill 2 — Dividing by a square root. Compute the square root *first*, then divide.

$$\frac{15}{\sqrt{25}} = \frac{15}{5} = 3 \qquad \frac{12}{\sqrt{40}} = \frac{12}{6.3246} \approx 1.897$$

Skill 3 — Evaluating $\frac{\sigma}{\sqrt{n}}$. This is the only formula you'll repeat. If $\sigma = 20$ and $n = 100$:

$$\frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$$

Skill 4 — A quick z-score refresher. From Chapter 6, a z-score counts how many standard deviations a value sits above $(+)$ or below $(-)$ the center: $z = \frac{\text{value} - \text{center}}{\text{spread}}$. Then a normal table (or jamovi) turns z into a probability.

Practice (answers at the end of the chapter):

1. Compute $\sqrt{64}$ and $\sqrt{50}$ (round to three decimals).
2. Evaluate $\frac{18}{\sqrt{36}}$.
3. If $\sigma = 30$ and $n = 25$, evaluate $\frac{\sigma}{\sqrt{n}}$.

7.1 Statistics Have Their Own Distributions

▪ READ THIS FIRST

Remember the pot of soup. You stirred it, took one spoonful, and judged the whole pot from that taste.

Now do it again. Take a *second* spoonful from the same pot. It won't taste exactly identical — maybe a hair saltier, maybe a touch less. A third spoonful: different again. None of them is "wrong." They're all honest tastes of the same soup. They just wobble a little around the real flavor of the pot.

Here's the quiet insight: those spoonful-to-spoonful differences aren't random chaos. If you wrote down the saltiness of a hundred spoonfuls, the numbers would pile up into a recognizable, predictable pattern. The wobble has a *shape*.

▪ LET'S TALK ABOUT IT

Imagine you and four friends each grabbed a separate spoonful from the same big pot and rated its saltiness from 1 to 10. Would all five of you say the exact same number? Probably not — but would they be wildly far apart, like a 2 and a 9? Also probably not. Where do you think most of the ratings would land?

▪ NOW WE NAME IT

A sample statistic — like the sample mean $\bar{x}$ — is a number you calculate from one sample. Take a *different* sample, and you'll get a slightly different $\bar{x}$. Take many samples, and you collect many slightly different $\bar{x}$ values.

The **sampling distribution** of a statistic is the distribution of all the values that statistic could take across all possible samples of a given size n. It is, quite literally, the pattern of the wobble.

This is a genuinely new idea, so go slow. In Chapter 4 a distribution described how *data values* spread out (the heights of 80 people). A sampling distribution describes how a *statistic* spreads out (the average height from 80 people, computed again and again on fresh samples of 80). One is about individuals; the other is about averages. Same word, one level up.

Why care? Because your real life only ever gives you *one* sample. You taste the soup once. The sampling distribution tells you how far your one taste might reasonably sit from the truth — and that is the whole foundation of trusting a sample.

▪ WATCH IT WORK

Question: A tiny population has just three values: $2, 4, 6$. Its true mean is $\mu = 4$. We take samples of size $n = 2$ (with replacement) and compute $\bar{x}$ each time. What does the sampling distribution of $\bar{x}$ look like?

Step 1 — List the possible samples. With replacement, the equally likely samples of size 2 are: $(2,2), (2,4), (2,6), (4,2), (4,4), (4,6), (6,2), (6,4), (6,6)$ — nine in all.

Step 2 — Compute $\bar{x}$ for each. The averages are 2, 3, 4, 3, 4, 5, 4, 5, 6.

Step 3 — Tally the $\bar{x}$ values. $\bar{x} = 2$ once, 3 twice, 4 three times, 5 twice, 6 once.

Step 4 — Read the shape. The averages pile up in the *middle* (around 4) and thin out at the edges (2 and 6). The single value $\bar{x} = 4$ — the true population mean — is the most common one. The sample means cluster around the truth.

That's a sampling distribution, built by hand. The averages wobble, but they wobble *around* μ, and they pile up in a bell-ish heap.

▪ YOUR TURN

Stay with the population $2, 4, 6$ (true mean $\mu = 4$).

1. From the nine samples in the Watch above, what was the most common value of $\bar{x}$?
2. Was any single sample's $\bar{x}$ as low as 2 or as high as 6? Were those common or rare?
3. In one sentence: do the sample means tend to land near the true mean, or far from it?

Work space:

▪ CHECK YOUR VOICE

If a voice just said "wait, distributions of averages, this is where it gets too abstract for me" — pause and breathe. You did not need any new math to see those nine averages pile up around 4. You added two numbers and divided by two, nine times. The *idea* feels big, but you just built it with arithmetic you've had since middle school. Abstract is not the same as hard.

7.2 The Central Limit Theorem

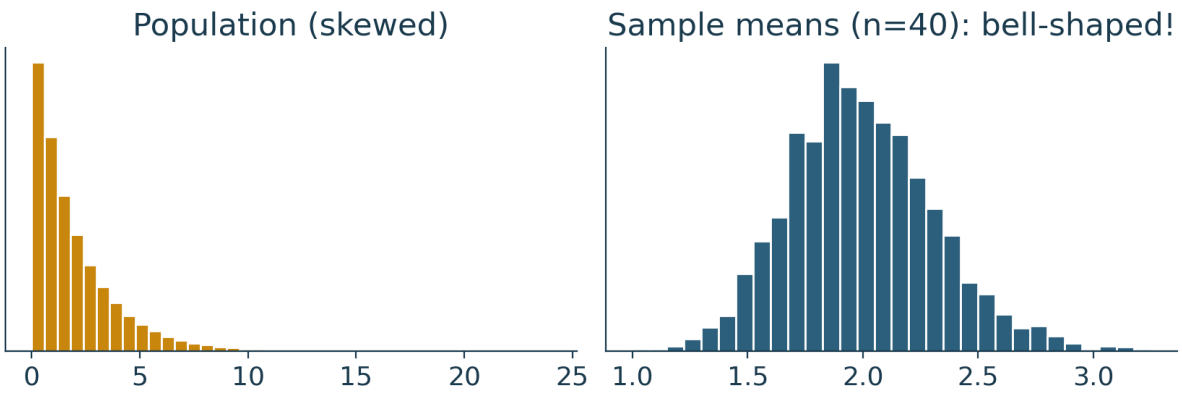


Figure 7.1. Even when the population is skewed (left), the distribution of sample means is bell-shaped (right).

▪ READ THIS FIRST

You've seen that sample means wobble around the truth and pile up in the middle. Now comes the part that feels almost too good to be true — so much so that mathematicians gave it a grand name.

Here it is in everyday terms: no matter how weird and lopsided the original population is — incomes that are skewed, soup that's wildly inconsistent, anything — the *averages* from large enough samples settle into a smooth, predictable bell shape, centered right on the truth, with a spread you can calculate exactly.

The mess of the world gets tamed by averaging. That promise has a name: the Central Limit Theorem. It's not a trap. It's the reason any of this works.

▪ LET'S TALK ABOUT IT

Picture a population that is *not* bell-shaped at all — say, the amount of money in everyone's checking account, where most people have a little and a few have a lot (heavily skewed right). If you repeatedly grabbed 50 random people and averaged their balances, do you think those *averages* would also be heavily skewed — or would they smooth out? Sit with your guess before reading on.

▪ NOW WE NAME IT

Take samples of size n from a population with mean μ and standard deviation σ. The sampling distribution of $\bar{x}$ has three properties:

1. Its center. The mean of the sampling distribution equals the population mean: $\mu_{\bar{x}} = \mu$. On average, the sample mean hits the truth. (We say $\bar{x}$ is *unbiased*.)

2. Its spread. The standard deviation of the sampling distribution is called the **standard error** of the mean:

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Notice the $\sqrt{n}$ on the bottom: bigger samples make the standard error *smaller*. More data means the average wobbles less. (That matches your gut — a bigger, better-stirred spoonful gives a steadier taste.)

3. Its shape — the Central Limit Theorem. If the sample size is large (a common rule of thumb is $n \geq 30$) *or* the population itself is normal, then the sampling distribution of $\bar{x}$ is approximately *normal* — a bell — *regardless of the population's shape*.

Read that last clause again, because it's the gift: the population can be skewed, spiky, or lumpy, and the distribution of *averages* still comes out bell-shaped, as long as n is big enough. That is why we can use the familiar normal-curve z-score machinery on sample means.

▪ WATCH IT WORK

Question: A coffee shop's drink sizes have a population mean of $\mu = 16$ ounces and standard deviation $\sigma = 3$ ounces (the machine isn't perfect). A manager pulls a random sample of $n = 36$ drinks. (a) What is the standard error of $\bar{x}$? (b) What is the probability the sample mean is more than 16.5 ounces?

Step 1 — Find the center and the standard error. The center is $\mu_{\bar{x}} = \mu = 16$. The standard error is

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = \frac{3}{6} = 0.5 \text{ ounces.}$$

So sample means of 36 drinks wobble around 16 with a typical wobble of about half an ounce.

Step 2 — Check the shape. Is $n \geq 30$? Yes, $n = 36$. By the Central Limit Theorem, $\bar{x}$ is approximately normal — so we may use a z-score.

Step 3 — Convert to a z-score. For a sample mean we use

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{16.5 - 16}{0.5} = \frac{0.5}{0.5} = 1.00.$$

The value 16.5 sits exactly one standard error above the truth.

Step 4 — Turn z into a probability. We want $P(\bar{x} > 16.5) = P(z > 1.00)$. From a normal table (or jamovi), the area below $z = 1.00$ is about 0.8413, so the area *above* is

$$1 - 0.8413 = 0.1587.$$

Step 5 — Say it in words. There's about a 15.9% chance that the average of 36 randomly chosen drinks exceeds 16.5 ounces. Not rare, not common — and now you can put a number on it.

▪ YOUR TURN

A bakery's loaves have population mean weight $\mu = 500$ grams and standard deviation $\sigma = 40$ grams. A random sample of $n = 64$ loaves is taken.

1. Compute the standard error, $\dfrac{\sigma}{\sqrt{n}}$.
2. Does the Central Limit Theorem let you treat $\bar{x}$ as approximately normal? Why?
3. Find the z-score for a sample mean of $\bar{x} = 510$ grams.
4. What is $P(\bar{x} > 510)$? (You'll want the area above your z.)

Work space:

▪ CHECK YOUR VOICE

If your inner critic insists the Central Limit Theorem is "one of those things only math people really get," notice the trick it's playing: it's confusing the fancy *name* with the difficulty of the *idea*. You just computed a standard error by dividing 3 by 6, and a z-score by dividing 0.5 by 0.5. The hardest arithmetic on this page is long division. You're not faking it — you're doing it.

7.3 The Sampling Distribution of a Proportion

▪ READ THIS FIRST

Sometimes the thing you're measuring isn't an amount — it's a yes-or-no. What fraction of voters support a measure? What fraction of chargers are defective? What fraction of customers come back?

Each sample gives you a *proportion* — "48% of the people we polled said yes." And just like the sample mean, that proportion wobbles from sample to sample. Poll a different 1,000 people and you might get 46% or 51%. Good news: this wobble follows the very same kind of predictable bell pattern.

▪ LET'S TALK ABOUT IT

Suppose the truth is that exactly half of all voters — 50% — support a measure. You poll 1,000 of them. Would you expect to get *exactly* 500 "yes" answers, or something close like 487 or 512? And if you polled only 10 people, would you trust "5 out of 10" as much as "500 out of 1,000"?

▪ NOW WE NAME IT

Let p be the true population proportion (the parameter) and $\hat{p}$ ("p-hat") be the sample proportion (the statistic) — the fraction of "yes" in your sample. The sampling distribution of $\hat{p}$ has three properties that mirror the mean's:

1. Its center. $\mu_{\hat{p}} = p$. On average, the sample proportion hits the true proportion.

2. Its spread. The standard error of a proportion is

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}}.$$

Again, the n is on the bottom, so larger samples shrink the wobble.

3. Its shape. The sampling distribution of $\hat{p}$ is approximately *normal* when the sample is large enough — specifically when

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10.$$

In words: you need at least about 10 expected "yes" *and* 10 expected "no." That's why "5 out of 10" is shaky and "500 out of 1,000" is solid.

▪ WATCH IT WORK

Question: Suppose the true proportion of voters supporting a measure is $p = 0.60$. A pollster surveys $n = 400$ voters. (a) Find the standard error of $\hat{p}$. (b) What's the probability the sample proportion is below 0.56?

Step 1 — Check the normal conditions. $np = 400(0.60) = 240 \geq 10$ and $n(1-p) = 400(0.40) = 160 \geq 10$. Both pass, so $\hat{p}$ is approximately normal.

Step 2 — Find the center and standard error. The center is $\mu_{\hat{p}} = p = 0.60$. The standard error is

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(0.40)}{400}} = \sqrt{\frac{0.24}{400}} = \sqrt{0.0006} \approx 0.02449.$$

Step 3 — Convert to a z-score. For a proportion,

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{0.56 - 0.60}{0.02449} = \frac{-0.04}{0.02449} \approx -1.633.$$

Step 4 — Turn z into a probability. We want $P(\hat{p} < 0.56) = P(z < -1.633)$. Rounding to $z < -1.63$ for a normal table (or using jamovi directly), that area is about 0.0516.

Step 5 — Say it in words. There's roughly a 5.2% chance that a poll of 400 voters shows support below 56%, even though the truth is 60%. A reasonably solid poll rarely strays that far — exactly why bigger polls feel more trustworthy.

▪ YOUR TURN

A factory's true defect rate is $p = 0.05$. An inspector checks a random sample of $n = 500$ items.

1. Check the two conditions $np \geq 10$ and $n(1-p) \geq 10$. Do they pass?
2. Compute the standard error, $\sqrt{p(1-p)/n}$ (round to four decimals).
3. Find the z-score for a sample proportion of $\hat{p} = 0.07$.
4. Roughly, is a sample defect rate of 7% a little unusual or very common? (You don't need an exact probability — just read your z.)

Work space:

▪ CHECK YOUR VOICE

Notice that the proportion section was the *same four moves* as the mean section: check the shape, find the standard error, compute a z, read a probability. If it felt easier the second time, that's not because it got simpler — it's because you already own the pattern. You're not relearning; you're recognizing. That's what fluency feels like.

7.4 Why This Is the Whole Ballgame

▪ READ THIS FIRST

Step back and look at what you now hold. You can take a single sample, compute one statistic, and say — with a real number attached — how far that statistic might sit from the truth. That little number, the standard error, is the most important quantity in the rest of this book. Everything ahead is built on it. So let's name why.

▪ LET'S TALK ABOUT IT

If someone told you "the average wait time in our sample was 12 minutes," what's the very next question a careful person should ask? (Hint: it's not "what's the average?" — you were just told that. It's something about *how sure* we are.)

▪ NOW WE NAME IT

The standard error is the bridge from a sample to a claim about the truth. Two huge ideas in the chapters ahead run entirely on it:

Confidence intervals (Part IV, Chapter 8). Instead of reporting a single guess like "$\bar{x} = 12$ minutes," we report a *range*: "the true average is somewhere around 12, give or take a couple of standard errors." The width of that range *is* the standard error, scaled up. A small standard error gives a tight, confident interval; a large one gives a wide, cautious one.

Hypothesis tests (Part V). When we ask "could the truth really be 10 minutes, or is our sample of 12 too far off for that to be believable?", we measure the distance in standard errors — exactly the z-score move you just practiced. A result many standard errors from the claim is hard to explain by chance.

So the standard error answers the question every honest data statement needs: *how much does this sample statistic naturally bounce around the truth?* Measure the bounce, and you can finally say how much to trust the spoonful.

▪ WATCH IT WORK

Question: A sample of $n = 100$ commute times has standard error $\text{SE} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{10}{\sqrt{100}} = 1$ minute, with sample mean $\bar{x} = 12$ minutes. Without the formal machinery of Chapter 8, sketch what a confidence statement would look like.

Step 1 — Anchor on the estimate. Our single best guess for the true average commute is $\bar{x} = 12$ minutes.

Step 2 — Attach the wobble. The standard error is 1 minute — that's how much $\bar{x}$ typically strays from the truth.

Step 3 — Build a rough range. Reaching about two standard errors on each side gives roughly 12 ± 2, i.e. from 10 to 14 minutes.

Step 4 — Say it in words. "We estimate the true average commute is about 12 minutes, and we're fairly confident it lies between 10 and 14." That sentence — estimate plus a standard-error-sized cushion — is the entire idea of a confidence interval. Chapter 8 just makes the cushion precise.

▪ YOUR TURN

A sample of $n = 49$ customer ratings has population standard deviation $\sigma = 14$ and sample mean $\bar{x} = 70$.

1. Compute the standard error, $\dfrac{\sigma}{\sqrt{n}}$.
2. Build a rough "estimate, give or take two standard errors" range around $\bar{x} = 70$.
3. In one sentence, explain why a larger sample would make that range *narrower*.

Work space:

▪ CHECK YOUR VOICE

If part of you is bracing for confidence intervals and hypothesis tests to be where you finally fall behind — hear this. You just *built* a confidence interval, by hand, in four plain steps, before we even named it formally. The hard engine of inference is the thing you spent this whole chapter learning. The chapters ahead are mostly putting a polished frame around the work you can already do.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 7 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 7 — Your Turn Answers

Math Skills: (1) $\sqrt{64} = 8$; $\sqrt{50} \approx 7.071$. (2) $\dfrac{18}{\sqrt{36}} = \dfrac{18}{6} = 3$. (3) $\dfrac{30}{\sqrt{25}} = \dfrac{30}{5} = 6$.

Section 7.1: (1) $\bar{x} = 4$ was most common (it occurred three times). (2) $\bar{x} = 2$ happened once and $\bar{x} = 6$ once — both rare (one time each out of nine). (3) The sample means tend to land *near* the true mean of 4, piling up in the middle.

Section 7.2: (1) SE $= \dfrac{40}{\sqrt{64}} = \dfrac{40}{8} = 5$ grams. (2) Yes — $n = 64 \geq 30$, so by the Central Limit Theorem $\bar{x}$ is approximately normal regardless of the population's shape. (3) $z = \dfrac{510 - 500}{5} = \dfrac{10}{5} = 2.00$. (4) $P(\bar{x} > 510) = P(z > 2.00) = 1 - 0.9772 = 0.0228$, about 2.3%.

Section 7.3: (1) $np = 500(0.05) = 25 \geq 10$ and $n(1-p) = 500(0.95) = 475 \geq 10$ — both pass. (2) SE $= \sqrt{\dfrac{0.05(0.95)}{500}} = \sqrt{\dfrac{0.0475}{500}} = \sqrt{0.000095} \approx 0.0097$. (3) $z = \dfrac{0.07 - 0.05}{0.0097} = \dfrac{0.02}{0.0097} \approx 2.05$. (4) A z of about 2.05 is a bit over two standard errors above the truth — a little unusual, not extreme (it happens roughly 2% of the time).

Section 7.4: (1) SE $= \dfrac{14}{\sqrt{49}} = \dfrac{14}{7} = 2$. (2) Two standard errors is 4, so the rough range is 70 ± 4, from 66 to 74. (3) A larger n makes $\sqrt{n}$ bigger, which makes the standard error $\dfrac{\sigma}{\sqrt{n}}$ smaller — so the give-or-take cushion shrinks and the range gets narrower.

Part III Checkpoint — Study Guide & Practice (Chapters 5–7)

This Study Guide pulls together everything from Chapters 5–7 — chance, distributions, and the wobble of a sample — and lets you rehearse before an exam the way the exam will actually feel.

STUDY GUIDE & PRACTICE

You just learned the machinery that the rest of statistics runs on: how to count chances, how randomness takes a *shape*, and why one little sample can speak for a whole population. This isn't a test — it's a dress rehearsal. We'll gather the big ideas in one place, check the vocabulary, and then run twelve mixed problems that jump across all three chapters, exactly the way a real exam does. When your inner voice says "I forgot all of this," keep going anyway — the answers are right here, fully worked, waiting for you.

The Big Ideas, in One Place

Chapter 5 — The Logic of Probability.

- A **probability** is a number from 0 to 1: $0 \le P(A) \le 1$. (0 means impossible, 1 means certain.) The probabilities of all outcomes in the sample space add to exactly 1.
- The **complement rule**: the chance an event does *not* happen is $P(\text{not } A) = 1 - P(A)$.
- The **addition rule** ("either/or"): $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. Subtract the overlap so nothing is counted twice. If the events are **mutually exclusive** (can't both happen), the overlap is 0 and it's just $P(A) + P(B)$.
- The **conditional probability** of A given B is $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$ — zoom into the B world, then ask what fraction also have A.
- The **multiplication rule** ("this and that"): $P(A \text{ and } B) = P(A)\,P(B \mid A)$.
- Two events are **independent** when knowing one doesn't change the other ($P(A \mid B) = P(A)$). Then the rule simplifies to $P(A \text{ and } B) = P(A)\,P(B)$.

- A **two-way table** gives all three at a glance. The whole trick is the *denominator*: marginal and joint probabilities divide by the *grand total*; a conditional probability divides by *just the row or column you're given*.

Chapter 6 — Predictable Randomness.

- A **discrete probability distribution** lists each value x with its probability $P(x)$; every $P(x)$ is in $[0, 1]$ and they sum to 1. Its **expected value** (mean) is $\mu = \sum x\,P(x)$, and its **standard deviation** is $\sigma = \sqrt{\sum (x-\mu)^2\,P(x)}$.

- A **binomial distribution** counts successes in a fixed number of yes/no trials. Conditions (**B-I-N-S**): **B**inary, **I**ndependent, **N**umber of trials fixed, **S**ame probability p. The chance of exactly k successes is

$$P(X = k) = \binom{n}{k} p^k\,(1-p)^{n-k}.$$

  Its mean is $\mu = np$ and its standard deviation is $\sigma = \sqrt{np(1-p)}$.

- A **normal distribution** is the symmetric bell curve, fixed by its mean μ (center) and standard deviation σ (spread). The **Empirical Rule** (68–95–99.7): about 68% of data lies within 1σ of the mean, about 95% within 2σ, and about 99.7% within 3σ. Symmetry splits each leftover tail evenly.

- A **z-score** places any value on the standard bell curve: $z = \dfrac{x-\mu}{\sigma}$. Positive is above average, negative is below, 0 is exactly average. The area to its left is the *percentile*. Go backward with $x = \mu + z\sigma$.

Chapter 7 — From Sample to Truth.

- A **sampling distribution** is the pattern of a statistic (like $\bar{x}$) across all possible samples of size n — the shape of the wobble. It centers on the truth: $\mu_{\bar{x}} = \mu$.

- The **standard error** of the mean is $\text{SE} = \dfrac{\sigma}{\sqrt{n}}$. The $\sqrt{n}$ on the bottom means bigger samples wobble *less*.

- The **Central Limit Theorem**: if n is large (rule of thumb $n \geq 30$) *or* the population is normal, the sampling distribution of $\bar{x}$ is approximately normal — no matter the population's shape.

- A probability about a sample mean uses $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, then read the area off the normal curve.

- For a **sample proportion** $\hat{p}$: center $\mu_{\hat{p}} = p$, standard error SE $= \sqrt{\dfrac{p(1-p)}{n}}$, and the shape is approximately normal when $np \geq 10$ *and* $n(1-p) \geq 10$ (at least about 10 expected "yes" and 10 expected "no").

Vocabulary Check

Can you say each of these in one plain sentence? If one feels fuzzy, that's exactly the one to reread — not a sign you're behind.

- sample space, outcome, event, probability
- complement rule; addition rule; mutually exclusive (disjoint)
- conditional probability; multiplication rule; independent events
- two-way table; marginal, joint, and conditional probability
- random variable; probability distribution; expected value, standard deviation
- binomial distribution; the B-I-N-S conditions; $\mu = np$, $\sigma = \sqrt{np(1-p)}$
- normal distribution; Empirical Rule (68–95–99.7); z-score; percentile
- sampling distribution; standard error; Central Limit Theorem
- sample proportion $\hat{p}$; the $np \geq 10$ and $n(1-p) \geq 10$ conditions

Mixed Practice

These twelve problems jump around on purpose — just like an exam. Try each one before peeking. The full worked answers come right after.

1. The chance it rains tomorrow is $P(\text{rain}) = 0.30$. (a) What is the chance it does *not* rain? (b) Which rule did you use?
2. You draw one card from a standard 52-card deck. What is the probability it is a king *or* a heart? (Hint: there are 4 kings, 13 hearts, and watch the overlap.)
3. At a clinic, $P(\text{flu}) = 0.20$ and $P(\text{flu and fever}) = 0.16$. What is $P(\text{fever} \mid \text{flu})$, the probability a flu patient also has a fever?

4. A group of 200 people was surveyed on whether they own a pet and whether they rent or own their home.

	Owns a Pet	No Pet	Total
Renter	40	60	100
Homeowner	70	30	100
Total	110	90	200

(a) What is $P(\text{owns a pet})$? (b) What is $P(\text{owns a pet} \mid \text{homeowner})$? (c) Are "owns a pet" and "homeowner" independent? Explain.

5. A small game has payouts X with this distribution: $P(0) = 0.1$, $P(1) = 0.3$, $P(2) = 0.4$, $P(3) = 0.2$. (a) Confirm it's a valid distribution. (b) Find the expected value μ.

6. A student guesses on $n = 8$ true/false questions, each with success probability $p = 0.5$, independently. (a) Find the probability of getting exactly 5 correct. (b) Find the mean μ and standard deviation σ of the number correct.

7. Adult systolic blood pressure is roughly normal with mean $\mu = 120$ and standard deviation $\sigma = 10$. Using the Empirical Rule, about what percent of adults have a reading between 100 and 140?

8. On the same blood-pressure distribution ($\mu = 120$, $\sigma = 10$), a patient reads 135. (a) Find the z-score. (b) Say in words what that z-score means.

9. A machine fills bags with population mean weight $\mu = 60$ grams and standard deviation $\sigma = 12$ grams. A random sample of $n = 36$ bags is taken. (a) Compute the standard error of $\bar{x}$. (b) Does the Central Limit Theorem let you treat $\bar{x}$ as approximately normal? Why?

10. Using the bag-filling setup from Problem 9 ($\mu = 60$, $\sigma = 12$, $n = 36$), find the probability that the sample mean weight exceeds 63 grams. (The area below $z = 1.50$ is about 0.9332.)

11. A city believes the true proportion of residents who recycle is $p = 0.40$. A survey of $n = 600$ residents is planned. (a) Check the conditions $np \geq 10$ and $n(1-p) \geq 10$. (b) Compute the standard error of $\hat{p}$.

12. Two events A and B are independent, with $P(A) = 0.5$ and $P(B) = 0.2$. (a) Find $P(A \text{ and } B)$. (b) Find $P(A \text{ or } B)$ using the addition rule.

Answers

Worked out in full — read these even for the ones you got right, because the *reasoning* is what the exam rewards.

1. (a) $P(\text{no rain}) = 1 - P(\text{rain}) = 1 - 0.30 = 0.70$. (b) The *complement rule*: the chance something doesn't happen is 1 minus the chance it does.

2. Use the addition rule and subtract the overlap. $P(\text{king}) = \dfrac{4}{52}$, $P(\text{heart}) = \dfrac{13}{52}$, and the overlap is the *king of hearts*, one card, so $P(\text{king and heart}) = \dfrac{1}{52}$.

$$P(\text{king or heart}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} \approx 0.308.$$

   Without subtracting the king of hearts you'd wrongly get $\dfrac{17}{52}$, double-counting that one card.

3. Use the conditional-probability formula:

$$P(\text{fever} \mid \text{flu}) = \frac{P(\text{flu and fever})}{P(\text{flu})} = \frac{0.16}{0.20} = 0.80.$$

   So 80% of flu patients also have a fever. (Notice the answer lands between 0 and 1, as every probability must.)

4. (a) Marginal: use the column total over the grand total, $P(\text{owns a pet}) = \dfrac{110}{200} = 0.55$. (b) Conditional: zoom into the homeowner row (100 people), $P(\text{owns a pet} \mid \text{homeowner}) = \dfrac{70}{100} = 0.70$ — denominator 100, not 200. (c) *Not independent*: $P(\text{owns a pet}) = 0.55$ but $P(\text{owns a pet} \mid \text{homeowner}) = 0.70$. Since these differ, knowing someone is a homeowner changes the chance they own a pet, so the two are not independent.

5. (a) $0.1 + 0.3 + 0.4 + 0.2 = 1.0$, and every probability is between 0 and 1 — valid. (b) $\mu = \sum x\,P(x)$:

$$\mu = (0)(0.1) + (1)(0.3) + (2)(0.4) + (3)(0.2) = 0 + 0.3 + 0.8 + 0.6 = 1.7.$$

   The expected payout is 1.7.

6. This is binomial with $n = 8$, $p = 0.5$, $k = 5$. (a)

$$P(X = 5) = \binom{8}{5}(0.5)^5(0.5)^{8-5} = \binom{8}{5}(0.5)^5(0.5)^3.$$

The combination is $\binom{8}{5} = \dfrac{8!}{5!\,3!} = \dfrac{40320}{120 \times 6} = \dfrac{40320}{720} = 56$. The powers: $(0.5)^5 = 0.03125$ and $(0.5)^3 = 0.125$. So

$$P(X = 5) = 56 \times 0.03125 \times 0.125 = 56 \times 0.00390625 = 0.21875 \approx 21.9\%.$$

(b) $\mu = np = 8(0.5) = 4$ questions correct on average; $\sigma = \sqrt{np(1-p)} = \sqrt{8(0.5)(0.5)} = \sqrt{2} \approx 1.41$.

7. The interval 100 to 140 runs from $120 - 2(10) = 100$ to $120 + 2(10) = 140$ — exactly $\mu \pm 2\sigma$. By the Empirical Rule, about 95% of adults fall within 2 standard deviations of the mean, so about $\mathbf{95}\%$ read between 100 and 140.

8. (a) $z = \dfrac{x - \mu}{\sigma} = \dfrac{135 - 120}{10} = \dfrac{15}{10} = 1.5$. (b) A reading of 135 sits 1.5 standard deviations *above* the mean — higher than average, but not extreme.

9. (a) SE $= \dfrac{\sigma}{\sqrt{n}} = \dfrac{12}{\sqrt{36}} = \dfrac{12}{6} = 2$ grams. (b) Yes — $n = 36 \geq 30$, so by the Central Limit Theorem the sampling distribution of $\bar{x}$ is approximately normal regardless of the population's shape.

10. Use the sample-mean z-score with SE $= 2$ from Problem 9:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{63 - 60}{2} = \frac{3}{2} = 1.50.$$

We want $P(\bar{x} > 63) = P(z > 1.50)$. The area below $z = 1.50$ is about 0.9332, so the area above is

$$1 - 0.9332 = 0.0668.$$

There's about a 6.7% chance the sample mean exceeds 63 grams.

11. (a) $np = 600(0.40) = 240 \geq 10$ and $n(1-p) = 600(0.60) = 360 \geq 10$ — both pass, so $\hat{p}$ is approximately normal. (b)

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.40(0.60)}{600}} = \sqrt{\frac{0.24}{600}} = \sqrt{0.0004} = 0.02.$$

The standard error of the sample proportion is 0.02.

12. (a) Since A and B are independent, $P(A \text{ and } B) = P(A)\,P(B) = (0.5)(0.2) = 0.10$. (b)

The addition rule still applies, and here we know the overlap from part (a):

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.5 + 0.2 - 0.10 = 0.60.$$

(Independent is not the same as mutually exclusive — here the overlap is 0.10, not 0, so we still subtract it.)

You just rehearsed every big idea from Part III — complements and the addition rule, conditional probability and independence, reading a two-way table, expected value, the binomial formula with its mean and SD, the Empirical Rule, z-scores, standard error, and the Central Limit Theorem. If some were shaky, now you know exactly where to look. That's not failing the rehearsal; that's what the rehearsal is *for*. You're ready.

Part III — Formula Sheet: Probability & Distributions

Every formula from Part III on one page, each symbol named in plain words with a note on when *to use it.*

FORMULA SHEET

The symbols, in plain words:

- $P(A)$ = the probability of event A n = number of trials or sample size k = number of successes
- p = probability of success on one trial (or a population proportion) μ = mean σ = standard deviation
- $\bar{x}$ = sample mean $\hat{p}$ = sample proportion z = standard score

Probability rules (Chapter 5)

$$0 \le P(A) \le 1 \qquad P(\text{not } A) = 1 - P(A)$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad \text{(disjoint: the last term is } 0)$$

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \qquad P(A \text{ and } B) = P(A)\,P(B \mid A) \quad \text{(independent: } = P(A)\,P(B))$$

When: complement for "at least one"; addition for "or"; multiplication for "and"; conditional when one event's outcome is already known.

Discrete probability distribution (Chapter 6)

$$\mu = \sum x\,P(x) \qquad \sigma = \sqrt{\sum (x-\mu)^2\,P(x)}$$

When: a variable takes listed values with listed probabilities (the probabilities must sum to 1). μ is the long-run average (expected value).

Binomial distribution (Chapter 6) — use when there are a fixed number n of independent

trials, each a success/failure with the same probability p.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

Normal distribution (Chapter 6)

$$z = \frac{x - \mu}{\sigma} \qquad\qquad \textbf{Empirical Rule: } \approx 68\%\ (1\sigma),\ 95\%\ (2\sigma),\ 99.7\%\ (3\sigma)$$

When: for bell-shaped data; the z-score says how many standard deviations a value sits from the mean, and unlocks normal probabilities (areas).

Sampling distributions — the Central Limit Theorem (Chapter 7)

$$\text{Sample mean } \bar{x}: \quad \text{center } \mu, \quad \text{standard error } \frac{\sigma}{\sqrt{n}}, \quad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\text{Sample proportion } \hat{p}: \quad \text{center } p, \quad \text{standard error } \sqrt{\frac{p(1-p)}{n}}$$

When: the sampling distribution of $\bar{x}$ is approximately normal if $n \geq 30$ *or* the population is normal (CLT). For $\hat{p}$, approximately normal when $np \geq 10$ and $n(1-p) \geq 10$. The **standard error** is how much a sample statistic naturally bounces around the truth — the engine behind Parts IV and V.

Part IV

Estimating with Confidence

Chapter 8

How Sure Are We?

Confidence Intervals — Estimating a Population Truth From One Sample

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll be able to take a single sample — one survey, one batch, one set of measurements — and report not just a single guess about the whole population, but a *range* with an honest label on how trustworthy it is. You'll stop saying "the average is 7.5" and start saying "we're 95% confident the true average is between 6.7 and 8.3."

The Story. You almost never get to measure everyone. You measure a slice (a sample) and use it to estimate the whole (the population). But a single number from a slice is never exactly right. So instead of pretending it is, statistics does something honest: it reports a number *plus a cushion*. That cushion is the margin of error, and the number-plus-cushion is a confidence interval.

The Skills You'll Have:

- Build a confidence interval as "point estimate $\pm$ margin of error"
- Say what "95% confident" actually means (and dodge the #1 misconception)
- Find a confidence interval for a mean, a proportion, and a standard deviation
- Explain why a bigger sample gives a tighter, more useful interval

The Confidence Moment. When a news story says "52% support it, margin of error ± 3 points," your brain will calmly translate: "so the real number is probably somewhere from 49% to 55% — they measured a slice, not everyone, and they're being honest about it."

The Bridge. A confidence interval asks, "what range of truths is plausible?" Part V flips the question to, "is one *specific* claim plausible?" — that's hypothesis testing. The two are the same machinery seen from two angles, so everything you build here makes the next part feel familiar.

MATH SKILLS YOU'LL NEED

This chapter asks you to look up one number in a table, take a square root, plug numbers into a formula, and round. That's genuinely it. Let's warm up each move.

Skill 1 — Reading a critical value. A confidence interval needs a "how wide" number called a *critical value* (t^* or z^*). For now, just know you *look it up*: pick your confidence level (say 95%) and, for t, your degrees of freedom. You'll often use $z^* = 1.96$ for 95%. Reading a value off a table is not a calculation — it's finding a row and a column, exactly like Chapter 1.

Skill 2 — Square roots. A square root undoes squaring: $\sqrt{0.0064} = 0.08$ because $0.08 \times 0.08 = 0.0064$. Your calculator has a $\sqrt{\ }$ button — that's all you need.

Skill 3 — Substituting into a formula and rounding. "Substitute" means swap each letter for its number, then do the arithmetic. For example, in $\frac{s}{\sqrt{n}}$ with $s = 1.2$ and $n = 12$: $\frac{1.2}{\sqrt{12}} = \frac{1.2}{3.464} = 0.346$. Round at the *end*, not in the middle.

Practice (answers at the end of the chapter):

1. For a 95% interval with degrees of freedom $df = 11$, the critical value is $t^* = 2.201$. Is t^* bigger or smaller than the $z^* = 1.96$ used when σ is known?
2. Find $\sqrt{0.0009}$.
3. Substitute into $z^* \frac{s}{\sqrt{n}}$ with $z^* = 1.96$, $s = 10$, $n = 100$. Round to one decimal place.

8.1 The Big Idea: Point Estimate + Margin of Error

▪ READ THIS FIRST

A friend texts: "Be there in 20 minutes." You don't actually believe they'll arrive at exactly minute 20. You hear it as "somewhere around 20 — give or take." If they're usually punctual, you mentally pencil in "18 to 22." If they're famously late, you widen it to "20 to 40."

You just built a confidence interval. The 20 is your best single guess. The "give or take" is your margin of error. And how wide you make it depends on how much you trust the estimate.

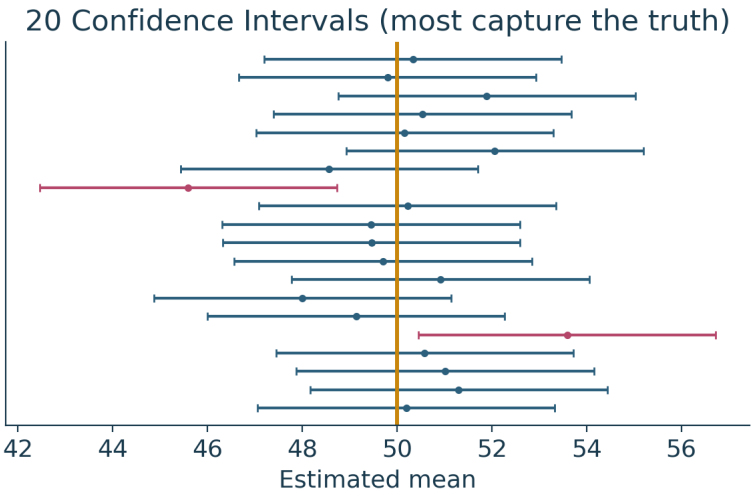


Figure 8.1. Twenty 95% confidence intervals for the same true mean (gold line). Most capture it; a few miss — that is what "95% confident" means.

▪ LET'S TALK ABOUT IT

Think of a time you gave someone an estimate — a price, an arrival time, how long a task would take. Did you give one exact number, or did you instinctively add "ish" or "give or take"? What made you widen the range versus narrow it?

▪ NOW WE NAME IT

A **point estimate** is your single best guess for a population value, calculated from your sample — for example, the sample mean $\bar{x}$ as a guess for the true population mean.

A **margin of error** is the "give or take" cushion you add and subtract around the point estimate.

A **confidence interval** puts them together:

$$\text{point estimate} \pm \text{margin of error}$$

The **confidence level** (usually 90%, 95%, or 99%) is the success rate of the *method*. Here is the part everyone gets wrong, so read it twice:

"95% confident" does NOT mean "there's a 95% probability the true value is in *this* interval." The true value is a fixed number — it's either in your interval or it isn't. What's really true is this: if you repeated the whole process many times (new sample each time), about 95% of the intervals you'd build would capture the true value. You happen to have one of them, and you're betting it's one of the good 95%.

▪ WATCH IT WORK

Question: A poll estimates 52% of voters support a measure, with a margin of error of 3 percentage points at the 95% confidence level. State the interval and interpret it correctly.

Step 1 — Identify the point estimate. The single best guess is 52%.

Step 2 — Apply the margin of error. $52\% \pm 3\%$ gives the interval from 49% to 55%.

Step 3 — Interpret it the honest way. "We are 95% confident that the true percentage of *all* voters who support the measure is between 49% and 55%." Meaning: the method that produced this range captures the truth about 95% of the time.

Step 4 — Say what it does NOT mean. It does *not* mean "there's a 95% chance the truth is between 49 and 55." The truth is one fixed number; it's our *interval* that varies from sample to sample.

Two more facts to lock in: a *higher* confidence level (99% vs. 90%) makes the interval *wider* (more cushion to be more sure), and a *larger sample* n makes it *narrower* (more information, less guessing).

▪ YOUR TURN

A study reports that the average daily screen time in a sample is 4.2 hours, with a margin of error of 0.5 hours at 95% confidence.

1. What is the point estimate?
2. Write out the confidence interval.
3. Fill in the blank correctly: "We are 95% confident that _____."
4. True or false: "There is a 95% probability the true mean is inside this exact interval."

Work space:

▪ CHECK YOUR VOICE

If the "it's not a 95% probability for THIS interval" part felt slippery — that's normal, and it does not mean you're bad at this. It is the single most misunderstood idea in all of intro statistics, and professionals slip on it too. You're not confused because you're behind; you're wrestling with it because you're actually paying attention. Talk back: "I noticed the subtle part, which is exactly what a careful thinker does."

8.2 Confidence Interval for a Mean

▪ READ THIS FIRST

You want to know the average number of hours students at your college sleep. You can't ask all 9,000 of them, so you ask 12 and get an average of 7.5 hours. Fine — but 7.5 is just *your 12 people*. Ask a different 12 and you'd get a slightly different number.

So you don't report 7.5 alone. You report 7.5 plus a cushion. And because you measured the spread *from your sample* (you didn't know the true spread ahead of time), you use a slightly more forgiving tool than the normal curve: the t-distribution.

▪ LET'S TALK ABOUT IT

Why should "I asked 12 people" come with a bigger cushion than "I asked 1,200 people"? Sit with the intuition before any formula: which sample would you personally trust more, and why?

▪ NOW WE NAME IT

When you want a confidence interval for a population mean μ and you do *not* know the population standard deviation σ — which is almost always the case — you use the sample standard deviation s and the **t-distribution**:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

The piece $\frac{s}{\sqrt{n}}$ is the **standard error** (the typical wobble of a sample mean), and t^* is the critical value you look up for your confidence level using **degrees of freedom** $df = n - 1$.

The t-distribution looks like the normal curve but with slightly heavier tails — its way of admitting "we estimated the spread, so let's be a touch more cautious." With small samples it's noticeably wider; with large samples it's nearly identical to the normal curve.

The rare case (σ known). If you somehow *did* know the true population standard deviation σ, you'd swap in the normal curve and use $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$. In practice you rarely know σ, so the t version above is your everyday tool.

▪ WATCH IT WORK

Question: A sample of $n = 12$ students sleeps an average of $\bar{x} = 7.5$ hours with sample standard deviation $s = 1.2$ hours. Build a 95% confidence interval for the true mean sleep μ.

Step 1 — Degrees of freedom and critical value. $df = n - 1 = 12 - 1 = 11$. Looking up the 95% value with $df = 11$ gives $t^* = 2.201$.

Step 2 — Standard error. $\frac{s}{\sqrt{n}} = \frac{1.2}{\sqrt{12}} = \frac{1.2}{3.464} = 0.346$ hours.

Step 3 — Margin of error. $t^* \frac{s}{\sqrt{n}} = 2.201 \times 0.346 = 0.762$ hours.

Step 4 — Assemble the interval. 7.5 ± 0.762, which runs from $7.5 - 0.762 = 6.74$ to $7.5 + 0.762 = 8.26$.

Step 5 — Interpret. "We are 95% confident that the true average sleep for all students is between 6.74 and 8.26 hours." Notice the interval is fairly wide — that's the price of asking only 12 people. Ask 120 and the cushion shrinks.

▪ YOUR TURN

A sample of $n = 16$ commuters has an average commute of $\bar{x} = 28$ minutes with $s = 8$ minutes. Build a 95% confidence interval. (The critical value is $t^* = 2.131$ for $df = 15$.)

1. What is df?
2. Compute the standard error $\frac{s}{\sqrt{n}}$ (round to two decimals).
3. Compute the margin of error and write the interval.
4. Interpret it in one sentence.

Work space:

▪ CHECK YOUR VOICE

If you got a little lost in the chain — look up t^*, divide for the standard error, multiply for the margin, add and subtract — that's four steps, and four-step problems feel heavy until they don't. The trick is that it's the *same* four steps every single time. You're not memorizing a new dance for each problem; you're rehearsing one dance. Say it: "I can follow a recipe, and this is just a recipe."

8.3 Confidence Interval for a Proportion

▪ READ THIS FIRST

Sometimes the thing you're estimating isn't an amount — it's a *percentage*. What fraction of customers would recommend the app? What share of voters support the measure? You don't have a mean here; you have a yes-or-no answer counted up into a proportion.

The idea is identical: take your sample proportion, add and subtract a cushion. Only the formula for the cushion changes, because percentages wobble differently than averages.

▪ LET'S TALK ABOUT IT

A headline says "68% of people approve." What's the one question you should always ask before trusting that number? (Hint: it's the same question from the soup pot in Chapter 1 — how big was the spoonful?)

▪ NOW WE NAME IT

When you're estimating a population proportion p (a percentage), the confidence interval is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Here $\hat{p}$ ("p-hat") is your **sample proportion** — the number of "yes" answers divided by n — and the square-root piece is the standard error for a proportion. Because proportions don't need a separately estimated spread, we use the normal curve's critical value z^*, no degrees of freedom required.

The three z^* values you'll use constantly:

Confidence level	z^*
90%	1.645
95%	1.960
99%	2.576

Quick sanity check before you start: this method assumes you have at least a handful of "yes" *and* "no" outcomes (a common rule is at least 10 of each), so the normal curve is a fair approximation.

▪ WATCH IT WORK

Question: In a sample of $n = 500$ customers, 180 say they would recommend a product. Build a 95% confidence interval for the true proportion p who would recommend it.

Step 1 — Sample proportion. $\hat{p} = \dfrac{180}{500} = 0.36$. (So $1 - \hat{p} = 0.64$.)

Step 2 — Critical value. For 95% confidence, $z^* = 1.96$.

Step 3 — Standard error. $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{0.36 \times 0.64}{500}} = \sqrt{\dfrac{0.2304}{500}} = \sqrt{0.0004608} = 0.0215.$

Step 4 — Margin of error. $z^* \times 0.0215 = 1.96 \times 0.0215 = 0.042.$

Step 5 — Assemble and interpret. 0.36 ± 0.042 runs from 0.318 to 0.402. "We are 95% confident that the true proportion of all customers who would recommend the product is between 31.8% and 40.2%."

▪ YOUR TURN

In a sample of $n = 400$ voters, 220 support a measure. Build a 95% confidence interval for the true proportion p. (Use $z^* = 1.96$.)

1. Find $\hat{p}$.
2. Compute the standard error $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (round to four decimals).
3. Compute the margin of error and write the interval.
4. Interpret it, and convert the endpoints to percentages.

Work space:

▪ CHECK YOUR VOICE

Notice that this section's formula *looked* scarier — a square root with a fraction inside — but the steps were the same shape as before: get your point estimate, get a cushion, add and subtract. If the square root made your stomach drop for a second, that's a learned flinch, not a verdict on your ability. You pressed one button on a calculator and it cooperated. Say it: "Big-looking formulas are still just buttons in order."

8.4 Confidence Interval for a Variance or Standard Deviation

▪ READ THIS FIRST

Usually we care about the *average* — but sometimes the *spread* is the whole point. A coffee machine that pours 12 ounces *on average* is useless if it swings wildly between 8 and 16. A factory cares about consistency. So we sometimes want to estimate the population's standard deviation, with a cushion around it, just like everything else.

The math here leans on a curve called chi-square. Don't let the symbol scare you — you'll look up two numbers and plug into a formula. That's the entire job.

▪ LET'S TALK ABOUT IT

Can you think of a situation where you'd care more about *consistency* than about the average? (Think: a bus that comes "on average every 10 minutes" but sometimes makes you wait 25.) Why might the spread matter more than the center there?

▪ NOW WE NAME IT

To estimate a population variance σ^2 (the squared spread), we use the **chi-square distribution**, written χ^2. Unlike the symmetric t and z curves, chi-square is lopsided, so we look up *two different* critical values — a left one and a right one — using $df = n - 1$.

The confidence interval for the variance σ^2 is:

$$\left(\frac{(n-1)s^2}{\chi^2_{right}} \ , \ \frac{(n-1)s^2}{\chi^2_{left}} \right)$$

Two things that trip people up, made simple: (1) the bigger critical value (χ^2_{right}) goes in the *bottom* of the *left* endpoint — dividing by a bigger number gives the smaller end. (2) To get an interval for the **standard deviation** σ instead of the variance, just take the square root of both endpoints.

A note on reading the width: intervals for spread are often quite wide, especially with small samples. That's honest, not broken — estimating variability well takes a lot of data.

▪ WATCH IT WORK

Question: A sample of $n = 10$ fills from a coffee machine has standard deviation $s = 4$ ounces. Build a 95% confidence interval for the population standard deviation σ.

Step 1 — Setup. $df = n - 1 = 9$, and $s^2 = 4^2 = 16$, so $(n-1)s^2 = 9 \times 16 = 144$.

Step 2 — Look up the two critical values. For 95% confidence with $df = 9$: $\chi^2_{right} = 19.023$ and $\chi^2_{left} = 2.700$.

Step 3 — Interval for the variance σ^2.

$$\left(\frac{144}{19.023} \ , \ \frac{144}{2.700} \right) = (7.57 \ , \ 53.33).$$

Step 4 — Take square roots to get σ. $\sqrt{7.57} = 2.75$ and $\sqrt{53.33} = 7.30$.

Step 5 — Interpret. "We are 95% confident that the true standard deviation of the machine's fills is between 2.75 and 7.30 ounces." That's a wide range — the honest signal that 10 fills isn't much to judge consistency on. More data would tighten it.

▪ YOUR TURN

A sample of $n = 10$ batteries has a standard deviation of $s = 3$ hours of life. Build a 95% confidence interval for σ. (Use $df = 9$, $\chi^2_{right} = 19.023$, $\chi^2_{left} = 2.700$.)

1. Compute $(n-1)s^2$.
2. Find the interval for the variance σ^2 (round to two decimals).
3. Take square roots to get the interval for σ.
4. In one sentence, is this interval wide or narrow, and what would shrink it?

Work space:

▪ CHECK YOUR VOICE

This is usually the section that feels like "okay, now it's getting hard." But look at what you actually did: you squared a number, multiplied by $n-1$, looked up two values, and divided. Same recipe energy as every other section — it just has a Greek letter on it. The symbol χ^2 is new; the moves are old friends. Say it: "A new symbol is not a new difficulty."

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 8 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 8 — Your Turn Answers

Math Skills: (1) Bigger — $t^* = 2.201 > 1.96$; the t-distribution adds extra cushion because we estimated the spread from the sample. (2) $\sqrt{0.0009} = 0.03$. (3) $1.96 \times \dfrac{10}{\sqrt{100}} = 1.96 \times \dfrac{10}{10} = 1.96 \times 1 = 2.0$.

Section 8.1: (1) The point estimate is 4.2 hours. (2) 4.2 ± 0.5, i.e. from 3.7 to 4.7 hours. (3) Example: "...the true mean daily screen time for the whole population is between 3.7 and 4.7 hours." (4) *False* — the true mean is a fixed number; the method (not this single interval) succeeds about 95% of the time.

Section 8.2: (1) $df = 16 - 1 = 15$. (2) $\dfrac{8}{\sqrt{16}} = \dfrac{8}{4} = 2.00$ minutes. (3) Margin $= 2.131 \times 2.00 = 4.26$; interval 28 ± 4.26, from 23.74 to 32.26 minutes. (4) "We are 95% confident the true average commute is between about 23.7 and 32.3 minutes."

Section 8.3: (1) $\hat{p} = \dfrac{220}{400} = 0.55$. (2) $\sqrt{\dfrac{0.55 \times 0.45}{400}} = \sqrt{\dfrac{0.2475}{400}} = \sqrt{0.00061875} = 0.0249$. (3) Margin $= 1.96 \times 0.0249 = 0.0488$; interval 0.55 ± 0.0488, from 0.501 to 0.599. (4) "We are 95% confident the true support is between about 50.1% and 59.9%."

Section 8.4: (1) $(n-1)s^2 = 9 \times 9 = 81$. (2) Variance interval $\left(\dfrac{81}{19.023}, \dfrac{81}{2.700}\right) = (4.26,\ 30.00)$. (3) σ interval $= (\sqrt{4.26}, \sqrt{30.00}) = (2.06,\ 5.48)$ hours. (4) It's wide; a larger sample n would shrink it.

Part IV Checkpoint — Study Guide & Practice (Confidence Intervals)

This Study Guide gathers everything about confidence intervals into one place and lets you rehearse before an exam the way the exam will actually feel — not a test, a dress rehearsal.

STUDY GUIDE & PRACTICE

You just learned how to take one sample and report an honest range instead of a single guess. That's a big deal, and it comes with a few moving parts: a point estimate, a cushion, a critical value, and the right way to say what you found. So let's pull it together. We'll line up the big ideas, check the vocabulary, and then run eight mixed problems that jump around exactly the way a real exam does. When your inner voice says "I'll never keep these formulas straight," keep going anyway — every answer is right here, fully worked, waiting for you.

The Big Ideas, in One Place

The one shape behind everything.

- Every confidence interval has the same shape: **point estimate** $\pm$ **margin of error**. The point estimate is your single best guess from the sample; the margin of error is the "give or take" cushion. Get those two pieces and you're done.
- The **confidence level** (usually 90%, 95%, or 99%) is the success rate of the *method*, not a probability about your one interval.

What "95% confident" really means (the #1 thing exams test).

- The true population value is a *fixed number* — it's either in your interval or it isn't. So you can NOT say "there's a 95% probability the truth is in *this* interval."
- The honest reading: if you repeated the whole process many times (a new sample each time), about 95% of the intervals you'd build would capture the true value. You're betting yours is one of the good ones.

- Safe sentence template: "We are 95% confident that the true [population mean / proportion / standard deviation] is between _____ and _____."

Confidence interval for a mean (t-interval).

- When you don't know the population spread (almost always), use the sample standard deviation s and the t-distribution:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- The piece $\frac{s}{\sqrt{n}}$ is the **standard error**. The critical value t^* is looked up using **degrees of freedom** $df = n - 1$.

- Recipe: (1) find $df = n-1$ and look up t^*; (2) compute the standard error $\frac{s}{\sqrt{n}}$; (3) multiply to get the margin $t^* \frac{s}{\sqrt{n}}$; (4) add and subtract from $\bar{x}$.

Confidence interval for a proportion (z-interval).

- When the thing you're estimating is a percentage, use the sample proportion $\hat{p}$ and the normal critical value z^* (no degrees of freedom):

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Memorize three z^* values: **90% → 1.645**, **95% → 1.96**, **99% → 2.576**. The everyday one is $z^* = 1.96$.

- $\hat{p}$ is the number of "yes" answers divided by n. The method assumes at least about 10 "yes" and 10 "no" outcomes.

Confidence interval for a variance or standard deviation (chi-square).

- To estimate a population variance σ^2, use the lopsided **chi-square distribution** χ^2, which needs *two* critical values (a left and a right) at $df = n - 1$:

$$\left(\frac{(n-1)s^2}{\chi^2_{right}} \, , \quad \frac{(n-1)s^2}{\chi^2_{left}} \right)$$

- The *bigger* value χ^2_{right} goes under the *left* (smaller) endpoint — dividing by a bigger number gives the smaller end.

- To get an interval for the **standard deviation** σ, take the $\sqrt{\ }$ of both endpoints.

How width changes (intuition exams love to check).

- A *larger sample* n makes the interval **narrower** — more information, less guessing. (The $\sqrt{n}$ in the bottom does the shrinking.)
- A *higher confidence level* (99% vs. 90%) makes the interval **wider** — more cushion is the price of being more sure.
- Intervals for spread (chi-square) are often very wide with small samples. That's honest, not broken.

Vocabulary Check

Can you say each of these in one plain sentence? If one feels fuzzy, that's exactly the one to reread — not a sign you're behind.

- point estimate, margin of error, confidence interval, confidence level
- the correct interpretation of "95% confident" (method success rate, not probability for one interval)
- standard error, critical value
- t-distribution, degrees of freedom $df = n - 1$
- sample proportion $\hat{p}$, critical value z^*
- the three common z^* values: 1.645, 1.96, 2.576
- chi-square distribution χ^2, variance σ^2, standard deviation σ
- how sample size n and confidence level change the width

Mixed Practice

These eight problems jump around on purpose — just like an exam. Try each one before peeking. The full worked answers come right after.

1. A nurse records the body temperatures of $n = 25$ healthy adults and finds $\bar{x} = 98.6\,^\circ\text{F}$ with sample standard deviation $s = 0.7\,^\circ\text{F}$. Build a 95% confidence interval for the true mean temperature μ. (Use $t^* = 2.064$ for $df = 24$.)

2. In a sample of $n = 600$ online shoppers, 312 say they abandoned a cart in the last month. Build a 95% confidence interval for the true proportion p who abandoned a cart. (Use $z^* = 1.96$.)

3. A study reports: "We are 95% confident the true mean is between 6.74 and 8.26 hours." A classmate rewrites this as: "There is a 95% probability the true mean is between 6.74 and 8.26 hours." Is the rewrite correct? If not, fix it and explain the difference in one or two sentences.

4. Two researchers study the same population. Researcher A surveys 100 people; Researcher B surveys $1{,}000$ people. Both build 95% confidence intervals for the same proportion. Whose interval will be *narrower*, and why? Then: if Researcher A switched from a 95% to a 99% confidence level (same sample), would her interval get wider or narrower?

5. A sample of $n = 36$ exam scores has $\bar{x} = 72$ points with $s = 12$ points. Build a 95% confidence interval for the true mean score μ. (Use $t^* = 2.030$ for $df = 35$.)

6. In a sample of $n = 150$ voters, 90 support a measure. Build a *99%* confidence interval for the true proportion p. (Use $z^* = 2.576$.)

7. A quality engineer computes a 95% confidence interval for the standard deviation of a part's length and gets $(1.83 \text{ mm},\ 3.94 \text{ mm})$ from a sample of $n = 15$ parts. (a) In plain words, what does this interval tell the engineer? (b) Is this interval wide or narrow, and what single change would tighten it?

8. A sample of $n = 15$ light bulbs has a sample standard deviation of $s = 2.5$ hundred-hours of life. Build a 95% confidence interval for the population standard deviation σ. (Use $df = 14$, $\chi^2_{right} = 26.119$, $\chi^2_{left} = 5.629$.)

Answers

Worked out in full — read these even for the ones you got right, because the *reasoning* is what the exam rewards.

1. This is a t-interval for a mean. **Step 1 — standard error:** $\dfrac{s}{\sqrt{n}} = \dfrac{0.7}{\sqrt{25}} = \dfrac{0.7}{5} = 0.14$. **Step 2 — margin of error:** $t^* \dfrac{s}{\sqrt{n}} = 2.064 \times 0.14 = 0.289$. **Step 3 — assemble:** 98.6 ± 0.289, which runs from 98.31 to $98.89\,^\circ$F. **Interpret:** "We are 95% confident that the true mean body temperature of all healthy adults is between 98.31 and $98.89\,^\circ$F."

2. This is a z-interval for a proportion. **Step 1 — sample proportion:** $\hat{p} = \dfrac{312}{600} = 0.52$ (so

$1 - \hat{p} = 0.48$). **Step 2 — standard error:** $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{0.52 \times 0.48}{600}} = \sqrt{\dfrac{0.2496}{600}} = \sqrt{0.000416} = 0.0204$. **Step 3 — margin of error:** $z^* \times 0.0204 = 1.96 \times 0.0204 = 0.040$. **Step 4 — assemble:** 0.52 ± 0.040 runs from 0.48 to 0.56. **Interpret:** "We are 95% confident that the true proportion of all shoppers who abandoned a cart is between 48% and 56%."

3. *No, the rewrite is wrong.* The true mean is a fixed number, so it doesn't have a probability of being inside one particular interval — it's either in there or it isn't. The correct statement is the original: "We are 95% confident the true mean is between 6.74 and 8.26 hours," which means the *method* that built this interval captures the truth about 95% of the time over many repeated samples. The confidence is about the long-run success rate of the procedure, not about this single interval.

4. **Researcher B's interval is narrower.** A larger sample ($1{,}000$ vs. 100) carries more information, and the $\sqrt{n}$ in the bottom of the standard error grows, shrinking the margin of error. More data $\rightarrow$ tighter, more useful interval. **Second part:** switching from 95% to 99% confidence makes Researcher A's interval *wider*, because a higher confidence level uses a bigger critical value ($z^* = 2.576$ instead of 1.96) — you pay for extra certainty with extra width.

5. This is a t-interval for a mean. **Step 1 — standard error:** $\dfrac{s}{\sqrt{n}} = \dfrac{12}{\sqrt{36}} = \dfrac{12}{6} = 2.0$. **Step 2 — margin of error:** $t^* \dfrac{s}{\sqrt{n}} = 2.030 \times 2.0 = 4.06$. **Step 3 — assemble:** 72 ± 4.06, which runs from 67.94 to 76.06 points. **Interpret:** "We are 95% confident that the true mean exam score is between 67.94 and 76.06 points."

6. This is a z-interval for a proportion at 99% confidence. **Step 1 — sample proportion:** $\hat{p} = \dfrac{90}{150} = 0.60$ (so $1 - \hat{p} = 0.40$). **Step 2 — standard error:** $\sqrt{\dfrac{0.60 \times 0.40}{150}} = \sqrt{\dfrac{0.24}{150}} = \sqrt{0.0016} = 0.04$. **Step 3 — margin of error:** $z^* \times 0.04 = 2.576 \times 0.04 = 0.103$. **Step 4 — assemble:** 0.60 ± 0.103 runs from 0.497 to 0.703. **Interpret:** "We are 99% confident that the true proportion of all voters who support the measure is between about 49.7% and 70.3%." (Notice how wide it is — that's the cost of demanding 99% confidence from only 150 people.)

7. (a) "We are 95% confident that the true standard deviation of the part's length — a measure of how much the parts vary from one another — is between 1.83 and 3.94 millimeters." It's a statement about the population's *consistency*, not its average. (b) The interval is *wide* (the upper end is more than double the lower end), which is normal when estimating spread from a small sample. The single change that would tighten it is a *larger sample size* n — estimating

variability well takes a lot of data.

8. This is a chi-square interval for σ. **Step 1 — setup:** $df = n-1 = 14$, and $s^2 = 2.5^2 = 6.25$, so $(n-1)s^2 = 14 \times 6.25 = 87.5$. **Step 2 — interval for the variance σ^2:** remember the bigger critical value goes under the smaller (left) endpoint,

$$\left(\frac{87.5}{26.119}\ ,\ \frac{87.5}{5.629}\right) = (3.35\ ,\ 15.55).$$

   Step 3 — take square roots to get σ: $\sqrt{3.35} = 1.83$ and $\sqrt{15.55} = 3.94$. **Interpret:** "We are 95% confident that the true standard deviation of bulb life is between 1.83 and 3.94 hundred-hours." Wide, as spread intervals from small samples usually are — more bulbs would tighten it.

You just rehearsed every kind of confidence interval — a mean with t, a proportion with z, a standard deviation with chi-square — plus the interpretation that exams love to trick you on, and the way width responds to sample size and confidence level. If some were shaky, now you know exactly where to look. That's not failing the rehearsal; that's what the rehearsal is *for*. You're ready.

Part IV — Formula Sheet: Estimating with Confidence

Every confidence-interval formula on one page, each symbol named in plain words with a note on when *to use it.*

FORMULA SHEET

The symbols, in plain words:

- $\bar{x}$ = sample mean s = sample standard deviation n = sample size $\hat{p}$ = sample proportion
- t^*, z^* = the critical value for your confidence level df = degrees of freedom χ^2 = chi-square value
- **Margin of error** = (critical value) × (standard error). A confidence interval is always **estimate ± margin of error**.

Common critical values z^*: 1.645 (90%) 1.96 (95%) 2.576 (99%)

Confidence interval for a mean (Chapter 8)

(σ unknown — the usual case): $\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$, $df = n - 1$ (σ known): $\bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}}$

When: estimating a population mean from one sample. Use t when you only have the sample SD s (almost always).

Confidence interval for a proportion (Chapter 8)

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

When: estimating a population proportion (a percentage/"what fraction") from one sample.

Confidence interval for a variance / standard deviation (Chapter 8)

for σ^2 : $\left(\dfrac{(n-1)s^2}{\chi^2_{right}},\ \dfrac{(n-1)s^2}{\chi^2_{left}}\right)$ for σ : take the square root of each endpoint

When: estimating how spread out a population is. Note the chi-square distribution is not symmetric, so the two critical values differ.

Reading the interval: "We are 95% confident the true [parameter] is between _ and _." This means 95% of intervals built this way capture the truth — *not* that there is a 95% probability the parameter is in this one interval. Higher confidence $\Rightarrow$ wider interval; larger n $\Rightarrow$ narrower interval.

Part V

Testing Claims — the p-Value Approach

Chapter 9

Testing What We Believe

Hypothesis Testing for One Population Parameter — the p-Value Approach

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll hear a claim — "our chargers last 10 hours," "half of customers prefer the new flavor" — and know exactly how to put it on trial. You'll collect evidence, measure how surprising it is, and deliver a verdict you can defend.

The Story. You already test beliefs constantly. A friend says "this bus is always on time," it's late twice, and you start to doubt them. Hypothesis testing is that same doubt, made careful: assume the claim is true, then ask how shocked we should be by the data we actually got.

The Skills You'll Have:

- Write a null hypothesis H_0 and an alternative H_a in symbols
- Compute a test statistic and read off its *p-value*
- Compare the p-value to α and make the right decision
- State the conclusion in plain English, in context
- Tell a Type I error from a Type II error

The Confidence Moment. When you see "$p = 0.01$, so we reject H_0," your brain will calmly translate: "the data were too surprising to believe the claim — so we don't."

The Bridge. This whole chapter tests *one* number against a claimed value. Chapter 10 does the natural next thing: it compares *two* populations — two means, two proportions — to ask whether they really differ. Same six steps, one more group.

MATH SKILLS YOU'LL NEED

A hypothesis test is mostly one piece of arithmetic, one lookup, and one comparison. Let's warm up all three.

Skill 1 — Plugging into a test-statistic formula. Every test statistic has the same shape: *(what we saw)* minus *(what was claimed)*, divided by a *standard error*. Substitute carefully and do the top first, then the bottom.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{52 - 50}{8/\sqrt{25}} = \frac{2}{8/5} = \frac{2}{1.6} = 1.25$$

Skill 2 — Reading a p-value as an area. A p-value is just a *probability* — an area under a curve, always between 0 and 1. "The area to the left of $t = -2.5$ is 0.012" means $p = 0.012$. Technology hands you this area; your job is to read it as "how often chance alone would give data this extreme."

Skill 3 — Comparing a decimal to α. The whole decision is one comparison. With $\alpha = 0.05$: is the p-value *less than or equal to* 0.05, or bigger? $0.012 \leq 0.05$ (yes), but $0.20 \leq 0.05$ (no). Smaller decimals win the comparison.

Practice (answers at the end of the chapter):

1. Compute $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ for $\hat{p} = 0.60$, $p_0 = 0.50$, $n = 100$.
2. A test gives a p-value of 0.032. Using $\alpha = 0.05$, do we reject H_0?
3. A test gives a p-value of 0.18. Using $\alpha = 0.05$, do we reject H_0?

9.1 The Logic of a Hypothesis Test

▪ READ THIS FIRST

Imagine a courtroom. The defendant is presumed *innocent* — that's the starting assumption, the thing we hold onto until evidence forces us to let go. The prosecutor brings evidence. If the evidence is weak, we keep the presumption of innocence. If the evidence is overwhelming — so unlikely under "innocent" that we can't explain it away — we reject innocence and convict.

A hypothesis test is exactly this courtroom. The "presumption of innocence" is a starting claim we assume is true. The "evidence" is our data. And the verdict comes down to one question: *if the claim were true, how surprising would this evidence be?*

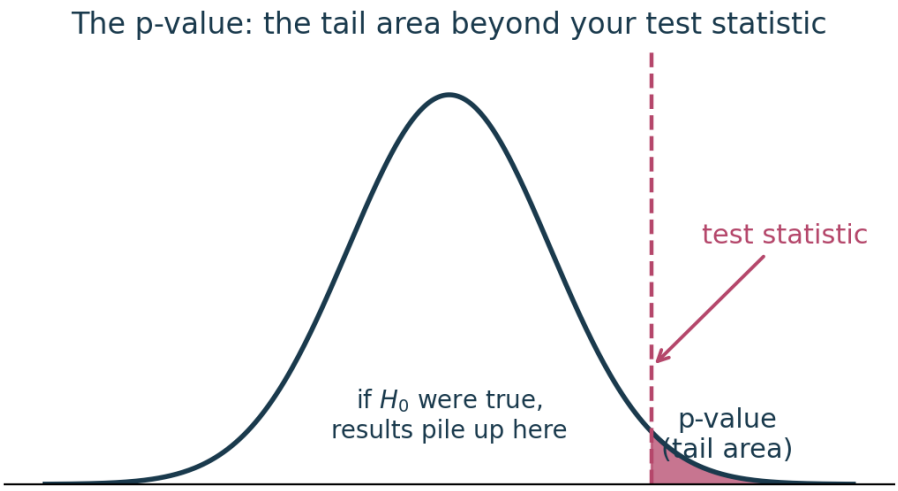


Figure 9.1. If H_0 were true, results pile up in the middle; the p-value is the tail area beyond your test statistic.

▪ LET'S TALK ABOUT IT

Think of a time you stopped believing something — a claim a product made, a promise about traffic, a rumor. How much evidence did it take before you flipped? Notice that you didn't need *proof* the claim was false; you just needed the world to look too weird if it were true. That instinct is the whole chapter.

▪ NOW WE NAME IT

Every test starts with two competing statements about a population parameter.

The **null hypothesis** H_0 is the "presumed innocent" claim — the status quo, the no-change, the boring default. It *always* contains an equals sign: $H_0 : \mu = 10$, or $H_0 : p = 0.5$. It's what we assume is true while we weigh the evidence.

The **alternative hypothesis** H_a is what we suspect instead — the thing we'd need evidence to believe. It uses $\neq$, $<$, or $>$:

- $H_a : \mu \neq 10$ — "different from" (two-sided)
- $H_a : \mu < 10$ — "less than" (one-sided)
- $H_a : \mu > 10$ — "greater than" (one-sided)

Now the key idea. We *assume* H_0 *is true*, then measure how far our data fall from it. The **p-value** is the probability of getting a test statistic as extreme as — or more extreme than — the one we observed, *assuming* H_0 *is true*. A tiny p-value means "if the claim were true, this data would be a freak event." A big p-value means "this data is totally ordinary under the claim."

The decision rule. Pick a **significance level** α first (usually 0.05) — your threshold for "too surprising." Then:

- If p-value $\leq \alpha$: *reject* H_0. The evidence is strong enough.
- If p-value $> \alpha$: *fail to reject* H_0. The evidence isn't strong enough.

Memory hook: **small p, reject** (the data were too surprising to keep believing the claim). Always end by stating the verdict in plain words, in context.

▪ WATCH IT WORK

Question: A bakery claims the mean weight of its loaves is 500 grams. A customer thinks the loaves are *lighter* than advertised. Set up the test and explain the logic — no numbers yet.

Step 1 — State the claim and the suspicion. The claim (status quo) is "mean $= 500$ g." The suspicion is "mean is *less than* 500 g."

Step 2 — Write H_0. The null is always the equals claim: $H_0 : \mu = 500$.

Step 3 — Write H_a. "Lighter" means less than, so $H_a : \mu < 500$. (One-sided.)

Step 4 — State the assumption. We *assume* $\mu = 500$ is true, then collect loaf weights.

Step 5 — State what the p-value will mean. It's the probability of seeing a sample mean as low as ours (or lower) *if the true mean really is 500*.

Step 6 — State the decision rule. If that probability is $\leq \alpha = 0.05$, the data are too surprising under the claim — reject H_0 and conclude the loaves are lighter. Otherwise, fail to reject: not enough evidence.

▪ YOUR TURN

A phone company advertises that its average customer-service wait is 5 minutes. A reporter suspects the true wait is *longer*.

1. Write H_0 in symbols.
2. Write H_a in symbols. Is it one-sided or two-sided?
3. In one sentence, what would a *small* p-value tell us?

Work space:

▪ CHECK YOUR VOICE

If "null" and "alternative" feel like backwards jargon — you're not confused, you're noticing that the names are clunky. The *idea* underneath is one you've used your whole life: assume the ordinary story, and only abandon it when the world gets too weird. You already own the logic. We're just bolting symbols onto it.

9.2 Testing a Claim About a Mean

▪ READ THIS FIRST

A flashlight company prints "battery lasts 10 hours" on every box. You buy a 16-pack, time them all, and your batch averages 9.5 hours. Is the company lying — or is 9.5 just normal wobble for a true 10-hour battery? You can't tell by staring. You need a number that says how surprising 9.5 is *if the truth really is 10*.

▪ LET'S TALK ABOUT IT

Before any math: would you be more suspicious if your 16 batteries averaged 9.5 hours, or if a single battery you tested lasted 9.5 hours? Why does the *number you tested* change how surprised you are? (You're feeling out sample size — the $\sqrt{n}$ that's about to show up.)

▪ NOW WE NAME IT

To test a claim about a population mean μ, we use the **t-test** for one mean. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where $\bar{x}$ is the sample mean, μ_0 is the claimed value from H_0, s is the sample standard deviation, and n is the sample size. The bottom, $s/\sqrt{n}$, is the standard error — how much sample means typically bounce around. We compare t to a *t-distribution* with $df = n - 1$ degrees of freedom, and read the p-value as the tail area (technology or a table gives it).

▪ WATCH IT WORK

Question: The flashlight box claims $\mu = 10$ hours. A consumer suspects the true mean is *less*. A sample of $n = 16$ batteries has $\bar{x} = 9.5$ hours and $s = 0.8$ hours. Test at $\alpha = 0.05$.

Step 1 — Hypotheses. $H_0 : \mu = 10$ versus $H_a : \mu < 10$ (one-sided, "less than").

Step 2 — Choose α. $\alpha = 0.05$.

Step 3 — Compute the test statistic.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{9.5 - 10}{0.8/\sqrt{16}} = \frac{-0.5}{0.8/4} = \frac{-0.5}{0.2} = -2.5.$$

The degrees of freedom: $df = n - 1 = 15$.

Step 4 — Find the p-value. Because H_a is "less than," the p-value is the area to the *left* of $t = -2.5$ on the t-distribution with $df = 15$. Technology gives $p \approx 0.012$.

Step 5 — Decide. Compare to α: $0.012 \leq 0.05$. Small p $\rightarrow$ *reject* H_0.

Step 6 — Conclude in context. There is significant evidence (at $\alpha = 0.05$) that the true mean battery life is *less than* 10 hours. If the box's claim were true, a sample this low would happen only about 1.2% of the time — too surprising to keep believing the claim.

▪ YOUR TURN

A gym advertises that the mean workout visit lasts 45 minutes. A member doubts this and thinks it's *different* (either way). A sample of $n = 25$ visits has $\bar{x} = 41$ minutes and $s = 10$ minutes. Use $\alpha = 0.05$.

1. Write H_0 and H_a. (Is it one- or two-sided?)
2. Compute t and state df.
3. The p-value comes out to about 0.057. What's your decision, and what do you conclude in context?

Work space:

▪ CHECK YOUR VOICE

That last one is sneaky on purpose: $p = 0.057$ is *just* above 0.05. If part of you wants to "round it down" and reject anyway — that's a very human urge, and it's exactly what the α line protects you from. The rule is the rule. "Fail to reject" isn't a failure on your part; it's an honest verdict. You read the number correctly, and that's the skill.

9.3 Testing a Claim About a Proportion

▪ READ THIS FIRST

A snack company reformulates its chips and announces "half of all customers prefer the new recipe." That's a claim about a *proportion* — a percentage of a whole population. You run a taste test with 200 people, and only 86 prefer the new chips. That's 43%, not 50%. Is the company's claim wrong, or is 43% just sampling wobble around a true 50%?

▪ LET'S TALK ABOUT IT

If you flipped a fair coin 200 times, would you expect *exactly* 100 heads? Probably not — maybe 94, maybe 103. So when 86 of 200 people prefer the new recipe, the real question isn't "is it exactly 100?" but "is 86 far enough from 100 to doubt the coin?" That gap is what the test measures.

▪ NOW WE NAME IT

To test a claim about a population proportion p, we use the **z-test** for one proportion. The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}},$$

where $\hat{p}$ is the sample proportion (successes divided by n), p_0 is the claimed proportion from H_0, and n is the sample size. Notice the standard error uses p_0 — the *claimed* value — because we're assuming H_0 is true while we measure surprise. We compare z to the *standard normal* distribution and read the p-value as a tail area.

▪ WATCH IT WORK

Question: The company claims $p = 0.5$ of customers prefer the new recipe. We suspect the true proportion is *different* from 0.5. In a sample of $n = 200$, exactly 86 prefer it. Test at $\alpha = 0.05$.

Step 1 — Hypotheses. $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$ (two-sided, "different from").

Step 2 — Choose α. $\alpha = 0.05$.

Step 3 — Compute the test statistic. First the sample proportion: $\hat{p} = \dfrac{86}{200} = 0.43$. Then

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.43 - 0.5}{\sqrt{(0.5)(0.5)/200}} = \frac{-0.07}{\sqrt{0.00125}} = \frac{-0.07}{0.03536} \approx -1.98.$$

Step 4 — Find the p-value. Because H_a is "$\neq$" (two-sided), we take the area in *both* tails: the area beyond -1.98 *and* beyond $+1.98$. The area left of -1.98 is about 0.0239, so $p \approx 2 \times 0.0239 = 0.048$.

Step 5 — Decide. Compare to α: $0.048 \leq 0.05$. Small p $\rightarrow$ *reject* H_0.

Step 6 — Conclude in context. There is (just barely) significant evidence at $\alpha = 0.05$ that the true proportion preferring the new recipe is *not* 50%. Our sample suggests it's lower. Notice how close this was — a reminder that "significant" and "barely significant" can look almost identical.

▪ YOUR TURN

An ad agency claims that 20% of viewers click a certain banner ad. A client suspects the true click rate is *higher*. In a sample of $n = 150$ viewers, 39 clicked. Use $\alpha = 0.05$.

1. Write H_0 and H_a. (One- or two-sided?)
2. Find $\hat{p}$, then compute z.
3. The p-value is about 0.033. State your decision and conclude in context.

Work space:

▪ CHECK YOUR VOICE

Two different formulas now (t for means, z for proportions), and your brain might be bracing to memorize a pile of them. Here's the relief: every test statistic is the *same sentence* — "how many standard errors is our data from the claim?" Top: observed minus claimed. Bottom: standard error. If you understand that one shape, you're not memorizing four formulas; you're filling in one.

9.4 Testing a Claim About a Variance (and Avoiding Mistakes)

▪ READ THIS FIRST

Sometimes the average is fine but the *consistency* is the problem. A coffee machine is supposed to pour cups that don't vary much — a standard deviation of 2 milliliters, say (so a variance of 4). If the pours start swinging wildly, customers get short-changed or overflowing cups even when the *average* pour looks perfect. To test consistency, we test the *variance*.

▪ LET'S TALK ABOUT IT

Can you think of a situation where you'd care more about *consistency* than about the average? (A bus that's reliably 5 minutes late vs. one that's "on average on time" but swings from 20 minutes early to 20 minutes late.) Variance is the math of "how spread out," and sometimes that's the whole ballgame.

▪ NOW WE NAME IT

To test a claim about a population variance σ^2, we use the **chi-square test** for one variance. The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

where s^2 is the sample variance, σ_0^2 is the claimed variance from H_0, and n is the sample size. We compare it to a *chi-square distribution* with $df = n - 1$, and read the p-value as a tail area. (Don't memorize the chi-square shape — just know it's the right tool when the claim is about *spread*, and let technology supply the p-value.)

▪ WATCH IT WORK

Question: The machine's maker claims the variance of pour size is $\sigma^2 = 4$ (ml^2). A manager suspects the pours are *more variable* than that. A sample of $n = 20$ pours has $s^2 = 6.5$. Test at $\alpha = 0.05$.

Step 1 — Hypotheses. $H_0 : \sigma^2 = 4$ versus $H_a : \sigma^2 > 4$ (one-sided, "more variable").

Step 2 — Choose α. $\alpha = 0.05$.

Step 3 — Compute the test statistic.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(20-1)(6.5)}{4} = \frac{(19)(6.5)}{4} = \frac{123.5}{4} = 30.875.$$

The degrees of freedom: $df = n - 1 = 19$.

Step 4 — Find the p-value. Because H_a is "greater than," the p-value is the area to the *right* of $\chi^2 = 30.875$ on the chi-square distribution with $df = 19$. Technology gives $p \approx 0.042$.

Step 5 — Decide. Compare to α: $0.042 \leq 0.05$. Small p $\rightarrow$ *reject* H_0.

Step 6 — Conclude in context. There is significant evidence at $\alpha = 0.05$ that the true variance of pour size is *greater* than 4 — the machine is pouring less consistently than claimed.

▪ NOW WE NAME IT

Two ways to be wrong (and one thing "fail to reject" never means). Because we decide from a *sample*, we can reach the wrong verdict two ways.

A **Type I error** is *rejecting a true* H_0 — convicting an innocent claim. We chose α precisely to control this: $\alpha = 0.05$ means we accept a 5% chance of this mistake.

A **Type II error** is *failing to reject a false* H_0 — letting a guilty claim walk free because our evidence wasn't strong enough.

	H_0 is really true	H_0 is really false
We reject H_0	Type I error	Correct
We fail to reject H_0	Correct	Type II error

And the most important habit in this chapter: "fail to reject H_0" does *not* prove H_0 is true. It only means we didn't find enough evidence against it — like a "not guilty" verdict, which means "not proven guilty," not "proven innocent." We never say "accept H_0" or "H_0 is true." We say "we fail to reject H_0" and "there is not enough evidence." Keeping that language honest is the mark of someone who truly understands testing.

▪ YOUR TURN

A bottling plant claims the variance of fill volume is $\sigma^2 = 25$. A quality engineer suspects the variance is *different* from 25. A sample of $n = 15$ bottles has $s^2 = 40$.

1. Write H_0 and H_a.
2. Compute χ^2 and state df.
3. A lab tech runs a test, gets a big p-value, and writes "we have proven the variance is 25." What's wrong with that sentence?

Work space:

▪ CHECK YOUR VOICE

If the difference between Type I and Type II made your eyes cross, try the courtroom: Type *I* is convicting the *innocent* (rejecting a true H_0); Type *II* is freeing the *guilty* (keeping a false H_0). And if "fail to reject doesn't mean accept" feels like splitting hairs — it isn't. It's the single most common mistake in all of statistics, and you now phrase it correctly. That's not pedantry; that's mastery.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 9 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 9 — Your Turn Answers

Math Skills: (1) $z = \dfrac{0.60 - 0.50}{\sqrt{(0.5)(0.5)/100}} = \dfrac{0.10}{\sqrt{0.0025}} = \dfrac{0.10}{0.05} = 2.0.$ (2) $0.032 \leq 0.05$, so *reject* H_0. (3) $0.18 > 0.05$, so *fail to reject* H_0.

Section 9.1: (1) $H_0 : \mu = 5$. (2) $H_a : \mu > 5$ — one-sided ("longer"). (3) A small p-value would mean: if the true mean wait really were 5 minutes, getting a sample wait time as high as ours would be very unlikely — so we'd doubt the 5-minute claim.

Section 9.2: (1) $H_0 : \mu = 45$ versus $H_a : \mu \neq 45$ — two-sided ("different"). (2) $t = \dfrac{41 - 45}{10/\sqrt{25}} = \dfrac{-4}{10/5} = \dfrac{-4}{2} = -2.0$, with $df = 24$. (3) $p \approx 0.057 > 0.05$, so *fail to reject* H_0: there is not enough evidence (at $\alpha = 0.05$) that the mean visit length differs from 45 minutes.

Section 9.3: (1) $H_0 : p = 0.20$ versus $H_a : p > 0.20$ — one-sided ("higher"). (2) $\hat{p} = \dfrac{39}{150} = 0.26$; $z = \dfrac{0.26 - 0.20}{\sqrt{(0.20)(0.80)/150}} = \dfrac{0.06}{\sqrt{0.001067}} = \dfrac{0.06}{0.03266} \approx 1.84.$ (3) $p \approx 0.033 \leq 0.05$, so *reject* H_0: there is significant evidence the true click rate is greater than 20%.

Section 9.4: (1) $H_0 : \sigma^2 = 25$ versus $H_a : \sigma^2 \neq 25$. (2) $\chi^2 = \dfrac{(15-1)(40)}{25} = \dfrac{(14)(40)}{25} = \dfrac{560}{25} = 22.4$, with $df = 14$. (3) A big p-value only means we *failed to reject* H_0 — not enough evidence against it. It never *proves* the variance is 25. The correct wording is "we fail to reject H_0; there is not enough evidence that the variance differs from 25."

Chapter 10

Comparing Two Groups

Hypothesis Testing for Two Population Parameters — Means, Proportions, and Pairs

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. Up to now you tested one number against a claim — "is the average really 10 minutes?" This chapter does the thing you've secretly wanted to do all along: *compare two groups*. Is the new method better than the old one? Do men and women answer differently? Did the training actually change anything? You'll learn to put a real number on "these two groups look different."

The Story. You compare groups constantly — two coffee shops, two routes to work, before-and-after a new habit. Your gut says "A seems better than B." Statistics asks the careful follow-up: *is that gap real, or just the luck of who we happened to measure?*

The Skills You'll Have:

- Test whether two *independent* groups have different means (Welch's t)
- Test *paired* data — the same people measured twice — and know why it's different
- Test whether two groups have different *proportions* (with the pooled $\hat{p}$ done right)
- Take a gentle first look at comparing two *spreads* with an F-test

The Confidence Moment. When someone says "the difference wasn't statistically significant," your brain will calmly translate: "the gap between the two groups was small enough to be explained by ordinary sampling luck."

The Bridge. Comparing *two* groups is the natural next step after testing one. And the very last section — comparing two spreads — quietly sets up Chapter 11, where we compare *three or more* groups at once with a tool called ANOVA.

MATH SKILLS YOU'LL NEED

Every test in this chapter is the same shape you already know: $\dfrac{\text{(estimate)} - \text{(what } H_0 \text{ claims)}}{\text{standard error}}$. The only new wrinkle is that the estimate is now a *difference* between two groups, and the standard error combines *two* sample sizes. Let's warm up the moves.

Skill 1 — Square roots of a sum. Add *first*, then take the root. Don't take roots of each piece separately.

$$\sqrt{\frac{64}{25} + \frac{100}{25}} = \sqrt{2.56 + 4} = \sqrt{6.56} \approx 2.561$$

Skill 2 — Working with two sample sizes and SDs. A formula like $\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}$ just says: square each group's SD, divide by that group's sample size, then add. If $s_1 = 8,\ n_1 = 25,\ s_2 = 10,\ n_2 = 25$:

$$\frac{8^2}{25} + \frac{10^2}{25} = \frac{64}{25} + \frac{100}{25} = 2.56 + 4 = 6.56$$

Skill 3 — The difference of two estimates. The numerator is almost always one estimate minus the other: $\bar{x}_1 - \bar{x}_2$ or $\hat{p}_1 - \hat{p}_2$. If $\bar{x}_1 = 78$ and $\bar{x}_2 = 72$, then $\bar{x}_1 - \bar{x}_2 = 6$. A positive answer means group 1 was higher; a negative answer means group 2 was higher. The *sign* carries meaning — keep track of which group is which.

Practice (answers at the end of the chapter):

1. Evaluate $\sqrt{\dfrac{36}{9} + \dfrac{16}{4}}$.
2. If $s_1 = 6,\ n_1 = 9,\ s_2 = 4,\ n_2 = 4$, compute $\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}$.
3. If $\hat{p}_1 = 0.45$ and $\hat{p}_2 = 0.30$, what is $\hat{p}_1 - \hat{p}_2$?

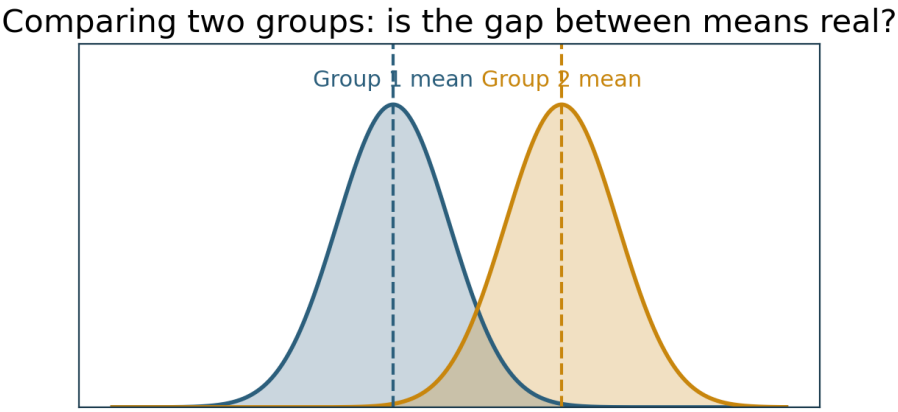


Figure 10.1. Two groups with different sample means — a hypothesis test asks whether the gap is real or just chance.

10.1 Two Independent Means

▪ READ THIS FIRST

You and a friend each take a different route to work for a couple of weeks and write down your commute times. Your route averaged 28 minutes; theirs averaged 33. "Mine's faster," you announce.

But wait — some of your mornings were lucky, some of theirs were unlucky. The two averages came from *two different sets of days*. The real question isn't "is 28 less than 33?" (obviously it is). It's: *is the 5-minute gap big enough that it couldn't easily be explained by ordinary day-to-day wobble?*

That's a two-group comparison. Two separate groups, measured separately — and a gap we need to weigh against chance.

▪ LET'S TALK ABOUT IT

Imagine the two routes were actually identical in true average time, and you just got slightly luckier on your two weeks. Could a 5-minute gap still show up by pure chance? What if the gap had been 30 minutes — would *that* feel like luck? Somewhere between "easily luck" and "no way that's luck" is the line we're learning to draw.

▪ NOW WE NAME IT

When you have **two independent samples** — two *separate* groups of subjects, with no link between an individual in one group and an individual in the other — you can test whether their population means differ.

The null hypothesis says the two population means are equal:

$$H_0 : \mu_1 = \mu_2 \qquad \text{(equivalently, } \mu_1 - \mu_2 = 0\text{)}.$$

The alternative $H_a : \mu_1 \neq \mu_2$ says they differ (a two-tailed test).

The test statistic is **Welch's** t, which weighs the observed difference against its standard error:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

The top is the gap we actually saw. The bottom is how much that gap would naturally wobble from sampling alone. A big t means the gap is large *compared to* the wobble — hard to chalk up to luck.

About the degrees of freedom. Welch's t uses a slightly messy df formula that technology computes for you. For *hand* work, a safe shortcut is to use the smaller of $n_1 - 1$ and $n_2 - 1$ — this is conservative, meaning it errs on the side of *not* declaring a difference. Always trust jamovi's reported df and p-value when you have it.

▪ WATCH IT WORK

Question: Two sections of the same class take the same final. Section A: $\bar{x}_1 = 78$, $s_1 = 8$, $n_1 = 25$. Section B: $\bar{x}_2 = 72$, $s_2 = 10$, $n_2 = 25$. At $\alpha = 0.05$, is there a difference in true mean scores?

Step 1 — Hypotheses. $H_0 : \mu_1 = \mu_2$ (the sections have the same true mean) versus $H_a : \mu_1 \neq \mu_2$ (they differ). Two-tailed.

Step 2 — Significance level. $\alpha = 0.05$.

Step 3 — Check conditions. The two sections are separate groups (independent), scores are quantitative, and each $n = 25$ is reasonably large, so the t-procedure is appropriate.

Step 4 — Test statistic. Build the standard error first:

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{64}{25} + \frac{100}{25} = 2.56 + 4 = 6.56, \qquad \sqrt{6.56} \approx 2.561.$$

Then

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{6.56}} = \frac{78 - 72}{2.561} = \frac{6}{2.561} \approx 2.34.$$

Step 5 — P-value. Using the conservative $df = \min(25-1,\, 25-1) = 24$, a two-tailed test with $t \approx 2.34$ gives a p-value of about 0.028. (jamovi, with Welch's exact df, reports a very similar value.)

Step 6 — Decide and say it in words. Since $0.028 < 0.05$, we *reject* H_0. In plain words: "The 6-point gap between the sections is larger than sampling luck comfortably explains — there is statistically significant evidence the two sections have different true mean scores."

▪ YOUR TURN

A store compares wait times (minutes) at two checkout designs, measured on different customers. Design 1: $\bar{x}_1 = 6.0$, $s_1 = 2$, $n_1 = 16$. Design 2: $\bar{x}_2 = 7.5$, $s_2 = 3$, $n_2 = 16$. Use $\alpha = 0.05$.

1. State H_0 and H_a.
2. Compute $\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}$ and its square root.
3. Compute the test statistic t.
4. With conservative $df = 15$, the two-tailed p-value is about 0.12. At $\alpha = 0.05$, what do you conclude?

Work space:

▪ CHECK YOUR VOICE

If the formula looked intimidating, look again at what you actually *did*: you squared two numbers, divided each by a sample size, added, took a square root, then did one division. That's it. The scary-looking fraction is just the one-sample t from the last chapter with a second group bolted on. You already owned the move — this is the same move, twice.

10.2 Paired (Matched) Means

▪ READ THIS FIRST

Now a different story. Ten people join a four-week walking program. You weigh each person *before* and weigh the *same person* after. You're not comparing two separate groups — you're comparing each person *to themselves*.

That changes everything. The natural number to look at isn't "average before" versus "average after." It's each person's own *change*: how much did *this* person's weight move? Subtract before from after, get one difference per person, and suddenly you're back to a one-group problem — a problem about the differences.

▪ LET'S TALK ABOUT IT

Why is measuring the same person twice better than grabbing two separate groups? Think about it this way: people start at wildly different weights. If you used two separate groups, that huge person-to-person variation would swamp the small program effect. But when each person is their own comparison, all that baseline difference *cancels out*. What's left is the change. Can you feel why pairing is so powerful?

▪ NOW WE NAME IT

Data is **paired** (or *matched*) when each value in one group is naturally linked to exactly one value in the other — the *same* subject measured twice (before/after), or two subjects matched as a pair (twins, left hand/right hand, husband/wife).

This is the classic confusion, so let's nail it: the difference between *paired* and *independent* is about the *study design*, not the numbers.

- **Independent (Section 10.1):** two *different* sets of subjects. Section A students are not the same people as Section B students. No pairing.
- **Paired (this section):** the *same* subjects measured twice, or deliberately matched. Person 3's "before" goes with Person 3's "after." Pairing.

Ask yourself: "Can I draw a line connecting each value on the left to one specific value on the right?" If yes — it's paired. If the two groups are just two unrelated piles — it's independent.

For paired data we compute each pair's difference d (for example, before $-$ after), and then we test those differences as a single group. (Either subtraction order works — just pick one, use it for every pair, and remember the *sign* of $\bar{d}$ then tells you the direction of the change.) The null says the average difference is zero:

$$H_0 : \mu_d = 0 \qquad \text{(no average change)}.$$

The test statistic is the one-sample t on the differences:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}, \qquad df = n - 1,$$

where $\bar{d}$ is the mean of the differences, s_d is their standard deviation, and n is the *number of pairs*.

▪ WATCH IT WORK

Question: Six people are weighed (pounds) before and after a program. Did weight change on average? Use $\alpha = 0.05$.

Person	1	2	3	4	5	6
Before	210	195	180	220	200	185
After	204	190	178	210	196	182
d (before − after)	6	5	2	10	4	3

Step 1 — Hypotheses. $H_0 : \mu_d = 0$ (no average weight change) versus $H_a : \mu_d \neq 0$.

Step 2 — Significance level. $\alpha = 0.05$.

Step 3 — Check the design. Each person was measured twice — the *same* person before and after. This is paired data, so we work with the six differences d.

Step 4 — Test statistic. First the mean difference:

$$\bar{d} = \frac{6 + 5 + 2 + 10 + 4 + 3}{6} = \frac{30}{6} = 5.$$

Now the SD of the differences. Deviations from 5 are $1, 0, -3, 5, -1, -2$; their squares are $1, 0, 9, 25, 1, 4$, summing to 40. So

$$s_d^2 = \frac{40}{6-1} = \frac{40}{5} = 8, \qquad s_d = \sqrt{8} \approx 2.828.$$

Then

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{5}{2.828/\sqrt{6}} = \frac{5}{2.828/2.449} = \frac{5}{1.155} \approx 4.33.$$

Step 5 — P-value. With $df = n - 1 = 5$, a two-tailed test with $t \approx 4.33$ gives a p-value of about 0.0075.

Step 6 — Decide and say it in words. Since $0.0075 < 0.05$, we *reject* H_0. In plain words: "People lost about 5 pounds on average, and that change is far too consistent to be sampling luck — there is significant evidence the program changed weight."

▪ YOUR TURN

Five students take a quiz before and after a study session. Their *improvements* (after − before) are: $4, 6, 3, 5, 2$ points. Use $\alpha = 0.05$.

1. Is this paired or independent data? Why?
2. Compute $\bar{d}$.
3. The standard deviation of these differences is $s_d \approx 1.581$. Compute $t = \dfrac{\bar{d}}{s_d/\sqrt{5}}$.
4. With $df = 4$, the two-tailed p-value is about 0.005. What do you conclude?

Work space:

▪ CHECK YOUR VOICE

If you've been quietly worried about telling paired from independent — good, that worry means you spotted the real trap. Here's your lifeline, forever: *can I connect each left value to one specific right value with a line?* Same person twice, or matched pairs → paired. Two unrelated piles of people → independent. You don't need to memorize a rule; you just answer that one question.

10.3 Two Proportions

▪ READ THIS FIRST

Sometimes the thing you're comparing is a yes-or-no, not an amount. Did the customer come back: yes or no? Did the patient recover: yes or no? Now you're comparing two *proportions* — the "yes" rate in group 1 versus the "yes" rate in group 2.

Say a new ad got 45% of viewers to click, and the old ad got 30%. A 15-point gap looks big. But each rate came from a limited sample. The familiar question returns: *is that gap real, or could two equal-quality ads land 15 points apart just by who happened to see them?*

▪ LET'S TALK ABOUT IT

Suppose both ads were truly equally good — same true click rate. You show each to 100 people. Would you expect the two observed click rates to come out *exactly* equal? Or would a few points of difference show up just from random chance? And would a 15-point gap feel like "a few points" to you, or like something more?

▪ NOW WE NAME IT

With **two proportions** we test whether two populations have the same true "yes" rate. The null says they're equal:

$$H_0 : p_1 = p_2 \qquad \text{versus} \qquad H_a : p_1 \neq p_2.$$

Here's the one new idea. *If H_0 is true*, the two groups share one common proportion — so our best estimate of it uses *all* the yes-answers from *both* groups combined. That's the **pooled proportion**:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

where x_1, x_2 are the counts of "yes" in each group. (Pool the *counts*, not the two percentages — add the numerators and add the denominators.)

The test statistic is a z, using that pooled $\hat{p}$ in the standard error:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The top is the observed gap in sample proportions. The bottom is how much that gap would wobble if the two true rates were actually equal. As always: big z means the gap is large compared to the wobble.

▪ WATCH IT WORK

Question: A new checkout page is tested. Of $n_1 = 100$ visitors who saw it, $x_1 = 45$ made a purchase. Of $n_2 = 100$ who saw the old page, $x_2 = 30$ purchased. At $\alpha = 0.05$, do the purchase rates differ?

Step 1 — Hypotheses. $H_0 : p_1 = p_2$ (same true purchase rate) versus $H_a : p_1 \neq p_2$. Two-tailed.

Step 2 — Significance level. $\alpha = 0.05$.

Step 3 — Sample proportions and conditions. $\hat{p}_1 = \dfrac{45}{100} = 0.45$ and $\hat{p}_2 = \dfrac{30}{100} = 0.30$. Each group has well over 10 "yes" and 10 "no," so the normal approximation is fine.

Step 4 — Pooled proportion, then the test statistic. Pool the counts:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{45 + 30}{100 + 100} = \frac{75}{200} = 0.375.$$

Build the standard error:

$$\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0.375(0.625)\left(\frac{1}{100} + \frac{1}{100}\right) = 0.234375 \times 0.02 = 0.0046875,$$

$$\sqrt{0.0046875} \approx 0.06847.$$

Then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{0.06847} = \frac{0.45 - 0.30}{0.06847} = \frac{0.15}{0.06847} \approx 2.19.$$

Step 5 — P-value. For a two-tailed z-test, $z \approx 2.19$ gives a p-value of about 0.029. (The area beyond $z = 2.19$ in one tail is about 0.0143; double it.)

Step 6 — Decide and say it in words. Since $0.029 < 0.05$, we *reject* H_0. In plain words: "The new page's 15-point higher purchase rate is larger than chance comfortably explains — there is significant evidence the two pages have different true purchase rates."

▪ YOUR TURN

A clinic compares recovery rates. Treatment group: $x_1 = 60$ recovered out of $n_1 = 80$. Control group: $x_2 = 50$ recovered out of $n_2 = 80$. Use $\alpha = 0.05$.

1. State H_0 and H_a.
2. Compute $\hat{p}_1$, $\hat{p}_2$, and the pooled $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$.
3. Compute the standard error $\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$ (round to four decimals) and then z.
4. The two-tailed p-value for your z is about 0.09. At $\alpha = 0.05$, what do you conclude?

Work space:

▪ CHECK YOUR VOICE

The single most common slip here is averaging the two percentages to get the pooled proportion. Don't. Pool the raw *counts*: add the yeses, add the totals, divide. If a voice says "I'll never remember which one to pool" — you just did, by reading this. Add $x_1 + x_2$ on top, $n_1 + n_2$ on the bottom. That's the whole rule.

10.4 Two Variances (a Brief Look)

▪ READ THIS FIRST

So far we've compared *centers* — means and proportions. But sometimes the interesting difference isn't the average; it's the *consistency*.

Picture two machines filling cereal boxes. Both average 500 grams. But Machine A is rock-steady — almost every box is right at 500 — while Machine B swings from 470 to 530. Same center, very different *spread*. If you care about reliability, Machine B is the problem, and no comparison of means would ever catch it. To compare spreads, we compare *variances*.

▪ LET'S TALK ABOUT IT

Which would you rather buy from: a machine that's usually right but occasionally wildly off, or one that's a tiny bit low but always the same? There's no universal right answer — but notice that "how much it varies" can matter as much as "what it averages." That's the whole reason this tool exists.

▪ NOW WE NAME IT

To compare two spreads we compare the two sample *variances* (s^2, the squared standard deviations) using the **F-test**. The null says the spreads are equal:

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

The test statistic is simply the ratio of the two sample variances — and by convention we put the *larger* variance on top so F is at least 1:

$$F = \frac{s_1^2}{s_2^2} \qquad \text{(larger } s^2 \text{ over smaller } s^2\text{)}.$$

If the two spreads were truly equal, this ratio would sit near 1. The further F climbs above 1, the more the spreads appear to differ. Technology turns F (with its two degrees of freedom, $n_1 - 1$ and $n_2 - 1$) into a p-value.

A gentle caution: the F-test for variances is sensitive to non-normal data, so treat it as a *first look* rather than a final word. We keep it short here on purpose.

Why it matters for what's next: this same F ratio — a comparison of variances — is the engine of Chapter 11's ANOVA, where we compare *three or more* group means at once by cleverly comparing variances. So this short section is really a doorway.

▪ WATCH IT WORK

Question: Two filling machines are sampled. Machine A: $s_1 = 10$ grams. Machine B: $s_2 = 5$ grams. Is Machine A noticeably more variable?

Step 1 — Hypotheses. $H_0 : \sigma_1^2 = \sigma_2^2$ (equal spreads) versus $H_a : \sigma_1^2 \neq \sigma_2^2$ (different spreads).

Step 2 — Significance level. Use $\alpha = 0.05$.

Step 3 — Check the idea. We're comparing consistency, not centers, so a variance comparison is the right tool. Put the larger variance on top.

Step 4 — Test statistic. Square each SD: $s_1^2 = 10^2 = 100$ and $s_2^2 = 5^2 = 25$. Then

$$F = \frac{s_1^2}{s_2^2} = \frac{100}{25} = 4.$$

Machine A's variance is four times Machine B's.

Step 5 — P-value. An F of 4 (with its two degrees of freedom) is fed to technology, which returns the p-value. A ratio this far above 1 is typically small-p — the spreads look genuinely different.

Step 6 — Say it in words. "Machine A varies about four times as much as Machine B; that's a large enough gap in spread to flag Machine A as the less consistent machine." For the exact p-value, we'd let jamovi do the F-distribution lookup.

▪ YOUR TURN

Two graders' scores are compared for consistency. Grader 1: $s_1 = 8$. Grader 2: $s_2 = 4$.

1. State H_0 in symbols.
2. Compute s_1^2 and s_2^2.
3. Compute F (larger over smaller).
4. In one sentence, what does an F well above 1 suggest about the two graders' consistency?

Work space:

▪ CHECK YOUR VOICE

This section was deliberately short and gentle — and if it felt easy, that's not a sign you missed something deep. The F-test really is just "divide one variance by the other and see how far from 1 it lands." Hold onto that simple picture; in the next chapter it grows into ANOVA, and you'll be glad the seed is already planted.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 10 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 10 — Your Turn Answers

Math Skills: (1) $\sqrt{\frac{36}{9}+\frac{16}{4}} = \sqrt{4+4} = \sqrt{8} \approx 2.828$. (2) $\frac{6^2}{9}+\frac{4^2}{4} = \frac{36}{9}+\frac{16}{4} = 4+4 = 8$. (3) $\hat{p}_1 - \hat{p}_2 = 0.45 - 0.30 = 0.15$.

Section 10.1: (1) $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$. (2) $\frac{2^2}{16}+\frac{3^2}{16} = \frac{4}{16}+\frac{9}{16} = 0.25+0.5625 = 0.8125$; $\sqrt{0.8125} \approx 0.901$. (3) $t = \frac{6.0-7.5}{0.901} = \frac{-1.5}{0.901} \approx -1.66$. (4) Since the p-value $\approx 0.12 > 0.05$, we *fail to reject* H_0: not enough evidence that the two designs have different mean wait times.

Section 10.2: (1) Paired — the *same* students were measured before and after, so each before-value links to one after-value. (2) $\bar{d} = \frac{4+6+3+5+2}{5} = \frac{20}{5} = 4$. (3) $t = \frac{4}{1.581/\sqrt{5}} = \frac{4}{1.581/2.236} = \frac{4}{0.707} \approx 5.66$. (4) Since p $\approx 0.005 < 0.05$, we *reject* H_0: significant evidence the study session improved scores (about 4 points on average).

Section 10.3: (1) $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$. (2) $\hat{p}_1 = \frac{60}{80} = 0.75$, $\hat{p}_2 = \frac{50}{80} = 0.625$, pooled $\hat{p} = \frac{60+50}{80+80} = \frac{110}{160} = 0.6875$. (3) SE $= \sqrt{0.6875(0.3125)\left(\frac{1}{80}+\frac{1}{80}\right)} = \sqrt{0.214844 \times 0.025} = \sqrt{0.005371} \approx 0.0733$; $z = \frac{0.75-0.625}{0.0733} = \frac{0.125}{0.0733} \approx 1.71$. (4) Since p $\approx 0.09 > 0.05$, we *fail to reject*

H_0: not quite enough evidence to call the recovery rates different at the 0.05 level.

Section 10.4: (1) $H_0 : \sigma_1^2 = \sigma_2^2$. (2) $s_1^2 = 8^2 = 64$, $s_2^2 = 4^2 = 16$. (3) $F = \dfrac{64}{16} = 4$. (4) An F well above 1 suggests Grader 1's scores are much more variable (less consistent) than Grader 2's.

Chapter 11

When Three or More Groups Disagree

Analysis of Variance (ANOVA) and Post-Hoc Tests

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll be able to compare three, four, or even five groups *at once* — in a single test — instead of squinting at a pile of separate comparisons and hoping you didn't get fooled.

The Story. You already know how to compare *two* groups (that was the t-test). But life rarely gives you exactly two. Three coffee brands. Four study methods. Five neighborhoods. ANOVA is the tool for asking, "Do these groups really differ, or is this just noise?" — without cheating.

The Skills You'll Have:

- Explain why running lots of t-tests is a trap
- State the ANOVA hypotheses correctly (and avoid the classic mistake)
- Read an ANOVA summary table and find the F value and p-value
- Use a post-hoc test to say *which* groups actually differ

The Confidence Moment. When you see "$F(2, 27) = 8.4$, $p = 0.001$," your brain will calmly say: "The groups vary more between each other than within themselves — something real is going on."

The Bridge to Part VI. So far we've mostly compared *numerical* measurements across groups. Next we turn to *categorical* data and *relationships* between variables — counts, tables, and the chi-square test. But first, let's finish the story of comparing means.

MATH SKILLS YOU'LL NEED

ANOVA sounds intimidating, but the arithmetic underneath is stuff you already do: averaging, squaring, and reading a value off a table. Let's warm it up.

Skill 1 — Averaging a small group. Add the values, divide by how many there are.

$$\frac{6+8+10}{3} = \frac{24}{3} = 8$$

Skill 2 — Squaring a difference. "Square" just means multiply a number by itself. Differences get squared so that negatives don't cancel positives.

$$(7-4)^2 = 3^2 = 9 \qquad (2-5)^2 = (-3)^2 = 9$$

Skill 3 — Reading an F value. An F value is just one number — always zero or positive. Bigger F means "the groups look more different." You don't compute it by hand in this book; you read it off the output and ask whether the p-value is small.

Practice (answers at the end of the chapter):

1. Find the average of $4, 9, 5$.
2. Compute $(10-6)^2$.
3. An output says $F = 0.3$, $p = 0.74$. Roughly, do the groups look very different or pretty similar?

11.1 Why Not Just Run Many t-Tests?

▪ READ THIS FIRST

Imagine you're testing whether a coin is fair by flipping it ten times. One "weird" streak wouldn't shock you — weird streaks happen by chance. Now imagine you do that same ten-flip experiment *twenty times in a row*. Somewhere in those twenty runs, you'll almost certainly hit a strange-looking streak, purely by luck.

The same trap hides in comparing groups. Each single comparison has a small chance of fooling you. Do *many* comparisons, and the chance that *at least one* fools you stops being small.

▪ LET'S TALK ABOUT IT

Say you have four groups and you compare every possible pair. How many comparisons is that? (Try listing them: 1&2, 1&3, 1&4, 2&3, 2&4, 3&4.) If each comparison carries its own little risk of a false alarm, does doing six of them feel safer or riskier than doing one?

▪ NOW WE NAME IT

Every hypothesis test carries a **Type I error** risk — the chance of a false alarm, declaring a difference that isn't really there. We usually set that risk at $\alpha = 0.05$ (a 5% chance) *per test*.

Here's the problem. If you run many separate t-tests on the same groups, each one gets its own 5% chance to misfire. The *overall* chance that at least one of them gives a false alarm — the **family-wise error rate** — climbs well above 5%. Six comparisons can push the real false-alarm risk past 25%. That's not a careful analysis anymore; that's fishing.

Analysis of Variance (ANOVA) solves this. Instead of many pairwise tests, it asks *one* question about *all* the group means together, keeping the overall error rate at 5%.

▪ WATCH IT WORK

Question: A researcher compares the average delivery time of three pizza shops, A, B, and C. Why not just run three t-tests (A vs. B, A vs. C, B vs. C)?

Step 1 — Count the tests. Three groups give three pairwise comparisons.

Step 2 — Track the risk. Each test has a 5% false-alarm chance. Across three tests, the chance that *at least one* misfires is roughly $1 - (0.95)^3 \approx 0.14$, or about 14% — nearly triple the 5% we intended.

Step 3 — Choose the honest tool. Run *one* ANOVA across all three shops. It holds the overall false-alarm risk at 5% while still testing whether *any* shop differs.

The more groups you have, the worse the pile-of-t-tests problem gets — and the more ANOVA earns its keep.

▪ NOW WE NAME IT

The hypotheses for ANOVA, with k groups, are:

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{(all the group means are equal)}$$

$$H_a: \ \text{at least one group mean is different.}$$

Read that alternative carefully. H_a does **not** say "all the means are different." It says *at least one* is different — maybe one oddball stands apart while the rest are alike. This is the single most common mistake students make, and now it won't be yours.

▪ YOUR TURN

A gym compares average weekly visits across four membership tiers: Bronze, Silver, Gold, and Platinum.

1. How many groups is k?
2. Write H_0 in words.
3. True or false: H_a claims all four tiers have different averages.

Work space:

▪ CHECK YOUR VOICE

If "family-wise error rate" made your shoulders tense — breathe. It's just a fancy name for an everyday idea: *the more times you roll the dice, the more likely a fluke shows up.* You already understood that from the coin-flip story. The vocabulary is new; the intuition was always yours.

11.2 One-Way ANOVA: The F-Test

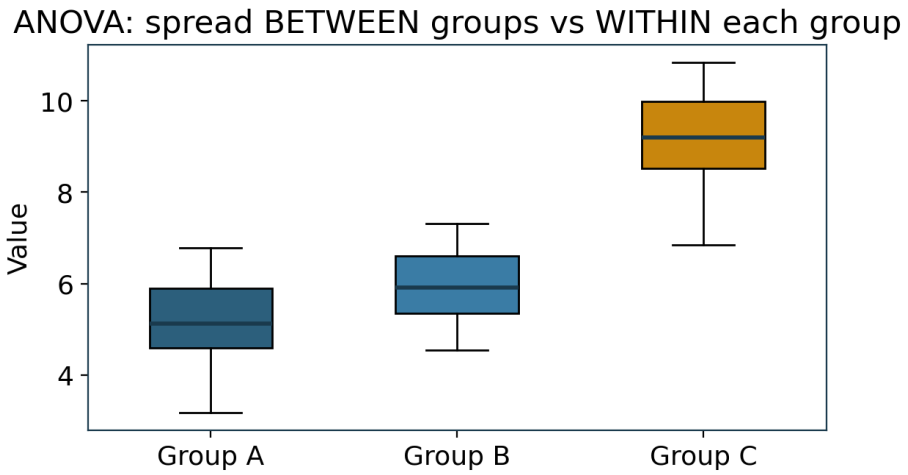


Figure 11.1. ANOVA compares the spread *between* group means to the spread *within* each group.

▪ READ THIS FIRST

Picture three archery teams, each shooting at their own target. To decide whether the teams are truly different in skill, you'd look at two things: how far apart the *teams'* average shots land from each other, and how scattered each *single team's* shots are around their own average.

If the teams' averages are far apart *compared to* how scattered each team is, the teams really differ. If the gap between teams is no bigger than the wobble within a team, the difference is probably just noise.

▪ LET'S TALK ABOUT IT

Two teams have averages 8 and 9 (on a 10-point target). Does that one-point gap mean a lot if each team's shots are tightly clustered within half a point of their average? What if each team's shots are scattered wildly across five points? Same gap — different conclusion. Why?

▪ NOW WE NAME IT

ANOVA compares two kinds of spread:

- **Between-group variation** — how far the group averages sit from one another (the gap between the teams).
- **Within-group variation** — how scattered the values are *inside* each group (the wobble within a team).

We turn each kind of spread into a single number called a *mean square* (MS), where MS $=$ SS$/df$ (a sum of squared differences, divided by its degrees of freedom). Then the test statistic is their ratio:

$$F = \frac{\text{MS}_{between}}{\text{MS}_{within}}.$$

The degrees of freedom, with k groups and N total observations, are:

$$df_{between} = k - 1, \qquad df_{within} = N - k.$$

A **big** F means the between-group spread is large compared to the within-group spread — the groups differ more than random noise would explain. That gives a *small* p-value, and we reject H_0. A small F (near 1) means the groups look about as different as random wobble, so we don't reject H_0.

▪ WATCH IT WORK

Question: Three study methods (Flashcards, Group Study, Rewriting Notes) are each tried by 10 students, so $N = 30$ and $k = 3$. We measure each student's quiz score. Here is the ANOVA summary table the software produced. Find F and interpret the p-value (use $\alpha = 0.05$).

Source	SS	df	MS	F
Between groups	120	2	60	9.0
Within groups	180	27	6.667	
Total	300	29		

Step 1 — Check the degrees of freedom. $df_{between} = k - 1 = 3 - 1 = 2$. $df_{within} = N - k = 30 - 3 = 27$. They add to $29 = N - 1$, the total df. Good.

Step 2 — Find each mean square. $\text{MS}_{between} = \text{SS}/df = 120/2 = 60$. $\text{MS}_{within} = 180/27 \approx 6.667$.

Step 3 — Compute F. $F = \dfrac{\text{MS}_{between}}{\text{MS}_{within}} = \dfrac{60}{6.667} \approx 9.0$.

Step 4 — Get the p-value. The software reports $p = 0.001$ for $F(2, 27) = 9.0$. That's the chance of seeing group differences this large if all three methods were truly equal.

Step 5 — Compare to α. $p = 0.001 < 0.05$, so we reject H_0.

Step 6 — Say it in plain words. "The three study methods do *not* all produce the same average quiz score — at least one method differs." Notice what we *can't* say yet: which one. That's the next section.

▪ YOUR TURN

An ANOVA compares average wait times at three clinics, with 12 patients timed at each clinic ($N = 36$, $k = 3$). The output reads $F(2, 33) = 0.45$, $p = 0.64$. Use $\alpha = 0.05$.

1. Confirm the degrees of freedom: what are $df_{between}$ and df_{within}?
2. Is F big or small? What does that say about between- vs. within-group spread?
3. Do you reject H_0? Write the conclusion in plain words.

Work space:

▪ CHECK YOUR VOICE

The word "variance" in the name scares people, but you just *used* ANOVA without computing a single sum of squares by hand — you read a table and divided. That's exactly how it works in real life: software fills the table, and your job is to understand the $F =$ between-over-within ratio and check the p-value. You did that. That's the skill.

11.3 Which Groups Differ? Post-Hoc Tests

▪ READ THIS FIRST

A smoke alarm going off tells you there's smoke *somewhere* in the house — it doesn't tell you which room. A significant ANOVA is exactly that alarm: it says "at least one group differs," but it won't point to which pair. To find the room, you need to walk through and check.

That walk-through is what a post-hoc test does — carefully, one pair at a time, without re-creating the many-tests problem from Section 11.1.

▪ LET'S TALK ABOUT IT

If your three study methods came back significant, which comparisons would you actually want to know about? Flashcards vs. Group Study? Flashcards vs. Rewriting? All of them? And why would it be a bad idea to go check those pairs only *because* the ANOVA was already significant — isn't that the safe order?

▪ NOW WE NAME IT

A significant ANOVA tells you "something differs," not "what differs." To find out, you run a **post-hoc test** — a follow-up that compares each pair of group means while *controlling the overall (family-wise) error rate*, so all those pairwise looks together still stay at about 5%.

The most common one is **Tukey's HSD** (Honestly Significant Difference). It checks every pair and flags which ones differ "honestly" — meaning the gap is bigger than what its built-in safety margin allows for chance.

Two rules to hold onto:

- Only run post-hoc tests *after* a significant ANOVA. If the ANOVA isn't significant, there's no alarm — don't go hunting through rooms.
- Post-hoc tests, not a fresh pile of plain t-tests, are the honest way to find *which* pairs differ.

▪ WATCH IT WORK

Question: The study-methods ANOVA was significant ($p = 0.001$), so we run Tukey's HSD. Here are the pairwise results. Which methods differ? Use $\alpha = 0.05$.

Comparison	Mean difference	p (Tukey)
Flashcards vs. Group Study	1.0	0.62
Flashcards vs. Rewriting Notes	5.0	0.001
Group Study vs. Rewriting Notes	4.0	0.004

Step 1 — We earned the right. The ANOVA was significant, so running post-hoc is appropriate.

Step 2 — Scan the p-values against $\alpha = 0.05$. Flashcards vs. Group Study: $p = 0.62 > 0.05$ — not significant. Flashcards vs. Rewriting: $p = 0.001 < 0.05$ — significant. Group Study vs. Rewriting: $p = 0.004 < 0.05$ — significant.

Step 3 — Identify the pattern. Flashcards and Group Study are statistically *alike*. Rewriting Notes differs from *both* of the others.

Step 4 — Read the direction. The mean differences are positive for the Rewriting comparisons, meaning Rewriting Notes scored lower (it's being subtracted), so Flashcards and Group Study both beat Rewriting Notes.

Step 5 — Plain-words conclusion. "Students using Flashcards or Group Study scored similarly to each other, and both scored significantly higher than students who only Rewrote their Notes."

Step 6 — Mind the limits. Tukey told us *which* pairs differ and kept the overall error at 5%. It does not prove Rewriting *causes* lower scores — that depends on how the study was designed (Chapter 2's question).

▪ YOUR TURN

After a significant ANOVA on three coffee brands' average customer ratings, Tukey's HSD reports:

Comparison	Mean difference	p (Tukey)
Brand A vs. Brand B	0.2	0.81
Brand A vs. Brand C	1.8	0.002
Brand B vs. Brand C	1.6	0.006

1. Which pairs differ significantly at $\alpha = 0.05$?
2. Which two brands look statistically alike?
3. Write a one-sentence plain-words summary of the pattern.

Work space:

▪ CHECK YOUR VOICE

If you noticed that the post-hoc step felt like "just comparing p-values to 0.05 again" — yes, exactly. The hard, clever part (keeping the overall error honest) is done *for* you inside Tukey's method. You're not missing some secret difficulty. Reading the output *is* the skill, and you have it.

11.4 Conditions for ANOVA

▪ READ THIS FIRST

Before you trust the soup from one spoonful, you stir the pot — some conditions have to be right for the taste to mean anything. ANOVA is the same. The F-test gives a trustworthy p-value only when a few reasonable conditions hold. They're gentle, not scary.

▪ LET'S TALK ABOUT IT

Think about the three pizza shops from earlier. If shop C only let you time *the same one driver* over and over, would those times really be independent measurements? What might go wrong?

▪ NOW WE NAME IT

ANOVA leans on three conditions:

- **Independence** — the observations within and across groups don't influence each other (separate people, separate measurements). This is the one you protect by *how you collect data*.
- **Approximate normality** — within each group, the values are roughly bell-shaped, with no wild skew. ANOVA is fairly forgiving here, especially with larger groups.
- **Similar spreads** — the groups have roughly equal variability (their within-group scatter is comparable). A common rule of thumb: the largest group's standard deviation shouldn't be more than about twice the smallest's.

If the spreads are wildly different or a group is badly skewed and tiny, the p-value can't fully be trusted — and there are adjusted versions of the test for those cases. For this course, check that the conditions are *reasonably* met, then proceed.

▪ WATCH IT WORK

Question: A study compares test scores across three classrooms (30 students each, all surveyed separately). The group standard deviations are 4.1, 4.5, and 3.9, and each group's scores look roughly bell-shaped. Are the ANOVA conditions reasonably met?

Step 1 — Independence. Students were measured separately, one score each, across distinct classrooms. Reasonable.

Step 2 — Normality. Each group's scores are roughly bell-shaped, and with 30 per group ANOVA is forgiving. Reasonable.

Step 3 — Similar spreads. Largest SD is 4.5, smallest is 3.9. Since 4.5 is far less than $2 \times 3.9 = 7.8$, the spreads are comfortably similar.

Step 4 — Verdict. All three conditions are reasonably met, so the ANOVA p-value can be trusted.

▪ YOUR TURN

A researcher runs ANOVA on reaction times for three age groups. The group standard deviations are 0.20, 0.22, and 0.55 seconds, and the third group's times are strongly right-skewed with only 8 people.

1. Which condition looks shaky here?
2. Does 0.55 exceed twice the smallest SD?
3. Should the researcher fully trust the p-value as-is, or be cautious?

Work space:

▪ CHECK YOUR VOICE

Checking conditions can feel like busywork before the "real" test — but it's the opposite. It's the move that lets you *trust* your answer instead of just hoping. Slowing down to look at spreads and shape is a sign of a careful statistician, which is exactly what you're becoming.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 11 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 11 — Your Turn Answers

Math Skills: (1) $\frac{4+9+5}{3} = \frac{18}{3} = 6$. (2) $(10-6)^2 = 4^2 = 16$. (3) $F = 0.3$ is small and $p = 0.74$ is large, so the groups look pretty similar (no real difference).

Section 11.1: (1) $k = 4$ groups. (2) H_0: the average weekly visits are equal across all four tiers ($\mu_1 = \mu_2 = \mu_3 = \mu_4$). (3) *False* — H_a only claims at least one tier's average differs, not that all four differ.

Section 11.2: (1) $df_{between} = k - 1 = 3 - 1 = 2$; $df_{within} = N - k = 36 - 3 = 33$. (2) $F = 0.45$ is small (near or below 1), so the between-group spread is no larger than the within-group spread. (3) $p = 0.64 > 0.05$, so do *not* reject H_0: there's no evidence the three clinics have different average wait times.

Section 11.3: (1) Brand A vs. Brand C ($p = 0.002$) and Brand B vs. Brand C ($p = 0.006$) differ significantly. (2) Brand A and Brand B look statistically alike ($p = 0.81$). (3) Example: "Brands A and B were rated similarly, and both were rated significantly higher than Brand C."

Section 11.4: (1) The "similar spreads" condition (and normality for the small, skewed third group). (2) Yes — $0.55 > 2 \times 0.20 = 0.40$, so the largest SD exceeds twice the smallest. (3) Be cautious: with unequal spreads and a small, skewed group, the standard p-value isn't fully trustworthy; use a Welch-adjusted ANOVA instead.

Part V Checkpoint — Study Guide & Practice (Hypothesis Testing)

This Study Guide gathers everything from Chapters 9–11 and lets you rehearse hypothesis testing the way an exam will actually feel.

STUDY GUIDE & PRACTICE

You've just built the engine of inferential statistics: putting a claim on trial, weighing the evidence, and delivering a verdict you can defend. This isn't a test — it's a dress rehearsal. We'll pull the big ideas into one place, check the vocabulary, then run twelve mixed problems that jump across all three chapters, exactly the way a real exam does. When your inner voice says "I forgot how to do all of this," keep going anyway — the answers are right here, fully worked, waiting for you.

The Big Ideas, in One Place

Chapter 9 — Testing One Parameter.

- Every test has two competing claims. The **null hypothesis** H_0 is the status-quo "presumed innocent" claim and *always* contains an equals sign ($H_0 : \mu = \mu_0$, $H_0 : p = p_0$). The **alternative hypothesis** H_a is what we'd need evidence to believe, and it uses $\neq$, $<$, or $>$.
- We *assume* H_0 *is true*, then measure surprise. The **p-value** is the probability of data as extreme as ours (or more) *if* H_0 *were true* — an area in the tail(s), always between 0 and 1.
- **The decision rule.** Pick α first (usually 0.05). If p-value $\leq \alpha$, *reject* H_0; if p-value $> \alpha$, *fail to reject* H_0. Memory hook: **small p, reject.** "Fail to reject" never means "H_0 is proven true" — only "not enough evidence."
- A **Type I error** is rejecting a true H_0 (convicting the innocent); its probability is α. A **Type II error** is failing to reject a false H_0 (freeing the guilty).
- Three one-parameter tests, all the same shape (observed $-$ claimed, over a standard error):

- One mean (t-test): $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$, with $df = n - 1$.
- One proportion (z-test): $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ — the standard error uses the *claimed* p_0.
- One variance (χ^2-test): $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$, with $df = n - 1$ — the tool when the claim is about *spread*.

Chapter 10 — Comparing Two Parameters.

- **Two independent means** (two separate groups) use **Welch's t**: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$, testing $H_0 : \mu_1 = \mu_2$. For hand work, use the conservative $df = \min(n_1 - 1,\, n_2 - 1)$.
- **Paired means** (the same subjects measured twice, or matched pairs) collapse to a one-sample t on the differences d: $t = \dfrac{\bar{d} - 0}{s_d/\sqrt{n}}$, with $df = n - 1$ and $n =$ number of pairs, testing $H_0 : \mu_d = 0$.
- **The paired-vs-independent distinction is about design, not numbers.** Ask: "Can I draw a line connecting each value on the left to one specific value on the right?" Same person twice or matched pairs → *paired*. Two unrelated piles of subjects → *independent*.
- **Two proportions** use a z with a **pooled proportion** $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ (pool the *counts*, never average the two percents): $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, testing $H_0 : p_1 = p_2$.
- **Two variances** use the **F-test**, the ratio of the two sample variances with the larger on top: $F = \dfrac{s_1^2}{s_2^2}$, testing $H_0 : \sigma_1^2 = \sigma_2^2$. If the spreads were equal, F sits near 1.

Chapter 11 — Three or More Groups (ANOVA).

- **The multiple-comparisons problem.** Running many separate t-tests piles up false-alarm risk: each carries its own α, so the overall **family-wise error rate** climbs well above 5%. ANOVA asks *one* question about all the means at once and holds the overall risk at 5%.
- **One-way ANOVA** compares **between-group variation** to **within-group variation**: $F = \dfrac{\text{MS}_{between}}{\text{MS}_{within}}$, where each MS $= \text{SS}/df$. The degrees of freedom with k groups and N total observations are $df_{between} = k - 1$ and $df_{within} = N - k$.
- The hypotheses are $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus H_a: "*at least one* group mean differs."

H_a does **not** say all means differ — that's the classic mistake.

- A **big** F means between-group spread dwarfs within-group spread $\rightarrow$ small p-value $\rightarrow$ reject H_0. A significant ANOVA tells you *that* something differs, not *which* groups differ.
- **Post-hoc** tests (most commonly **Tukey's HSD**) find *which* pairs differ while keeping the family-wise error near 5%. Run them *only after* a significant ANOVA — no alarm, no room-to-room search.

Vocabulary Check

Can you say each of these in one plain sentence? If one feels fuzzy, that's exactly the one to reread — not a sign you're behind.

- null hypothesis H_0, alternative hypothesis H_a, one-sided vs. two-sided
- p-value, significance level α, the decision rule (p $\leq \alpha \rightarrow$ reject)
- Type I error, Type II error
- one-mean t-test, one-proportion z-test, one-variance χ^2-test
- two independent means (Welch's t), standard error of a difference
- paired (matched) data vs. independent samples, $\bar{d}$, paired t-test
- two-proportion z-test, pooled proportion $\hat{p}$
- two-variance F-test
- multiple-comparisons problem, family-wise error rate
- between-group variation, within-group variation, mean square (MS)
- one-way ANOVA, $F = \text{MS}_{between}/\text{MS}_{within}$, $df_{between} = k - 1$, $df_{within} = N - k$
- post-hoc test, Tukey's HSD

Mixed Practice

These twelve problems jump around on purpose — just like an exam. Try each one before peeking. The full worked answers come right after.

1. A coffee shop claims its drive-through mean service time is $\mu = 4$ minutes. A regular suspects it's actually *longer*. Write H_0 and H_a in symbols, and say whether the test is one-sided or two-sided.

2. A bakery advertises that its loaves weigh $\mu = 500$ grams on average. A customer thinks they're *lighter*. A sample of $n = 16$ loaves has $\bar{x} = 494$ grams and $s = 12$ grams. Using $\alpha = 0.05$, compute the one-sample t, state df, and decide via the p-value (the left-tail area for $t = -2.0$ at $df = 15$ is about 0.032).

3. A politician claims $p = 0.60$ of voters support a measure. A pollster suspects the true support is *different* from 0.60. In a sample of $n = 100$ voters, 51 support it. Compute $\hat{p}$ and the one-proportion z, then decide at $\alpha = 0.05$ (the two-tailed p-value for $z \approx -1.84$ is about 0.066).

4. For each scenario, decide whether the data are *paired* or *independent*, and say why: (a) A trainer records each of 12 runners' mile times before a training block and again after. (b) A researcher compares mile times of 12 runners on Team Red against a *different* 12 runners on Team Blue.

5. Seven patients have their blood pressure measured before and after a new medication. The drops (before $-$ after) are: $8, 5, 12, 3, 7, 6, 10$. The mean drop is $\bar{d} = 7.286$ and $s_d \approx 3.039$. Using $\alpha = 0.05$, test $H_0 : \mu_d = 0$ against $H_a : \mu_d \neq 0$. Compute t, state df, and decide (the two-tailed p-value for $t \approx 6.34$ at $df = 6$ is about 0.0007).

6. A company tests two landing pages. Of $n_1 = 120$ visitors to Page A, $x_1 = 54$ signed up; of $n_2 = 130$ visitors to Page B, $x_2 = 39$ signed up. At $\alpha = 0.05$, test whether the sign-up rates differ. Compute $\hat{p}_1$, $\hat{p}_2$, the pooled $\hat{p}$, the standard error, and z, then decide (the two-tailed p-value for $z \approx 2.45$ is about 0.014).

7. A nutritionist compares the mean daily calorie intake across four diet plans, with 20 people on each plan ($N = 80$, $k = 4$). The ANOVA table shows $\text{SS}_{between} = 1800$ and $\text{SS}_{within} = 6840$. Fill in the degrees of freedom, compute both mean squares and F, and decide at $\alpha = 0.05$ (the software reports $p \approx 0.0005$).

8. An ANOVA comparing average commute times across three neighborhoods ($N = 45$, $k = 3$) reports $F(2, 42) = 0.78$, $p = 0.46$. Use $\alpha = 0.05$. Confirm the degrees of freedom, and state your decision and conclusion in plain words.

9. In your own words, what does a *significant* ANOVA tell you — and what does it *not* tell you?

10. A study compares average sleep hours across five different work shifts and gets a significant ANOVA ($p = 0.003$). A colleague says, "Great, now just run a t-test on each pair of shifts to see which differ." What's the better approach, and why is running a fresh pile of t-tests a bad idea?

11. A researcher runs an ANOVA across three teaching methods and gets $F(2, 57) = 1.20$, $p = 0.31$ at $\alpha = 0.05$. She wants to run Tukey's HSD to find which methods differ. Should she? Explain.

12. Classify the correct test for each claim: (a) "The mean cholesterol of patients differs from 200." (b) "The proportion of left-handed students differs between two schools." (c) "The average test score is the same across four tutoring centers." (d) "The same students' anxiety scores changed from before to after therapy."

Answers

Worked out in full — read these even for the ones you got right, because the *reasoning* is what the exam rewards.

1. $H_0 : \mu = 4$ versus $H_a : \mu > 4$. Because "longer" points in one direction (greater than), this is a *one-sided* test. (The null always carries the equals sign; the suspicion drives the direction of H_a.)

2. **Hypotheses:** $H_0 : \mu = 500$ versus $H_a : \mu < 500$ (one-sided, "lighter"). The test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{494 - 500}{12/\sqrt{16}} = \frac{-6}{12/4} = \frac{-6}{3} = -2.0, \qquad df = n - 1 = 15.$$

   Because H_a is "less than," the p-value is the left-tail area: $p \approx 0.032$. Since $0.032 \leq 0.05$, *reject* H_0. There is significant evidence (at $\alpha = 0.05$) that the true mean loaf weight is less than 500 grams.

3. **Hypotheses:** $H_0 : p = 0.60$ versus $H_a : p \neq 0.60$ (two-sided, "different"). First $\hat{p} = \dfrac{51}{100} = 0.51$. Then

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.51 - 0.60}{\sqrt{(0.60)(0.40)/100}} = \frac{-0.09}{\sqrt{0.0024}} = \frac{-0.09}{0.04899} \approx -1.84.$$

   The two-tailed p-value is about 0.066. Since $0.066 > 0.05$, *fail to reject* H_0. There is not enough evidence (at $\alpha = 0.05$) that true support differs from 60%.

4. (a) *Paired* — the *same* 12 runners are measured twice (before and after), so each before-time links to one specific after-time. (b) *Independent* — Team Red and Team Blue are two *different* sets of runners with no person-to-person link between the groups. The line test settles it: in (a) you can connect each runner's two times; in (b) you can't.

5. **Hypotheses:** $H_0 : \mu_d = 0$ (no average change) versus $H_a : \mu_d \neq 0$. This is *paired* (same patients before and after), so we test the differences as one sample:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{7.286}{3.039/\sqrt{7}} = \frac{7.286}{3.039/2.646} = \frac{7.286}{1.149} \approx 6.34, \qquad df = n - 1 = 6.$$

   The two-tailed p-value is about 0.0007. Since $0.0007 \leq 0.05$, *reject* H_0. There is significant evidence the medication changed blood pressure — on average it dropped about 7.3 points.

6. **Hypotheses:** $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$ (two-sided). Sample proportions: $\hat{p}_1 = \frac{54}{120} = 0.45$ and $\hat{p}_2 = \frac{39}{130} = 0.30$. Pool the *counts*:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{54 + 39}{120 + 130} = \frac{93}{250} = 0.372.$$

   Standard error:

$$\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.372(0.628)\left(\frac{1}{120} + \frac{1}{130}\right)}$$

$$= \sqrt{0.233616 \times 0.016026} \approx \sqrt{0.003744} \approx 0.06119.$$

   Then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{0.06119} = \frac{0.45 - 0.30}{0.06119} = \frac{0.15}{0.06119} \approx 2.45.$$

   The two-tailed p-value (for $z \approx 2.45$, with one-tail area about 0.0071, doubled) is about 0.014. Since this is ≤ 0.05, *reject* H_0. There is significant evidence the two pages have different sign-up rates — Page A's is higher.

7. Degrees of freedom: $df_{between} = k - 1 = 4 - 1 = 3$ and $df_{within} = N - k = 80 - 4 = 76$ (they add to $79 = N - 1$). Mean squares:

$$\mathsf{MS}_{between} = \frac{\mathsf{SS}_{between}}{df_{between}} = \frac{1800}{3} = 600, \qquad \mathsf{MS}_{within} = \frac{\mathsf{SS}_{within}}{df_{within}} = \frac{6840}{76} = 90.$$

   Then

$$F = \frac{\mathsf{MS}_{between}}{\mathsf{MS}_{within}} = \frac{600}{90} \approx 6.67.$$

The software reports $p \approx 0.0005$. Since $0.0005 \leq 0.05$, *reject* H_0. There is significant evidence that the four diet plans do *not* all produce the same mean calorie intake — at least one differs. (Notice: we can't yet say *which*; that needs a post-hoc test.)

8. Degrees of freedom: $df_{between} = k - 1 = 3 - 1 = 2$ and $df_{within} = N - k = 45 - 3 = 42$ — matching the reported $F(2, 42)$. Here $F = 0.78$ is small (near 1), meaning the between-group spread is no larger than the within-group spread. Since $p = 0.46 > 0.05$, *fail to reject* H_0. There is no evidence the three neighborhoods have different mean commute times.

9. A significant ANOVA tells you that *at least one* group mean differs from the others — the groups are not all equal. It does **not** tell you *which* group (or groups) differ, by how much, or in which direction. It's a smoke alarm: it says there's smoke somewhere, not which room. To find which pairs differ, you follow up with a post-hoc test like Tukey's HSD.

10. The better approach is a *post-hoc* test (such as Tukey's HSD), run only because the ANOVA was significant. Running a fresh pile of plain t-tests on every pair is exactly the *multiple-comparisons problem* ANOVA was built to avoid: with five shifts there are 10 pairwise comparisons, each with its own 5% false-alarm risk, so the family-wise error rate balloons far above 5% and you'll likely flag a "difference" that's just chance. Tukey's HSD compares the pairs while holding the overall error rate near 5%.

11. No — she should *not* run Tukey's HSD. Post-hoc tests are only appropriate *after* a significant ANOVA. Here $p = 0.31 > 0.05$, so we fail to reject H_0: there's no evidence any method differs. With no alarm sounding, there's no room to go searching. Running post-hoc tests anyway would be hunting for differences the overall test didn't support.

12. (a) One-mean *t-test* (a single mean against a claimed value 200). (b) Two-proportion *z-test* (comparing a "yes" rate across two independent schools). (c) One-way *ANOVA* (F-test) (comparing means across four groups at once). (d) Paired *t-test* (the *same* students measured before and after — paired data, so test the differences $\bar{d}$).

You just rehearsed every big idea from Part V — writing hypotheses, running one-sample t, z, and χ^2 tests, telling paired from independent, comparing two means and two proportions, reading an ANOVA table, and knowing when (and when not) to reach for a post-hoc test. If some were shaky, now you know exactly where to look. That's not failing the rehearsal; that's what the rehearsal is *for*. You're ready.

Part V — Formula Sheet: Testing Claims

Every hypothesis test from Part V on one page (p-value approach), each symbol named in plain words with a note on when *to use it.*

FORMULA SHEET

The logic (every test): state H_0 (the "$=$" / no-effect claim) and H_a ($\neq$, $<$, or $>$); compute the test statistic; find the **p-value** (the probability of a result this extreme *if* H_0 *is true*); **if p-value** $\leq \alpha$**, reject** H_0; otherwise fail to reject. Always state the conclusion in context. "Fail to reject" never *proves* H_0.

ONE population parameter (Chapter 9)

mean: $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1$$

proportion: $$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

variance: $$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}, \quad df = n - 1$$

TWO population parameters (Chapter 10)

two independent means: $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

paired means: $$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad df = n - 1$$

two proportions: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ two variances: $F = \dfrac{s_1^2}{s_2^2}$

When: **paired** $=$ the same (or matched) subjects measured twice — test the differences d. **Independent** $=$ two separate groups.

THREE OR MORE means — one-way ANOVA (Chapter 11)

$$F = \frac{\text{MS}_{between}}{\text{MS}_{within}}, \quad \text{MS} = \frac{\text{SS}}{df}, \quad df_{between} = k - 1, \quad df_{within} = N - k$$

When: comparing the means of $k \geq 3$ groups (N = total observations). $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$; H_a: *at least one* mean differs. A small p-value says "at least one differs" — then a **post-hoc test (Tukey's HSD)** finds *which* pairs differ, while controlling the overall error rate. Only run post-hoc after a significant ANOVA.

Part VI

Categorical Data & Relationships

Chapter 12

Counts, Categories, and Surprises

Chi-Square Tests — Goodness-of-Fit, Independence, and Homogeneity

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll look at a pile of *counts* — how many reds, how many "yes" answers, how many people in each box of a table — and calmly ask: "Is this what I'd expect if nothing weird were going on? And if not, how surprised should I be?"

The Story. Every test in this chapter does the exact same thing: it compares the counts you *actually* got to the counts you'd *expect*, and measures the gap. One formula, one idea, three situations. If you learn the idea once, you've learned all three.

The Skills You'll Have:

- Compute *expected counts* and the chi-square statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$
- Run a *goodness-of-fit* test (one variable vs. a claimed distribution)
- Run a *test of independence* (two variables, one sample)
- Run a *test of homogeneity* and explain how it differs from independence

The Confidence Moment. When you see a two-way table of counts, you'll think: "I know what to do — find expected counts, square the gaps, divide, add them up, read the p-value." Same recipe every time.

The Bridge. Chi-square tests handle *categorical* data — labels and counts. Chapter 13 turns to *quantitative* relationships with regression, asking not "are these categories associated?" but "how does one number predict another?" Different data, same spirit: looking for a real pattern instead of random noise.

MATH SKILLS YOU'LL NEED

The only arithmetic in this whole chapter is the inside of one formula: $\dfrac{(O-E)^2}{E}$. Let's warm up the three little moves it asks for.

Skill 1 — The $(O-E)^2/E$ move. Subtract, square the result, then divide by E. Do it left to right, one step at a time.

$$O = 18,\ E = 12: \quad (18-12)^2 = 6^2 = 36, \quad \frac{36}{12} = 3$$

Skill 2 — Adding the pieces up. Once you have one $(O-E)^2/E$ for each category, the chi-square statistic is just their sum. Nothing fancy — you add a short list of numbers.

$$\chi^2 = 3 + 0.5 + 1.5 = 5$$

Skill 3 — Expected counts as products and quotients. An expected count is always either $n \times$ (a proportion) or a row total times a column total, divided by the grand total. Both are just multiply-then-(maybe)-divide.

$$100 \times 0.30 = 30 \qquad \frac{(40)(60)}{200} = \frac{2400}{200} = 12$$

Practice (answers at the end of the chapter):

1. Compute $\dfrac{(O-E)^2}{E}$ when $O = 25$ and $E = 20$.
2. A bag has $n = 80$ candies. The company claims 25% are red. What is the expected count of red candies?
3. In a table, a cell's row total is 50, its column total is 90, and the grand total is 150. Find the expected count for that cell.

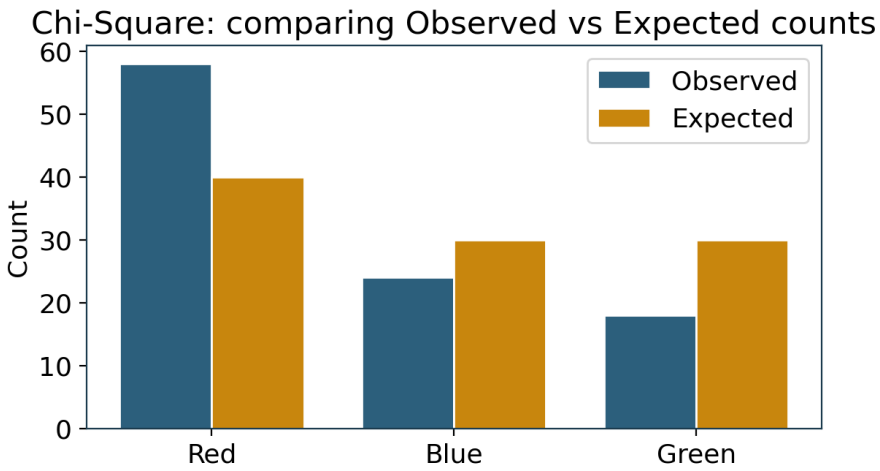


Figure 12.1. Chi-square measures how far the observed counts fall from what we'd expect.

12.1 Observed vs. Expected: The Chi-Square Idea

▪ READ THIS FIRST

Imagine you buy a big bag of candy and the wrapper promises "equal amounts of every color." You pour it out and there are way more reds than anything else, and almost no blues. You squint and think: "That doesn't look equal."

You just did a chi-square test in your head. You had counts you *expected* (roughly equal colors) and counts you *observed* (a flood of red). The bigger the gap between expected and observed, the more your eyebrow goes up.

That single instinct — "how far is what I got from what I expected?" — is the entire chapter. Everything else is just putting a number on the eyebrow-raise.

▪ LET'S TALK ABOUT IT

Think of a time something "felt off" just from the counts — a die that seemed to land on six too often, a class that seemed to have way more of one major than you'd guess, a playlist that kept "shuffling" to the same artist. What were you secretly comparing? (You were comparing what happened to what you expected. That comparison has a name now.)

▪ NOW WE NAME IT

Chi-square tests work with **observed** counts — written O — which are simply how many cases actually landed in each category. We compare them to **expected** counts — written E — the counts we would expect if the null hypothesis were true.

The **chi-square statistic** measures the total gap:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

For each category: subtract $(O-E)$, square it (so over- and under-shooting both count as positive), divide by E (so a gap of 5 matters more when you only expected 2 than when you expected 200), and then sum across all categories.

The logic is short and sweet: big differences between O and E $\rightarrow$ big χ^2 $\rightarrow$ small p-value $\rightarrow$ reject H_0. A small χ^2 means observed and expected are close, so the data look like what H_0 predicted, and we don't reject.

Two things to hold onto. First, chi-square is for *counts of categorical data* — labels you can tally, not measurements you average. Second, the p-value comes from a curve called the **chi-square distribution**, which technology (or a table) reads for us once we hand it χ^2 and the degrees of freedom.

▪ WATCH IT WORK

Question: A four-sided spinner should land equally on A, B, C, D. In $n = 40$ spins we expect 10 of each. We observe: A $= 13$, B $= 9$, C $= 11$, D $= 7$. Just to feel the formula, compute χ^2.

Step 1 — List O and E for each category. $E = 10$ for all four (equal spinner, $40 \div 4$).

Step 2 — Find each $(O-E)^2/E$.

$$\text{A: } \frac{(13-10)^2}{10} = \frac{9}{10} = 0.9 \qquad \text{B: } \frac{(9-10)^2}{10} = \frac{1}{10} = 0.1$$

$$\text{C: } \frac{(11-10)^2}{10} = \frac{1}{10} = 0.1 \qquad \text{D: } \frac{(7-10)^2}{10} = \frac{9}{10} = 0.9$$

Step 3 — Add them up. $\chi^2 = 0.9 + 0.1 + 0.1 + 0.9 = 2.0$.

That's it — a number describing how far this spinner's counts strayed from "perfectly equal." A small value like 2.0 says "barely strayed." In the next sections we'll attach a p-value and an actual decision; here, just notice the machine: subtract, square, divide, add.

▪ YOUR TURN

A coin is flipped $n = 50$ times. If it's fair, we expect 25 heads and 25 tails. We observe 30 heads and 20 tails.

1. What is E for heads? For tails?
2. Compute $(O - E)^2/E$ for heads and for tails.
3. Add them to get χ^2.

Work space:

▪ CHECK YOUR VOICE

If the formula looked scary at first glance — all those symbols — notice what just happened. You did four subtractions, four squarings, four divisions, and one addition. That's grade-school arithmetic wearing a fancy Greek letter. The symbol χ^2 is not a threat; it's shorthand for "the adding-up you just did." Say it: "I can do the arithmetic, so I can do chi-square."

12.2 Goodness-of-Fit Test

▪ READ THIS FIRST

Back to the candy bag. The company prints a claim on the back: "40% brown, 30% yellow, 30% red." You dump out your bag and count the colors. Now you want to know: do my actual counts *fit* the company's claimed percentages, or did they fudge the numbers?

That question — "does my one pile of counts fit a claimed set of proportions?" — is exactly what a goodness-of-fit test answers. "Goodness of fit" literally means "how well do the observed counts fit the claim."

▪ LET'S TALK ABOUT IT

What's a claimed distribution you've heard? "Our customers are evenly split across the week." "This region is 60% / 40%." "Each number on the die is equally likely." Pick one. If you collected real counts, how would you decide whether reality matches the claim?

▪ NOW WE NAME IT

A **goodness-of-fit test** looks at *one* categorical variable and asks whether its data follow a *claimed distribution* of proportions.

The six-step rhythm, set up for goodness-of-fit:

- **Step 1 — Hypotheses.** H_0: the data follow the claimed proportions. H_a: they do not (at least one proportion is different).
- **Step 2 — Expected counts.** For each category, $E = n \times$ (claimed proportion).
- **Step 3 — Statistic.** $\chi^2 = \sum \dfrac{(O-E)^2}{E}$.
- **Step 4 — Degrees of freedom.** $df =$ (number of categories) $- 1$.
- **Step 5 — p-value.** The chi-square-curve probability (from technology) of a χ^2 at least this big.
- **Step 6 — Decide and say it plainly.** If p-value $\leq \alpha$, reject H_0; the data don't fit the claim. Otherwise, don't reject; the data are consistent with the claim.

One sanity check that catches mistakes: your expected counts should add up to n, the same total as your observed counts.

▪ WATCH IT WORK

Question: The candy company claims 40% brown, 30% yellow, 30% red. You count a bag of $n = 100$: brown $= 58$, yellow $= 24$, red $= 18$. Test the claim at $\alpha = 0.05$.

Step 1 — Hypotheses. H_0: the colors follow $40\%/30\%/30\%$. H_a: at least one proportion differs from the claim.

Step 2 — Expected counts ($E = n \times$ proportion):

$$\text{brown: } 100(0.40) = 40 \qquad \text{yellow: } 100(0.30) = 30 \qquad \text{red: } 100(0.30) = 30$$

(Check: $40 + 30 + 30 = 100 = n$. Good.)

Color	Observed O	Expected E	$(O-E)^2/E$
Brown	58	40	$(18)^2/40 = 324/40 = 8.1$
Yellow	24	30	$(-6)^2/30 = 36/30 = 1.2$
Red	18	30	$(-12)^2/30 = 144/30 = 4.8$

Step 3 — Statistic. $\chi^2 = 8.1 + 1.2 + 4.8 = 14.1$.

Step 4 — Degrees of freedom. Three categories, so $df = 3 - 1 = 2$.

Step 5 — p-value. For $\chi^2 = 14.1$ with $df = 2$, technology gives p-value ≈ 0.0009.

Step 6 — Decide. Since $0.0009 \leq 0.05$, we *reject* H_0. In plain words: this bag's color counts do *not* fit the company's claimed $40\%/30\%/30\%$ — there's far more brown (and less red) than the claim predicts.

▪ YOUR TURN

A carnival spinner has four equal-looking colors and the booth claims each is equally likely. You spin $n = 80$ times and observe: red $= 26$, blue $= 18$, green $= 22$, yellow $= 14$. Test the "all equal" claim at $\alpha = 0.05$.

1. Write H_0 and H_a.
2. Find the expected count for each color.
3. Compute χ^2 and state df. (The p-value is about 0.26 — what do you conclude?)

Work space:

▪ CHECK YOUR VOICE

Notice the test didn't just say "yes" or "no" — it told you a *story*: "too much brown to be an accident." If you got a little thrill from catching the company, hold onto that. That feeling — "the data revealed something" — is the whole reason statistics exists. You're not memorizing a ritual; you're learning to catch surprises.

12.3 Test of Independence

▪ READ THIS FIRST

Picture a survey of one group of students. For each person you record two things: their year (underclass or upperclass) and their morning drink (coffee, tea, or neither). Now you wonder: "Are these two things *related*? Do upperclass students lean more toward coffee, or is drink choice basically unrelated to year?"

You have *one sample* of people, and you measured *two categorical variables* on each of them. The question is whether the two variables travel together. That's a test of independence.

▪ LET'S TALK ABOUT IT

Think of two labels that get recorded about the same people — major and whether they have a campus job; phone brand and age group; favorite genre and whether they own a pet. Pick a pair. Do you suspect they're linked, or unrelated? A test of independence is how you'd check instead of guess.

▪ NOW WE NAME IT

A **test of independence** uses *two* categorical variables measured on *one* sample, arranged in a two-way **contingency table**. It asks whether the two variables are associated.

The six-step rhythm, set up for independence:

- **Step 1 — Hypotheses.** H_0: the two variables are *independent* (not associated). H_a: the two variables are associated (related).
- **Step 2 — Expected counts.** For each cell,

$$E = \frac{(\text{row total})(\text{column total})}{\text{grand total}}.$$

- **Step 3 — Statistic.** $\chi^2 = \sum \frac{(O-E)^2}{E}$, summed over every cell.
- **Step 4 — Degrees of freedom.** $df = (r-1)(c-1)$, where $r =$ number of rows and $c =$ number of columns.
- **Step 5 — p-value.** The chi-square-curve probability of a χ^2 at least this big.
- **Step 6 — Decide and say it plainly.** If p-value $\leq \alpha$, reject H_0: the variables *are* associated. Otherwise, don't reject: no evidence of association.

Where does the expected-count formula come from? If the variables really were independent, a cell's share would just be (its row's share) $\times$ (its column's share) of the total — which rearranges into row total times column total over grand total. You don't have to derive it; just trust the recipe and check that your expected counts still add up to the same row and column totals as the observed ones.

▪ WATCH IT WORK

Question: A survey of $n = 200$ students records year and morning drink. Are year and drink choice associated? Use $\alpha = 0.05$.

	Coffee	Tea	Neither	Row total
Underclass	60	30	10	100
Upperclass	40	30	30	100
Column total	100	60	40	200

Step 1 — Hypotheses. H_0: year and drink choice are independent. H_a: year and drink choice are associated.

Step 2 — Expected counts $\left(E = \dfrac{(\text{row total})(\text{column total})}{\text{grand total}}\right)$:

$$\text{Under, Coffee: } \frac{(100)(100)}{200} = 50 \qquad \text{Under, Tea: } \frac{(100)(60)}{200} = 30$$

$$\text{Under, Neither: } \frac{(100)(40)}{200} = 20$$

By the same arithmetic, the Upperclass row also expects 50, 30, 20. (Check: each expected row still totals 100, each expected column still totals its observed column total.)

Cell	Observed O	Expected E	$(O-E)^2/E$
Under, Coffee	60	50	$(10)^2/50 = 100/50 = 2$
Under, Tea	30	30	$0/30 = 0$
Under, Neither	10	20	$(-10)^2/20 = 100/20 = 5$
Upper, Coffee	40	50	$(-10)^2/50 = 100/50 = 2$
Upper, Tea	30	30	$0/30 = 0$
Upper, Neither	30	20	$(10)^2/20 = 100/20 = 5$

Step 3 — Statistic. $\chi^2 = 2 + 0 + 5 + 2 + 0 + 5 = 14$.

Step 4 — Degrees of freedom. $r = 2$ rows, $c = 3$ columns, so $df = (2-1)(3-1) = (1)(2) = 2$.

Step 5 — p-value. For $\chi^2 = 14$ with $df = 2$, technology gives p-value ≈ 0.0009.

Step 6 — Decide. Since $0.0009 \leq 0.05$, we *reject* H_0. In plain words: year and morning drink *are* associated — upperclass students are noticeably more likely to choose "neither," and underclass students lean toward coffee.

▪ YOUR TURN

A single sample of $n = 200$ adults is asked whether they exercise regularly (Yes/No) and whether they sleep well (Yes/No). Test for association at $\alpha = 0.05$.

	Sleeps well	Doesn't	Row total
Exercises	40	60	100
Doesn't	60	40	100
Column total	100	100	200

1. Write H_0 and H_a in terms of association.
2. Find the expected count for each of the four cells.
3. Compute χ^2 and state df. (The p-value is about 0.005 — what do you conclude?)

Work space:

▪ CHECK YOUR VOICE

If the table looked like a lot at first — six cells, totals everywhere — look back at what you actually did: one little $(O - E)^2/E$ per cell, exactly like Section 12.1, just six times instead of four. A bigger table is not a harder idea; it's the same idea repeated. You already had the skill before the table got bigger.

12.4 Test of Homogeneity (and How It Differs)

▪ READ THIS FIRST

Now a slightly different setup. Instead of one group measured on two things, imagine three *separate* groups — students at three different campuses — and from each campus you take its own sample of 100 and ask one question: "Do you prefer online or in-person classes?"

You're no longer asking "are two variables linked?" You're asking: "Do these three campuses have the *same* breakdown of preferences, or do they differ?" That's a test of homogeneity — "homogeneity" just means "sameness."

▪ LET'S TALK ABOUT IT

Here's the subtle part worth pausing on: in the last section we had *one* sample and recorded *two* things about each person. Here we have *several separate samples* (one per campus) and record *one* thing (preference). Same-looking table, different way of collecting the data. Can you feel the difference between "one group, two questions" and "several groups, one question"?

▪ NOW WE NAME IT

A **test of homogeneity** compares the distribution of *one* categorical variable across *several separate populations or groups* — each with its own sample. It asks: "Do these groups have the same distribution?"

Here's the reassuring news: the *mechanics are identical* to the test of independence.

- Same expected-count formula: $E = \dfrac{(\text{row total})(\text{column total})}{\text{grand total}}$.
- Same statistic: $\chi^2 = \sum \dfrac{(O-E)^2}{E}$.
- Same degrees of freedom: $df = (r-1)(c-1)$.

The arithmetic you'd type into a calculator or jamovi is exactly the same. What changes is only the *design* — how the data were gathered — and therefore the *wording* of the hypotheses.

The crystal-clear distinction:

- **Independence:** *one* sample, *two* categorical variables. H_0: the two variables are independent. ("Is drink choice associated with year, within one group of students?")
- **Homogeneity:** *several* samples (one per group), *one* categorical variable. H_0: all the groups have the same distribution of that variable. ("Do these three campuses share the same online/in-person split?")

If the row totals were *fixed in advance* (you decided "100 from each campus"), it's homogeneity. If you took one big sample and the totals fell where they fell, it's independence.

▪ WATCH IT WORK

Question: From each of three campuses, a researcher samples 100 students and asks: online or in-person? Do the campuses have the same preference distribution? Use $\alpha = 0.05$.

	Online	In-person	Row total
Campus A	70	30	100
Campus B	60	40	100
Campus C	50	50	100
Column total	180	120	300

Step 1 — Hypotheses. H_0: the three campuses have the *same* distribution of preference. H_a: at least one campus differs.

Step 2 — Expected counts (same formula as independence):

$$\text{Online column: } \frac{(100)(180)}{300} = 60 \text{ for each campus}$$

$$\text{In-person column: } \frac{(100)(120)}{300} = 40 \text{ for each campus}$$

So every cell expects 60 online and 40 in-person. (Check: each row still totals 100, columns still total 180 and 120.)

Cell	Observed O	Expected E	$(O-E)^2/E$
A, Online	70	60	$(10)^2/60 = 100/60 \approx 1.667$
A, In-person	30	40	$(-10)^2/40 = 100/40 = 2.5$
B, Online	60	60	$0/60 = 0$
B, In-person	40	40	$0/40 = 0$
C, Online	50	60	$(-10)^2/60 = 100/60 \approx 1.667$
C, In-person	50	40	$(10)^2/40 = 100/40 = 2.5$

Step 3 — Statistic. $\chi^2 \approx 1.667 + 2.5 + 0 + 0 + 1.667 + 2.5 = 8.33$.

Step 4 — Degrees of freedom. $r = 3$ rows, $c = 2$ columns, so $df = (3-1)(2-1) = (2)(1) = 2$.

Step 5 — p-value. For $\chi^2 = 8.33$ with $df = 2$, technology gives p-value ≈ 0.0155.

Step 6 — Decide. Since $0.0155 \leq 0.05$, we *reject* H_0. In plain words: the campuses do *not* share the same preference distribution — Campus A leans strongly online while Campus C is split evenly.

Notice this looked exactly like a test of independence in every calculation. The only thing that made it "homogeneity" was the design: *three separate samples* of one variable, with 100 fixed per campus.

▪ YOUR TURN

Two separate clinics each sample 80 patients and record whether each patient is satisfied (Yes/No). Are the two clinics' satisfaction distributions the same? Use $\alpha = 0.05$.

	Satisfied	Not	Row total
Clinic 1	48	32	80
Clinic 2	32	48	80
Column total	80	80	160

1. Is this independence or homogeneity? How can you tell from the design?
2. Find the expected count for each cell, then compute χ^2 and state df.
3. The p-value is about 0.011. What do you conclude in plain words?

Work space:

▪ CHECK YOUR VOICE

If the independence-vs-homogeneity distinction feels slippery, that's completely normal — they share every formula, so the only difference lives in the story behind the numbers. Here's your anchor: *count the samples*. One sample with two questions $\rightarrow$ independence. Several samples with one question $\rightarrow$ homogeneity. You don't have to feel it instantly; you just have to ask "how many samples?" every time, and the labels sort themselves out.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 12 in each companion matches this chapter, and

every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 12 — Your Turn Answers

Math Skills: (1) $(25-20)^2/20 = 25/20 = 1.25$. (2) $80 \times 0.25 = 20$ red candies. (3) $\dfrac{(50)(90)}{150} = \dfrac{4500}{150} = 30$.

Section 12.1: (1) $E = 25$ for heads and 25 for tails. (2) Heads: $(30-25)^2/25 = 25/25 = 1$; tails: $(20-25)^2/25 = 25/25 = 1$. (3) $\chi^2 = 1 + 1 = 2$.

Section 12.2: (1) H_0: all four colors are equally likely (25% each); H_a: at least one color's proportion differs. (2) $E = 80 \times 0.25 = 20$ for each color. (3) $\chi^2 = \dfrac{(26-20)^2}{20} + \dfrac{(18-20)^2}{20} + \dfrac{(22-20)^2}{20} + \dfrac{(14-20)^2}{20} = 1.8 + 0.2 + 0.2 + 1.8 = 4.0$, with $df = 4 - 1 = 3$. Since the p-value $\approx 0.26 > 0.05$, we *fail to reject* H_0 — the spinner's colors are consistent with "all equally likely."

Section 12.3: (1) H_0: exercise and sleep quality are independent (not associated); H_a: they are associated. (2) Every cell has the same expected count: $\dfrac{(100)(100)}{200} = 50$. (3) $\chi^2 = \dfrac{(40-50)^2}{50} + \dfrac{(60-50)^2}{50} + \dfrac{(60-50)^2}{50} + \dfrac{(40-50)^2}{50} = 2 + 2 + 2 + 2 = 8.0$, with $df = (2-1)(2-1) = 1$. Since the p-value $\approx 0.005 \leq 0.05$, we *reject* H_0 — exercising regularly and sleeping well *are* associated.

Section 12.4: (1) *Homogeneity* — there are two separate samples (one per clinic, 80 each fixed in advance) and one variable (satisfaction). (2) Each cell expects $\dfrac{(80)(80)}{160} = 40$. $\chi^2 = \dfrac{(48-40)^2}{40} + \dfrac{(32-40)^2}{40} + \dfrac{(32-40)^2}{40} + \dfrac{(48-40)^2}{40} = 1.6 + 1.6 + 1.6 + 1.6 = 6.4$, with $df = (2-1)(2-1) = 1$. (3) Since the p-value $\approx 0.011 \leq 0.05$, we *reject* H_0 — the two clinics do *not* have the same satisfaction distribution; Clinic 1's patients are more satisfied.

Chapter 13

Drawing the Line

Simple Linear Regression and Correlation — Measuring and Testing Relationships

CHAPTER PREVIEW: THE BIG PICTURE

Where We're Going. By the end of this chapter you'll look at two columns of numbers — hours studied and exam scores, say — and calmly answer three questions: "Do they move together?", "What's the best straight line through them?", and "Is that relationship real, or could it just be luck?"

The Story. Until now we've mostly studied *one* variable at a time. This chapter is about *two* quantitative variables and how they relate. The whole idea is human-sized: you already believe that more sleep tends to go with a better mood, or that more rain goes with greener grass. Regression just draws the straight line that captures "tends to go with," and inference asks whether that line is trustworthy.

The Skills You'll Have:

- Read a scatterplot and describe a relationship in plain words
- Find and read the *correlation* r — strength and direction in one number
- Compute the *least-squares regression line* and use it to predict
- Measure how well the line fits using *residuals* and r^2
- Test whether the relationship is statistically significant

The Confidence Moment. When a report says "the slope was significant ($p = .003$)," your brain will translate: "the line is real — x and y genuinely move together, not by accident."

The Bridge. This is the capstone. Every chapter built toward two big skills: *describing* data and *making an inference* about it. Here we do both at once, for the relationship between two variables — and that closes the inference arc of the whole book. "Regression" is just two friendly steps: **find the best line, then ask if it's real.**

MATH SKILLS YOU'LL NEED

This chapter leans on a little algebra you already met in high school. Nothing new — let's wake it up.

Skill 1 — Plotting a point. A point (x, y) means "go across to x, then up to y." The point $(3, 70)$ sits at $x = 3$ on the horizontal axis and $y = 70$ on the vertical axis.

Skill 2 — Slope as "rise over run." In the line $y = mx + b$, the slope m is how much y changes for each 1-unit step in x. If $m = 4.8$, then every time x goes up by 1, y goes up by 4.8. The number b is the *intercept*: the value of y when $x = 0$.

Skill 3 — Substituting into a line. To predict, just plug in. If $y = 54.8 + 4.8x$ and $x = 5$:

$$y = 54.8 + 4.8(5) = 54.8 + 24 = 78.8$$

Skill 4 — Square roots. You'll square a number ($0.8^2 = 0.64$) and occasionally take a square root ($\sqrt{16} = 4$). A calculator is welcome.

Practice (answers at the end of the chapter):

1. For the line $y = 10 + 3x$, what is y when $x = 4$?
2. A line has slope $m = 2$. If x increases by 1, how much does y change?
3. Compute 0.7^2 and then $\sqrt{0.49}$.

13.1 Seeing the Relationship: Scatterplots and Correlation

▪ READ THIS FIRST

Think about hours studied and exam scores in a class. You don't need a formula to guess the pattern: students who studied more *tended* to score higher. Not every single one — somebody crammed for twelve hours and bombed, somebody barely studied and aced it — but *on the whole*, more hours went with more points.

That phrase "on the whole, more goes with more" is the entire idea of this chapter. We're just going to give it a picture and a number.

▪ LET'S TALK ABOUT IT

Name two things in your life that "go together": maybe hours of sleep and your mood, or screen time and how tired your eyes feel, or outdoor temperature and how much ice cream a shop sells. For one pair, does *more* of the first go with *more* of the second, or with *less*? Just notice the direction — that's the first thing statistics will measure.

▪ NOW WE NAME IT

When we study how two *quantitative* variables relate, we usually pick one as the **explanatory variable** (called x — the thing we think does the explaining, like hours studied) and one as the **response variable** (called y — the thing that responds, like the exam score).

A **scatterplot** is the picture: each individual becomes one dot at the point (x, y). Hours studied goes on the horizontal axis, score on the vertical axis, and each student is one dot. If the dots drift up to the right, more x goes with more y (a *positive* relationship). If they drift down to the right, more x goes with less y (a *negative* relationship).

The **correlation coefficient**, written r, packs the strength and direction of a *straight-line* (linear) relationship into a single number between -1 and 1:

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

- The **sign** of r gives the *direction*: positive r = uphill, negative r = downhill.
- The **size** of r gives the *strength*: near ± 1 the dots hug a straight line (strong); near 0 they scatter like spilled rice (weak); exactly 0 means no linear pattern at all.

So $r = 0.95$ is a strong uphill relationship, $r = -0.85$ is a strong downhill one, and $r = 0.1$ is barely-there.

▪ WATCH IT WORK

Question: A study records how many hours five students slept and a mood rating (1–10) the next day. Sleep tends to pair with higher mood. Sketched in words: the dots run from the lower-left up to the upper-right in a tight, nearly straight band. Is the correlation positive or negative, and roughly strong or weak?

Step 1 — Find the direction. The dots go *up* as we move right (more sleep, higher mood). Up-to-the-right means a *positive* relationship, so r is positive.

Step 2 — Judge the strength. The dots form a *tight* band that hugs a straight line. Tight $=$ strong, so $|r|$ is close to 1.

Step 3 — Put it together. A strong, positive, linear relationship: r is positive and near 1 (maybe around 0.9).

Step 4 — The crucial caveat. Does more sleep *cause* a better mood? Maybe — but this study can't prove it. Calmer days might cause *both* better sleep and better mood. **Correlation is not causation** (remember Chapter 2: only a well-designed experiment can establish cause). A strong r tells us two things move together; it does not tell us *why*.

▪ YOUR TURN

For each described scatterplot, say whether r is positive or negative, and whether it's closer to strong or weak.

1. Outdoor temperature and hot-cocoa sales: dots drift clearly *down* to the right in a fairly tight band.
2. Shoe size and exam score: dots are scattered all over with no tilt at all.
3. Hours practiced and free-throws made: dots climb steadily up to the right, hugging a line.

Work space:

▪ CHECK YOUR VOICE

If the words "correlation coefficient" sounded intimidating, look at what you just did: you read a cloud of dots and said "up and tight" or "down and loose." That *is* correlation. The number r only puts a precise value on the judgment your eyes already made. You're not learning a new instinct — you're naming one you already have.

13.2 The Least-Squares Regression Line

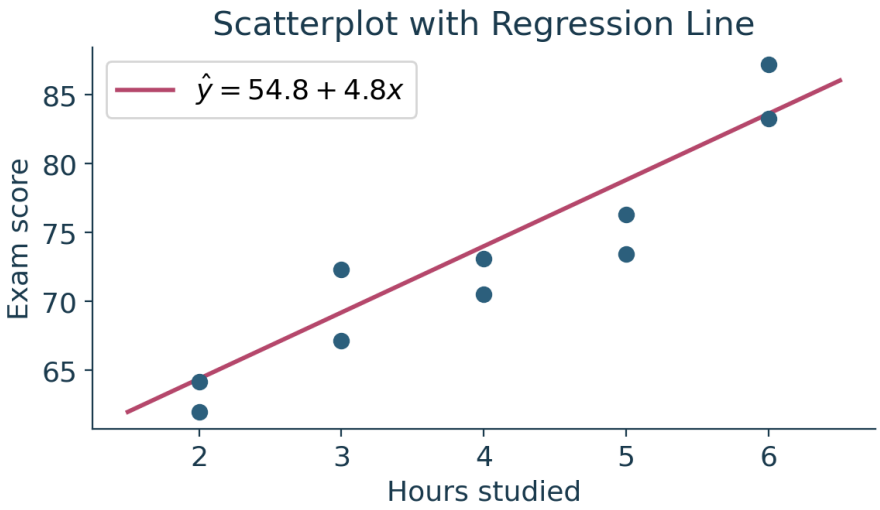


Figure 13.1. A scatterplot with its least-squares regression line.

▪ READ THIS FIRST

Picture the scatterplot of hours studied and exam scores. The dots trend uphill, but they don't sit on a perfect line. So here's a natural question: if I had to draw *one* straight line that best summarizes the trend, where would it go?

You could eyeball it with a ruler. Statistics does something more careful but the same in spirit: it finds the one line that comes closest to all the dots at once. That line has a name and a formula, and once you have it, you can *predict*.

▪ LET'S TALK ABOUT IT

If you drew a line through a cloud of dots by hand, what would "best" mean to you — the line that's closest to the most dots? The line where the misses above cancel the misses below? Hold that thought: "closest to all the dots" is exactly what the math is about to make precise.

▪ NOW WE NAME IT

The **least-squares regression line** is the straight line that makes the prediction errors as small as possible — specifically, it minimizes the sum of the *squared* vertical distances from the dots to the line. (Squaring keeps misses above and below from cancelling, and punishes big misses more.) We write it with a "hat" on y because it gives *predicted* values:

$$\hat{y} = b_0 + b_1 x$$

Here b_1 is the *slope* and b_0 is the *intercept*. The beautiful part: you can build the line straight from numbers you already know how to find:

$$b_1 = r\,\frac{s_y}{s_x} \qquad\qquad b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{x}, \bar{y}$ are the means and s_x, s_y are the standard deviations of x and y, and r is the correlation.
Reading the slope: for each 1-unit increase in x, the predicted $\hat{y}$ changes by b_1. **Reading the intercept:** b_0 is the predicted $\hat{y}$ when $x = 0$ — meaningful only if $x = 0$ is realistic.
One firm warning — don't extrapolate. The line is trustworthy only across the range of x values you actually observed. Predicting far outside that range (**extrapolation**) is guessing dressed up as math. If nobody in the data studied more than 6 hours, the line has nothing honest to say about 20 hours.

▪ WATCH IT WORK

Question: Five students report hours studied (x) and exam score (y):

Hours studied x	2	3	4	5	6
Exam score y	65	70	72	78	85

We are given the summary numbers $\bar{x} = 4$, $\bar{y} = 74$, $s_x = 1.5811$, $s_y = 7.7136$, and $r = 0.984$. Find the regression line and predict the score for a student who studies 5 hours.

Step 1 — Find the slope. Use $b_1 = r\,\dfrac{s_y}{s_x}$:

$$b_1 = 0.984 \times \frac{7.7136}{1.5811} = 0.984 \times 4.879 = 4.8$$

Step 2 — Find the intercept. Use $b_0 = \bar{y} - b_1\bar{x}$:

$$b_0 = 74 - (4.8)(4) = 74 - 19.2 = 54.8$$

Step 3 — Write the line.

$$\hat{y} = 54.8 + 4.8x$$

Step 4 — Interpret it in plain English. The slope 4.8 says: for each *additional hour studied*, the predicted exam score goes up by about 4.8 points. The intercept 54.8 is the predicted score for $x = 0$ hours — a student who didn't study at all. (Here $x = 0$ sits below the lowest x in our data, 2 hours, so this is an *extrapolation* — read the intercept cautiously.)

Step 5 — Predict. For a student who studies 5 hours, substitute $x = 5$:

$$\hat{y} = 54.8 + 4.8(5) = 54.8 + 24 = 78.8$$

We predict about a 79 on the exam.

Step 6 — Stay in bounds. Predicting for 5 hours is safe — it's inside our range of 2 to 6 hours. Predicting for 40 hours would be extrapolation, and we refuse to do it.

▪ YOUR TURN

Use the regression line $\hat{y} = 54.8 + 4.8x$ from the example above.

1. Predict the exam score for a student who studies 3 hours.
2. In one sentence, interpret what the slope 4.8 means in context.
3. Why would it be a bad idea to use this line to predict the score for someone who studied 25 hours?

Work space:

▪ CHECK YOUR VOICE

That was a lot of symbols — b_0, b_1, s_x, $\bar{y}$ — and you still got a clean line out the other end. If your stomach tightened at the formulas, notice they were just *recipes*: plug in the numbers you were handed, do the arithmetic, write the line. You didn't have to invent anything. "Find the best line" turned out to be two substitutions. Say it: "I can build a regression line."

13.3 How Well Does the Line Fit? Residuals and r^2

▪ READ THIS FIRST

Your line predicts 78.8 for a 5-hour student. But maybe the actual 5-hour student scored 78. The line missed by a little. Every dot has a tiny gap between where it *is* and where the line *said* it would be.

Those gaps are not failures — they're information. Add them up the right way and they tell you exactly how good your line is. A line whose gaps are all tiny is a great fit; a line with big gaps everywhere is barely better than guessing.

▪ LET'S TALK ABOUT IT

If you predicted a friend would score 80 and they scored 76, how far off were you, and in which direction — too high or too low? That single "how far off, and which way" is the whole idea of a residual. You've been computing them in your head your whole life.

▪ NOW WE NAME IT

A **residual** is the leftover — how far the actual point sits above or below the line:

$$\text{residual} = y - \hat{y} \qquad (\text{observed} - \text{predicted})$$

A *positive* residual means the point sits *above* the line (we under-predicted); a *negative* residual means it sits *below* (we over-predicted). The least-squares line is precisely the one that makes the squared residuals as small as possible.

To summarize fit in a single number, we use the **coefficient of determination**, r^2 — literally the correlation squared. It is the **proportion of the variation in y that the line explains**, and it always falls between 0 and 1:

$$0 \le r^2 \le 1$$

For example, if $r = 0.8$, then $r^2 = 0.64$, and we say "64% of the variation in y is explained by the linear relationship with x." The other 36% is due to everything else the line doesn't capture. Closer to 1 = the line explains almost everything (tight fit); closer to 0 = the line explains little (loose fit).

▪ WATCH IT WORK

Question: Using our line $\hat{y} = 54.8 + 4.8x$ and the same five students, find the residual for the student who studied 3 hours (and actually scored 70), and report and interpret r^2 given $r = 0.984$.

Step 1 — Predict for $x = 3$.

$$\hat{y} = 54.8 + 4.8(3) = 54.8 + 14.4 = 69.2$$

Step 2 — Subtract to get the residual. The student actually scored $y = 70$:

$$\text{residual} = y - \hat{y} = 70 - 69.2 = 0.8$$

Step 3 — Read the residual. The residual is $+0.8$, a small positive number. The student scored 0.8 points *above* what the line predicted — the line did a good job, missing by less than a point.

Step 4 — Square the correlation. With $r = 0.984$:

$$r^2 = (0.984)^2 = 0.968$$

Step 5 — Interpret r^2. About 96.8% of the variation in exam scores is explained by the linear relationship with hours studied. That's a very strong fit.

Step 6 — Sanity check. A tiny residual and an r^2 near 1 tell the same story two ways: this line fits these dots beautifully. Good — they should agree.

▪ YOUR TURN

Still using $\hat{y} = 54.8 + 4.8x$ for the five students.

1. The student who studied 6 hours actually scored 85. Find the predicted score and the residual.
2. Was the line's prediction for that student too high or too low?
3. If a different study found $r = 0.5$, what is r^2, and what percent of the variation in y does the line explain?

Work space:

▪ CHECK YOUR VOICE

Notice r^2 didn't require any scary new calculation — you *squared a number you already had*. And "residual" turned out to be plain subtraction: what happened minus what you guessed. The hardest part of this section was the vocabulary, and you just used all of it correctly. That glaze-over feeling from page one? Gone.

13.4 Is the Relationship Real? Testing the Slope and the Correlation

▪ READ THIS FIRST

Here's an honest worry. We found a slope of 4.8 and a correlation of 0.984 from just five students. But what if those five were a fluke? Flip enough coins and you'll see a streak; sample enough students and you might see a "relationship" that's really just luck. We need a way to ask: *is this relationship real, or could random chance alone have produced it?*

That's a hypothesis test — the same machinery from the last several chapters, now pointed at a slope. Nothing new conceptually: state a skeptical null, compute a test statistic, get a p-value, decide.

▪ LET'S TALK ABOUT IT

Suppose someone shows you a relationship built from only *three* people and says "see, more coffee means better grades!" What's your gut reaction — convinced, or suspicious that three people could line up by accident? That suspicion is exactly what the slope test measures and tames.

▪ NOW WE NAME IT

In the population, the true regression line has a true slope we call β_1 (Greek beta) and the true correlation is ρ (Greek rho). "No linear relationship" means the true line is flat — a slope of zero. So the **null hypothesis** is:

$$H_0 : \beta_1 = 0 \qquad \text{(equivalently } H_0 : \rho = 0)$$

against $H_a : \beta_1 \neq 0$ (the relationship is real). We test the slope with a t-statistic — estimate divided by its standard error — with $n - 2$ degrees of freedom:

$$t = \frac{b_1}{SE(b_1)}, \qquad df = n - 2$$

Software reports b_1, $SE(b_1)$, this t, and its p-value automatically. There's an equivalent test phrased in terms of the correlation, which you can compute by hand from just r and n:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n - 2$$

Here is the reassuring fact: **these two tests give the exact same t and the exact same p-value.** Testing "is the slope zero?" and testing "is the correlation zero?" are the same question. A small p-value (below your significance level, usually 0.05) means: chance alone is an unconvincing explanation, so we conclude the linear relationship is *statistically significant*.

▪ WATCH IT WORK

Question: For our five students, $r = 0.984$ and $n = 5$. Test, at $\alpha = 0.05$, whether there's a significant linear relationship between hours studied and exam score.

Step 1 — State the hypotheses. $H_0 : \rho = 0$ (no linear relationship) versus $H_a : \rho \neq 0$ (there is one). Equivalently $H_0 : \beta_1 = 0$.

Step 2 — Find the degrees of freedom. $df = n - 2 = 5 - 2 = 3$.

Step 3 — Compute the test statistic. First the pieces: $1 - r^2 = 1 - 0.968 = 0.032$, so $\sqrt{1 - r^2} = \sqrt{0.032} = 0.1789$. And $\sqrt{n - 2} = \sqrt{3} = 1.732$. Then:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.984 \times 1.732}{0.1789} = \frac{1.704}{0.1789} \approx 9.52$$

Step 4 — Get the p-value. With $t \approx 9.52$ and $df = 3$, software (or a t-table) gives a two-sided p-value of about 0.0024. (For comparison, the critical value for $df = 3$ at $\alpha = 0.05$ is about 3.182; our $t \approx 9.52$ blows past it.)

Step 5 — Decide. The p-value 0.0024 is far below 0.05, so we *reject* H_0.

Step 6 — Conclude in context. There is statistically significant evidence of a positive linear relationship between hours studied and exam score ($t(3) \approx 9.52$, $p = 0.0024$). It's very unlikely we'd see a correlation this strong from five students by chance alone. (Still — significant means *real*, not *causal*. We've shown the relationship isn't luck; we have *not* shown that studying *causes* the higher scores.)

▪ YOUR TURN

A researcher collects $n = 10$ pairs and finds a correlation of $r = 0.60$. Test whether the linear relationship is significant at $\alpha = 0.05$.

1. State H_0 and H_a in terms of ρ, and give the degrees of freedom.
2. Compute the test statistic $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. (You'll need $r^2 = 0.36$.)
3. The two-sided p-value turns out to be about 0.067. At $\alpha = 0.05$, do you reject H_0? What do you conclude?

Work space:

▪ CHECK YOUR VOICE

You just ran a full hypothesis test on a relationship — null, statistic, p-value, decision, conclusion — and it followed the *exact* rhythm of every test in this book. The only new thing was the formula for t, and you plugged numbers into it just fine. If part of you expected "regression inference" to be a wall, look back: it was a familiar staircase. You climbed it. That's the whole book, finished, in your hands.

▪ WANT TO TRY IT ON A COMPUTER?

Every calculation in this chapter is walked through, one step at a time, in the free **Technology Companions**: one each for **R & RStudio**, **jamovi**, **StatCrunch**, and the **TI-84**. You can find them at `learnwithoutwalls.com`. Chapter 13 in each companion matches this chapter, and every example comes with a ready-made dataset. Learn the idea here; the button-pushing is waiting for you there, whenever you want it.

Chapter 13 — Your Turn Answers

Math Skills: (1) $y = 10 + 3(4) = 22$. (2) y increases by 2. (3) $0.7^2 = 0.49$ and $\sqrt{0.49} = 0.7$.

Section 13.1: (1) Negative, and strong (down-to-the-right, tight band). (2) Essentially $r = 0$ — no linear relationship (a scatter with no tilt). (3) Positive, and strong (climbs steadily, hugs a line).

Section 13.2: (1) $\hat{y} = 54.8 + 4.8(3) = 54.8 + 14.4 = 69.2$, about a 69. (2) For each additional hour studied, the predicted exam score increases by about 4.8 points. (3) 25 hours is far outside the observed range (2 to 6 hours), so predicting there is extrapolation — the line has no evidence to back up a guess that far out.

Section 13.3: (1) $\hat{y} = 54.8 + 4.8(6) = 54.8 + 28.8 = 83.6$; residual $= 85 - 83.6 = 1.4$. (2) Too low — the actual score (85) was above the prediction (83.6), so the line under-predicted (positive residual). (3) $r^2 = 0.5^2 = 0.25$, so the line explains 25% of the variation in y.

Section 13.4: (1) $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$; $df = n - 2 = 10 - 2 = 8$. (2) $t = \dfrac{0.60\sqrt{8}}{\sqrt{1 - 0.36}} = \dfrac{0.60 \times 2.828}{\sqrt{0.64}} = \dfrac{1.697}{0.8} = 2.12$. (3) The p-value 0.067 is greater than 0.05, so we *fail to reject* H_0: with only 10 pairs, this correlation isn't strong enough to rule out chance — the linear relationship is not statistically significant at the 0.05 level.

Part VI Checkpoint — Study Guide & Practice (Categorical Data & Relationships)

This Study Guide gathers everything from Chapters 12 and 13 — chi-square tests and regression — and lets you rehearse before an exam the way the exam will actually feel.

STUDY GUIDE & PRACTICE

You just finished two of the most "grown-up" chapters in the whole book: testing whether categories are linked, and drawing the best line through a cloud of dots. This isn't a test — it's a dress rehearsal. We'll gather the big ideas in one place, check the vocabulary, and then run ten mixed problems that jump between the two chapters, exactly the way a real exam does. When your inner voice says "I forgot all of this," keep going anyway — every answer is right here, fully worked, waiting for you.

The Big Ideas, in One Place

Chapter 12 — Counts, Categories, and Surprises (Chi-Square).

- Every chi-square test does the *same* thing: compare the counts you *observed* (O) to the counts you'd *expect* (E) if nothing interesting were going on, and measure the total gap with

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

  For each category or cell: subtract, square, divide by E, then add them all up. Big gaps $\rightarrow$ big χ^2 $\rightarrow$ small p-value $\rightarrow$ reject H_0.

- **Goodness-of-fit test**: *one* categorical variable versus a *claimed* set of proportions. Expected counts are $E = n \times$ (claimed proportion), and the degrees of freedom are

$$df = (\text{number of categories}) - 1.$$

  Sanity check: the expected counts must add up to n, just like the observed counts.

- **Test of independence** and **test of homogeneity**: both use a two-way table, and both use the *identical* machinery. Expected count for each cell is

$$E = \frac{(\text{row total})(\text{column total})}{\text{grand total}}, \qquad df = (r-1)(c-1),$$

where r is the number of rows and c is the number of columns.

- **Independence vs. homogeneity is decided by *design*, not by arithmetic.** *Independence:* one sample, two categorical variables — "are these two things associated within one group?" *Homogeneity:* several separate samples (often with totals fixed in advance), one categorical variable — "do these groups share the same distribution?" Your anchor question: *how many samples?* One sample, two questions $\rightarrow$ independence. Several samples, one question $\rightarrow$ homogeneity.

Chapter 13 — Drawing the Line (Regression & Correlation).

- A **scatterplot** plots each individual as a dot (x, y), with the explanatory variable x across and the response y up. Dots drifting up-to-the-right $=$ positive; down-to-the-right $=$ negative.

- The **correlation** r packs strength and direction of a *straight-line* relationship into one number with $-1 \le r \le 1$. The *sign* gives direction (positive $=$ uphill, negative $=$ downhill); the *size* gives strength (near ± 1 $=$ tight/strong, near 0 $=$ scattered/weak).

- The **coefficient of determination** r^2 is literally r squared, and it is the *proportion of the variation in y explained by the linear relationship with x*, with $0 \le r^2 \le 1$. If $r = 0.8$, then $r^2 = 0.64$: "64% of the variation in y is explained."

- The **least-squares regression line** is written with a hat because it predicts:

$$\hat{y} = b_0 + b_1 x, \qquad b_1 = r\,\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1 \bar{x}.$$

The slope b_1 is how much predicted $\hat{y}$ changes per 1-unit increase in x; the intercept b_0 is the predicted $\hat{y}$ when $x = 0$.

- A **residual** is observed minus predicted: residual $= y - \hat{y}$. Positive $=$ the point sits above the line (under-predicted); negative $=$ below (over-predicted).

- **Two firm warnings.** *Don't extrapolate* — the line is only trustworthy across the range of x you actually observed. And *correlation is not causation* — a real relationship means x and y move together, not that one causes the other.

- To ask "is the relationship *real*, or just luck?" test the slope (equivalently, the correlation). The null is $H_0 : \beta_1 = 0$ (equivalently $H_0 : \rho = 0$). You can compute the test statistic by hand from r and n:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n - 2.$$

  A small p-value means chance alone is an unconvincing explanation: the relationship is statistically significant.

Vocabulary Check

Can you say each of these in one plain sentence? If one feels fuzzy, that's exactly the one to reread — not a sign you're behind.

- observed count O, expected count E, chi-square statistic χ^2
- goodness-of-fit test; claimed distribution
- contingency (two-way) table; test of independence; test of homogeneity
- degrees of freedom for goodness-of-fit (categories-1) vs. for a two-way table ($(r-1)(c-1)$)
- scatterplot; explanatory variable (x); response variable (y)
- correlation r; positive vs. negative; strong vs. weak
- coefficient of determination r^2
- least-squares regression line $\hat{y} = b_0 + b_1 x$; slope b_1; intercept b_0
- residual ($y - \hat{y}$)
- extrapolation
- testing the slope/correlation; statistically significant
- correlation $\neq$ causation

Mixed Practice

These ten problems jump around on purpose — just like an exam. Try each one before peeking. The full worked answers come right after.

1. A gamer suspects a six-sided die is unfair. She rolls it $n = 60$ times (a fair die would land 10 times on each face) and observes: $1 \to 14$, $2 \to 8$, $3 \to 12$, $4 \to 6$, $5 \to 11$, $6 \to 9$. Compute the goodness-of-fit χ^2 and state the degrees of freedom. The p-value turns out to be about 0.52. At $\alpha = 0.05$, decide whether the die is unfair.

2. A survey of 300 people records their phone brand (three brands) and their age group (three groups), arranged in a two-way table. In one cell, the row total is 80 and the column total is 120. Find the expected count for that cell.

3. A study cross-tabulates favorite music genre (4 categories) against region of the country (3 categories) for one sample of listeners. How many degrees of freedom does the chi-square test of independence use?

4. A researcher takes *one* sample of 250 employees and records both their department (Sales, Engineering, HR) and whether they work remotely (Yes/No). Is this a test of independence or a test of homogeneity? Explain how you can tell from the design.

5. A study reports a correlation of $r = -0.88$ between the number of hours per week a person spends on screens and the number of hours they sleep. Interpret this r in plain words — direction and strength.

6. Five study sessions are recorded: hours practiced $x = 1, 2, 3, 4, 5$ and quiz score $y = 3, 5, 6, 8, 8$. You are given the summary numbers $\bar{x} = 3$, $\bar{y} = 6$, $s_x = 1.5811$, $s_y = 2.1213$, and $r = 0.969$. Find the least-squares regression line $\hat{y} = b_0 + b_1 x$, then predict the score for a student who practices 4 hours.

7. In a different regression, the correlation between advertising spending and weekly sales is $r = 0.6$. Compute r^2 and interpret it in one sentence.

8. A report on $n = 40$ cities finds a significant positive slope when predicting average commute time from population size: $t(38) = 4.1$, $p = 0.0002$. In plain words, what does "the slope is significant" tell you here? (And what does it *not* tell you?)

9. Over many months, a town notices that the more firefighters it sends to a fire, the more property damage the fire causes — a strong positive correlation. A columnist concludes, "Firefighters cause damage; we should send fewer." Explain why this reasoning is wrong, and name a confounding variable that better explains the link.

10. A goodness-of-fit test is run on a categorical variable with 5 categories, and a chi-square test

of independence is run on a 3×4 contingency table. State the degrees of freedom for each test.

Answers

Worked out in full — read these even for the ones you got right, because the *reasoning* is what the exam rewards.

1. Every face has $E = 10$ (a fair die over 60 rolls). Compute each $(O - E)^2/E$ and add:

Face	Observed O	Expected E	$(O-E)^2/E$
1	14	10	$(4)^2/10 = 16/10 = 1.6$
2	8	10	$(-2)^2/10 = 4/10 = 0.4$
3	12	10	$(2)^2/10 = 4/10 = 0.4$
4	6	10	$(-4)^2/10 = 16/10 = 1.6$
5	11	10	$(1)^2/10 = 1/10 = 0.1$
6	9	10	$(-1)^2/10 = 1/10 = 0.1$

   $\chi^2 = 1.6 + 0.4 + 0.4 + 1.6 + 0.1 + 0.1 = 4.2$. With 6 categories, $df = 6 - 1 = 5$. Since the p-value $\approx 0.52 > 0.05$, we *fail to reject* H_0 — the counts are consistent with a fair die. (The face-to-face wobble is the kind of variation you'd expect from chance alone.)

2. Use $E = \dfrac{(\text{row total})(\text{column total})}{\text{grand total}}$:

$$E = \frac{(80)(120)}{300} = \frac{9600}{300} = 32.$$

   The expected count for that cell is 32.

3. For a two-way table, $df = (r-1)(c-1)$. Here genre gives $r = 4$ rows and region gives $c = 3$ columns, so

$$df = (4-1)(3-1) = (3)(2) = 6.$$

4. This is a *test of independence*. The tell is the design: there is *one* sample of 250 employees, and *two* categorical variables (department and remote status) are recorded on each person. The totals weren't fixed in advance — they fell where they fell. The question is "are these two variables associated within this one group?" If instead the researcher had drawn a fixed-size sample from *each* department separately and asked only about remote work, that would be homogeneity. Anchor: one sample, two questions $\rightarrow$ independence.

5. The correlation $r = -0.88$ describes a *strong, negative* linear relationship. The *negative* sign says the two variables move in *opposite* directions: more screen time tends to go with *less* sleep. The *size*, 0.88 (close to 1), says the relationship is *strong* — the dots would hug a downhill line fairly tightly. (Strong, but still not proof that screens *cause* less sleep.)

6. **Step 1 — Slope.** $b_1 = r\,\dfrac{s_y}{s_x} = 0.969 \times \dfrac{2.1213}{1.5811} = 0.969 \times 1.3416 = 1.3.$

   Step 2 — Intercept. $b_0 = \bar{y} - b_1\bar{x} = 6 - (1.3)(3) = 6 - 3.9 = 2.1.$

   Step 3 — The line. $\hat{y} = 2.1 + 1.3x.$

   Step 4 — Predict for $x = 4$. $\hat{y} = 2.1 + 1.3(4) = 2.1 + 5.2 = 7.3.$ We predict a quiz score of about 7.3. (And $x = 4$ is safely inside the observed range of 1 to 5 hours, so no extrapolation.)

7. $r^2 = (0.6)^2 = 0.36$. About 36% of the variation in weekly sales is explained by the linear relationship with advertising spending. (The other 64% is due to everything the line doesn't capture.)

8. "The slope is significant" means the relationship is *real, not a fluke*: the small p-value (0.0002) says it's very unlikely we'd see a positive slope this strong from 40 cities by chance alone, so we reject $H_0 : \beta_1 = 0$. Bigger cities genuinely tend to have longer commutes in this data. What it does *not* tell you: that larger population *causes* longer commutes — significant means real, not causal, and this is observational data, so confounders could be at work.

9. This confuses correlation with causation — two things rising together doesn't mean one causes the other. The *confounding variable* is the *size/severity of the fire*: big fires both draw more firefighters *and* cause more property damage. Fire size is tied to both, so it — not the firefighters — explains the link. Sending fewer firefighters wouldn't reduce damage; it would almost certainly make it worse.

10. *Goodness-of-fit* (5 categories): $df = (\text{number of categories}) - 1 = 5 - 1 = 4$. *Test of independence* (3×4 table): $df = (r-1)(c-1) = (3-1)(4-1) = (2)(3) = 6$. Notice the two tests use *different* degree-of-freedom rules — a classic exam trap, so always ask first which kind of test you're running.

You just rehearsed every big idea from Part VI — computing a chi-square statistic, finding expected counts, picking the right degrees of freedom, telling independence from homogeneity, reading a correlation, building and using a regression line, interpreting r^2, and keeping "significant" separate from "causal." If some were shaky, now you know exactly where to look. That's not failing the

rehearsal; that's what the rehearsal is *for*. You're ready.

Part VI — Formula Sheet: Categorical Data & Relationships

Every chi-square and regression formula on one page, each symbol named in plain words with a note on when *to use it.*

FORMULA SHEET

The symbols, in plain words:

- O = observed count E = expected count χ^2 = chi-square statistic df = degrees of freedom
- r = correlation coefficient $\hat{y}$ = predicted value b_0, b_1 = intercept, slope $\bar{x}, \bar{y}$ = sample means s_x, s_y = sample SDs

Chi-square tests — all use (Chapter 12)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$ (a value of 0 means a perfect match; bigger $\chi^2 \Rightarrow$ smaller p-value)

- **Goodness-of-fit** (one categorical variable vs. a claimed distribution): $E = n \times$ (claimed proportion), $df =$ (number of categories) $- 1$.
- **Independence** (two categorical variables, one sample) and **Homogeneity** (one variable across several samples) — same arithmetic: $E = \frac{\text{(row total)(column total)}}{\text{grand total}}$, $df = (r-1)(c-1)$.

When: chi-square is for *counts* in categories. Independence vs. homogeneity differ only in the study design (one sample/two variables vs. several samples/one variable), not the formula.

Correlation and regression (Chapter 13)

$$-1 \le r \le 1 \qquad \hat{y} = b_0 + b_1 x \qquad b_1 = r\,\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

r^2 = proportion of the variation in y explained by the line residual $= y - \hat{y}$

When: for two quantitative variables. r measures the strength/direction of a *linear* relationship; the least-squares line predicts y from x (don't extrapolate beyond the data); r^2 tells you how much of y's variation the line explains.

Is the relationship real? (Chapter 13)

$$\text{slope: } t = \frac{b_1}{SE(b_1)} \qquad \equiv \qquad \text{correlation: } t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n-2$$

When: test $H_0 : \beta_1 = 0$ (equivalently $H_0 : \rho = 0$) — "is there really a linear relationship, or just chance?" Both tests give the *same* p-value. Small p-value $\Rightarrow$ a statistically significant linear relationship (still not proof of causation).

Final Exam Prep: Mixed Practice

This pulls from the whole book at once. On a real final, nobody tells you which tool to reach for — choosing the right one is *the skill, so that's exactly what we'll rehearse.*

STUDY GUIDE & PRACTICE

Here we are, at the end. Take a breath. Everything in this Study Guide is something you have already done at least once in this book — we're just putting it all on one table so you can practice picking the right tool without being told which chapter it came from. That "which tool?" moment is the only thing a cumulative final adds, and we're going to practice it until it feels ordinary. When your inner voice says "I don't remember any of this," keep going anyway. The fully worked answers are right here, waiting for you.

How to Choose the Right Tool

Before you compute anything, ask one question: **what is this problem asking me to *do*?** Almost every problem in the book is doing one of three jobs.

- **Describing data** — "summarize what we have." You make a **graph** (bar chart or pie for categorical; dotplot, stem-and-leaf, or histogram for quantitative) and report *center* (mean, median), *spread* (standard deviation, IQR, range), and *shape* (symmetric, skewed). No claim about a larger population — just "here's what this data looks like."
- **Estimating** — "what's the true value out in the population, give or take?" You build a **confidence interval**: a best guess $\pm$ a margin of error. Signal words: *estimate*, *how confident*, *margin of error*, *within* $\pm$.
- **Testing** — "is there real evidence *for* a claim, or could this just be chance?" You run a **hypothesis test**, set up H_0 and H_a, get a test statistic and a p-value, and decide. Signal words: *is there evidence*, *significantly different*, *test the claim*, *at the* 5% *level*.

Once you know you're **testing**, pick the specific test by counting groups and noticing the variable type:

The question is about...	Use...
One quantitative mean vs. a claimed number	one-sample t-test (use z in the rare case where σ is given)
Two quantitative means (two groups)	two-sample t-test
Three or more quantitative means	ANOVA (F-test)
Counts in categories (one or two categorical variables)	chi-square test
Relationship between two quantitative variables	regression / correlation

Memory hook: **count the groups, then check the variable type.** One/two/three+ *quantitative* groups walk you from t to two-sample t to ANOVA; *categorical counts* send you to chi-square; a *relationship between two number variables* sends you to regression and correlation.

Mixed Practice (Cumulative)

These fifteen problems jump across the whole book on purpose — just like the final. For several of them, your *first* job is to name which procedure applies before you do any arithmetic. Try each one before you peek; the full worked answers come right after.

1. (Ch 1) Classify each variable as categorical or quantitative, and if quantitative, discrete or continuous: (a) zip code, (b) number of pets owned, (c) weight in kilograms, (d) movie genre.

2. (Ch 2) A website posts a poll on its homepage: "Click here to tell us if you love our new design!" Of those who clicked, 90% said yes. Name the most likely type of bias and explain why this can't be trusted as the opinion of all visitors.

3. (Ch 3) A histogram of household incomes for a city has most households bunched at the lower end, with a long tail stretching toward a few very high incomes. Name the shape, and say whether you'd expect the mean or the median to be larger.

4. (Ch 4) A set of quiz scores has mean $\bar{x} = 70$ and standard deviation $s = 8$. (a) Find the z-score of a score of 86 and say what it means. (b) The five-number summary is min $= 40$, $Q_1 = 62$, median $= 70$, $Q_3 = 78$, max $= 96$. Use the $1.5 \times IQR$ rule to check whether 96 is an outlier.

5. (Ch 5) A bag holds 20 marbles: 8 red, 7 blue, 5 green. You draw one at random. (a) Find $P(\text{red})$. (b) Find $P(\text{not green})$.

6. (Ch 6) A free-throw shooter makes each shot independently with probability $p = 0.3$. She

takes $n = 10$ shots. *First name the model*, then find the probability she makes exactly 2, and find the mean and standard deviation of the number she makes.

7. (Ch 6) Scores on a test are normally distributed with mean 500 and standard deviation 100. Using the empirical (68–95–99.7) rule, what percent of scores fall between 400 and 600? Between 300 and 700?

8. (Ch 7) A population has standard deviation $\sigma = 20$. You take random samples of size $n = 100$ and record the sample mean each time. (a) What is the standard error of $\bar{x}$? (b) In one sentence, why is the spread of sample means smaller than the spread of individual values?

9. (Ch 8) *Name the procedure first.* A random sample of $n = 64$ light bulbs has mean life $\bar{x} = 70$ hundred hours, and the population standard deviation is $\sigma = 16$. Build a 95% confidence interval for the true mean bulb life. *(Because σ is given, this is the rare "σ known" case from Chapter 8, so use the normal curve and $z^* = 1.96$ — not t. When σ is unknown, as usual, you would use t with $df = n - 1$.)*

10. (Ch 8) *Name the procedure first.* In a random sample of 400 voters, 240 support a measure. Build a 95% confidence interval for the true proportion of all voters who support it.

11. (Ch 9) *Name the test first.* A factory claims its bags of flour weigh 100 grams on average. A random sample of $n = 25$ bags has mean $\bar{x} = 106$; treat the population standard deviation as $\sigma = 15$ (known). At the 5% level, is there evidence the true mean is *not* 100 grams? *(Since σ is known here, use the z version — the rare case; with unknown σ you would use the one-sample t-test with $df = n - 1$.)*

12. (Ch 10) *Name the test first.* Two teaching methods are compared. Method A: $n_1 = 50$, $\bar{x}_1 = 82$, $s_1 = 10$. Method B: $n_2 = 50$, $\bar{x}_2 = 78$, $s_2 = 10$. At the 5% level, is there a significant difference in mean scores?

13. (Ch 11) An ANOVA comparing the mean test scores of *four* different study programs returns $F = 5.2$ with a p-value of 0.008. State the hypotheses in words and give the conclusion at the 5% level. What does a significant result here tell you — and what does it *not* tell you?

14. (Ch 12) *Name the test first.* A fair six-sided die is rolled 60 times, giving counts: face 1: 8, face 2: 12, face 3: 9, face 4: 11, face 5: 15, face 6: 5. At the 5% level (critical value $\chi^2 = 11.07$, $df = 5$), is there evidence the die is unfair?

15. (Ch 13) For a sample of students, hours studied and exam score have correlation $r = 0.85$,

and the least-squares line is $\hat{y} = 2 + 3x$ (with x in hours). (a) Describe the relationship in words. (b) Predict the score for someone who studies 4 hours. (c) What does r^2 tell you here?

Answers

Worked out in full — read these even for the ones you got right, because on a cumulative final the *reasoning* (especially *which tool*) is what earns the points.

1. (a) Zip code = *categorical* (it's a label made of digits; averaging zip codes is meaningless). (b) Number of pets = *quantitative, discrete* (a count of whole things). (c) Weight in kilograms = *quantitative, continuous* (a measurement that can land anywhere in a range). (d) Movie genre = *categorical* (a label).

2. *Voluntary-response bias*. Only people who felt strongly enough to click in responded, so the sample over-represents intense opinions and is not a representative slice of all visitors. A self-selected "yes" rate tells you about the clickers, not about everyone.

3. The distribution is *skewed right* (the tail stretches toward the high end). In a right-skewed distribution the few large values pull the *mean* up, so the **mean is larger than the median**.

4. (a) $z = \dfrac{86 - 70}{8} = \dfrac{16}{8} = 2.0$: the score 86 sits 2 standard deviations *above* the mean. (b) $IQR = Q_3 - Q_1 = 78 - 62 = 16$. Upper fence $= Q_3 + 1.5(IQR) = 78 + 1.5(16) = 78 + 24 = 102$. Since $96 < 102$, the value 96 is *not* an outlier.

5. Total $= 20$. (a) $P(\text{red}) = \dfrac{8}{20} = 0.40$. (b) $P(\text{not green}) = 1 - P(\text{green}) = 1 - \dfrac{5}{20} = 1 - 0.25 = 0.75$.

6. This is a *binomial model*: a fixed number of independent shots ($n = 10$), each a success/failure with the same probability ($p = 0.3$). Exactly 2 made:

$$P(X = 2) = \binom{10}{2}(0.3)^2(0.7)^8 = 45\,(0.09)(0.0576) \approx 0.233.$$

   Mean $= np = 10(0.3) = 3$ shots. Standard deviation $= \sqrt{np(1-p)} = \sqrt{10(0.3)(0.7)} = \sqrt{2.1} \approx 1.45$ shots.

7. By the empirical rule, about 68% of values fall within 1 standard deviation of the mean. Here 400 to 600 is 500 ± 100, i.e. ± 1 SD, so about **68%**. And 300 to 700 is $500 \pm 200 = \pm 2$ SD, so about **95%**.

8. (a) Standard error $= \dfrac{\sigma}{\sqrt{n}} = \dfrac{20}{\sqrt{100}} = \dfrac{20}{10} = 2$. (b) Averaging many values lets the high and low ones cancel out, so sample means vary less than single observations — bigger samples give steadier averages.

9. **Procedure: a confidence interval for a mean** (we're estimating one quantitative average, σ known). $SE = \dfrac{\sigma}{\sqrt{n}} = \dfrac{16}{\sqrt{64}} = \dfrac{16}{8} = 2$. Margin $= 1.96(2) = 3.92$. CI $= 70 \pm 3.92 = (66.08,\ 73.92)$ hundred hours. We're 95% confident the true mean bulb life is between about 66.08 and 73.92 hundred hours.

10. **Procedure: a confidence interval for a proportion** (we're estimating one categorical percentage). $\hat{p} = \dfrac{240}{400} = 0.60$. $SE = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{0.6(0.4)}{400}} = \sqrt{0.0006} \approx 0.0245$. Margin $= 1.96(0.0245) \approx 0.048$. CI $= 0.60 \pm 0.048 = (0.552,\ 0.648)$. We're 95% confident that between about 55.2% and 64.8% of all voters support the measure.

11. **Test: a one-sample test for a mean, σ known** (so we use z, not the usual t). $H_0 : \mu = 100$ versus $H_a : \mu \neq 100$ (two-sided). $SE = \dfrac{\sigma}{\sqrt{n}} = \dfrac{15}{\sqrt{25}} = \dfrac{15}{5} = 3$. Test statistic $z = \dfrac{106 - 100}{3} = \dfrac{6}{3} = 2.0$. The two-sided p-value is about 0.046, which is less than 0.05, so we **reject** H_0: there is evidence the true mean is not 100 grams.

12. **Test: a two-sample test for two means** (comparing two quantitative groups). $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$. $SE = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} = \sqrt{\dfrac{100}{50} + \dfrac{100}{50}} = \sqrt{2+2} = \sqrt{4} = 2$. Test statistic $t = \dfrac{82 - 78}{2} = \dfrac{4}{2} = 2.0$, with the p-value just under 0.05. So we **reject** H_0: there is a significant difference, with Method A scoring higher on average.

13. **ANOVA** compares three or more means at once. H_0: all four programs have the *same* mean score ($\mu_1 = \mu_2 = \mu_3 = \mu_4$). H_a: *at least one* program's mean is different. Since $p = 0.008 < 0.05$, we **reject** H_0: at least one program's mean differs from the others. What it does *not* tell you: *which* program(s) differ — a significant F only says "not all equal," so you'd need follow-up comparisons to find the specific differences.

14. **Test: a chi-square goodness-of-fit test** (counts across categories vs. what "fair" predicts). If the die is fair, each face is expected $60/6 = 10$ times. H_0: the die is fair (all faces equally likely); H_a: it is not. Compute $\chi^2 = \sum \dfrac{(O-E)^2}{E}$:

Face	Observed O	Expected E	$(O-E)^2/E$
1	8	10	$4/10 = 0.4$
2	12	10	$4/10 = 0.4$
3	9	10	$1/10 = 0.1$
4	11	10	$1/10 = 0.1$
5	15	10	$25/10 = 2.5$
6	5	10	$25/10 = 2.5$
Total	60	60	$\chi^2 = 6.0$

$\chi^2 = 6.0$. Since $6.0 < 11.07$, we **fail to reject** H_0: there is *not* enough evidence to call the die unfair. The wobble in counts is within what chance can produce.

15. (a) $r = 0.85$ is a *strong, positive, linear* relationship: more study hours go with higher scores. (b) $\hat{y} = 2 + 3(4) = 2 + 12 = \mathbf{14}$ points predicted. (c) $r^2 = (0.85)^2 = 0.7225$, so about **72%** of the variation in exam scores is explained by the linear relationship with hours studied (the other $\approx 28\%$ comes from everything else).

You just walked the entire book — naming variables, spotting bias, reading shape, computing z-scores and probabilities, using the empirical rule, finding a standard error, building two kinds of confidence intervals, running one-sample, two-sample, ANOVA, and chi-square tests, and interpreting a regression line. More importantly, you practiced the one move a final really tests: *looking at a bare problem and deciding which tool it wants.* If some were shaky, you now know exactly where to look — and that's what this rehearsal was for. You're ready.

A Final Note

So here you are, at the last page. I want to tell you something I believe completely: you were never "bad at math." You were a person who got handed the hard parts before anyone gave you the warm, plain-language version — and you kept going anyway. That's not a weakness. That's exactly the stubbornness this subject rewards.

Look at what you can do now. You can read a headline and ask where the data came from. You can look at a graph and catch the trick in a cut-off axis. You can hear "the sample mean" or "statistically significant" and calmly translate it into something true and human. You didn't memorize a pile of formulas — you learned to read the world as data, and to ask better questions of it. That's a way of thinking you get to keep for the rest of your life.

Carry this confidence forward. Into the next class, the next decision, the next moment someone waves a scary-looking number at you and hopes you'll just nod. You won't just nod anymore. You are not behind, and you were never broken — you are someone who learned a powerful new language, one careful idea at a time. I'm proud of you. Now go use it.