STATISTICS · STUDY GUIDE

Exam 1 Study Guide

This guide covers the first two chapters. You will learn the vocabulary every later chapter leans on: descriptive and inferential statistics, populations and samples, types of data and levels of measurement. Then you will see where good data comes from, why random samples work, how bias creeps in and what separates an experiment from an observational study. Work one chapter at a time. Read the "what to know" paragraphs, copy the formula tables by hand, then try each worked example with the steps covered. Finish each chapter with its ten self-test items. The cram sheet at the end puts every formula in one place.

WHAT THIS GUIDE COVERS

- Chapter 1 · descriptive versus inferential, population and sample, parameter and statistic, mean, median and range, sampling error versus bias, types of data, levels of measurement, and how statistics mislead.
- Chapter 2 · why random sampling works, the standard error $\sigma \div \sqrt{n}$, the four probability sampling methods, named biases and nonsampling error, observational studies versus experiments, and experimental design.

CHAPTER 1

What Statistics Is

Read it in full: **Chapter 1 in the textbook**. Practice it in the lab: **Population vs Sample** and **Levels of Measurement**.

What to know

Statistics exists because things vary. Two people on the same diet lose different amounts, two polls of the same race disagree, and two identical machines make slightly different parts. Statistics is the science of collecting, organizing, summarizing and drawing conclusions from data in the presence of that variation. Its payoff is a decision you can defend rather than one you merely feel sure about.

The field splits in two. **Descriptive statistics** summarizes the data you have: means, medians, ranges, tables and graphs. **Inferential statistics** uses a sample to make a claim about a larger group you did not measure, and it must report how uncertain that claim is. A descriptive statement cannot be wrong if the arithmetic is right. An inferential one is a bet about people you never measured, and it can be flatly wrong.

The mean is $\bar{x} = \Sigma x \div n$, the sum divided by the count. The median is the middle value once the data are sorted; with an even count it is the average of the two middle values. The range is largest minus smallest. In the book's nine commutes the mean is 31.1 minutes but the median is 24, because one 88 minute day pulls the mean up. The median answers "what is a typical day?" The mean answers "what happens on average over many days?", which is the right question for totals.

A **population** is the whole group your question is about. A **sample** is the part you actually observe. Measuring everyone is a **census**, which is usually too costly, too slow, impossible to reach, or destructive (you cannot crash test every car). A **parameter** describes a population and is fixed but unknown. A **statistic** describes a sample and changes from sample to sample. Greek letters go with populations (μ, σ, p, and N for size) and Roman letters with samples ($\bar{x}$, s, $\hat{p}$, n).

Population variance divides by N. Sample variance divides by $n - 1$, because deviations measured from $\bar{x}$ come out slightly too small on average and dividing by $n - 1$ corrects that tilt. **Sampling error** is the natural wobble between a statistic and the parameter; it is not a mistake and it shrinks as n grows. **Sampling bias** is a systematic tilt from how the sample was chosen, and a bigger sample does not fix it. The 1936 Literary Digest poll got about 2.4 million ballots back and still called the election wrong.

A variable is any characteristic that can differ between individuals. **Qualitative** (categorical) variables place each individual in a category, such as blood type. **Quantitative** (numerical) variables record amounts or counts. Quantitative variables are **discrete** (counted, with gaps, such as number of siblings) or **continuous** (measured, any value in an interval, such as weight). Digits do not make a quantity: ZIP codes, jersey numbers and coded categories are qualitative. The mean of a discrete variable can be a value no one has, such as 1.6 children.

The four **levels of measurement** each add one property. Nominal is unordered categories (mode only). Ordinal adds order but not equal gaps (mode and median). Interval adds equal, meaningful distances but no true zero (mean allowed, ratios not). Ratio adds a true zero, so "twice as much" is real. 20°C is not twice as hot as 10°C: the same pair gives a ratio of 2 in Celsius, 1.36 in Fahrenheit and 1.035 in kelvin. Differences survive a change of units; ratios on an interval scale do not.

Correlation allows four explanations: A causes B, B causes A, a **confounding variable** drives both, or coincidence. Only a randomized experiment singles out cause. Most misleading statistics are true numbers framed badly: truncated axes, cherry-picked start dates, relative risk with no baseline, missing denominators, survivorship bias and averages that hide the spread. A 50% relative risk reduction can be a 0.1 point absolute reduction with a number needed to treat of 1,000.

Two readings to get right forever. A p-value is the chance of data at least as extreme as yours if the null hypothesis were true; it is not the chance the hypothesis is true, and a large p-value does not show there is no effect. "95% confidence" describes the long-run success rate of the procedure, not a 95% chance that this one interval holds the truth. Both tools rely on random sampling, independent observations and a reasonable model.

LOCK THESE IN

- Descriptive describes the data in hand; inferential generalizes and must report uncertainty.
- Parameter goes with population and Greek letters (μ, σ, p, N); statistic goes with sample and Roman letters ($\bar{x}$, s, $\hat{p}$, n).
- The mean is pulled by outliers; the median is not. Report both when they disagree.
- Sampling error shrinks with bigger n. Sampling bias needs better design, not more data.
- Sample variance divides by $n - 1$; population variance divides by N.
- Counted is discrete, measured is continuous; digits used as names are qualitative.
- Nominal, ordinal, interval, ratio: each level adds order, then equal gaps, then a true zero.
- Relative risk reduction = $(r_c - r_t) \div r_c$; absolute = $r_c - r_t$; NNT = 1 ÷ absolute reduction.



Drop the 88 and the mean becomes 24.0, the same as the median.

One extreme value moves the mean, not the median. Nine commutes have a median of 24 minutes, but the single 88 drags the mean up to 31.1.

Formulas and when to use them

Quantity	Formula	When to use it
Mean	$\bar{x} = \Sigma x \div n$ (sample); $\mu = \Sigma x \div N$ (population)	Center of quantitative data; totals and long-run averages
Median	Middle sorted value; average of the two middle values if n is even	Typical value; skewed data or outliers
Range	Largest − smallest	Quickest spread; fragile, set by two values
Population SD	$\sigma = \sqrt{[\Sigma(x - \mu)^2 \div N]}$	You have every member of the population
Sample SD	$s = \sqrt{[\Sigma(x - \bar{x})^2 \div (n - 1)]}$	You have a sample and want to estimate σ
Relative risk reduction	$(r_c - r_t) \div r_c$	How large the change is relative to baseline
Absolute risk reduction	$r_c - r_t$	How many fewer cases per person treated
Number needed to treat	NNT = 1 ÷ ARR	People treated for one to benefit
Celsius to Fahrenheit	°F = (9 ÷ 5)°C + 32	Testing whether a ratio survives a unit change
Celsius to kelvin	K = °C + 273.15	Kelvin has a true zero, so its ratios are real

Symbols and levels of measurement

Item	Population or meaning	Sample or what it allows
Size	N	n
Mean	μ (mu)	$\bar{x}$ (x bar)
Standard deviation	σ (sigma)	s
Proportion	p	$\hat{p}$ (p hat)
Nominal	Unordered categories (blood type, ZIP code)	Counts, percents, mode
Ordinal	Ordered, unequal gaps (letter grade, cancer stage)	Adds the median
Interval	Equal gaps, no true zero (°C, calendar year, IQ)	Adds the mean and differences; no ratios
Ratio	True zero (weight, income, time)	Every statistic, plus meaningful ratios

Worked examples

WORKED EXAMPLE 1 · MEAN, MEDIAN AND RANGE WITH AN OUTLIER

A clinic records patient wait times in minutes: 12, 15, 9, 14, 11, 13, 58, 10. Find the mean, median and range, and say which better describes a typical wait.

STEP 1 · ADD AND DIVIDE

Sum = 142 and n = 8, so $\bar{x}$ = 142 ÷ 8 = **17.75 minutes**.

STEP 2 · SORT

Sorted: 9, 10, 11, 12, 13, 14, 15, 58.

STEP 3 · FIND THE MIDDLE

With n = 8 the median is the average of the 4th and 5th values: (12 + 13) ÷ 2 = **12.5 minutes**.

STEP 4 · RANGE

58 − 9 = **49 minutes**.

STEP 5 · INTERPRET

Seven of the eight waits are 15 minutes or less, yet the mean is 17.75 because of the single 58. The median of 12.5 describes a typical wait. Without the 58 the mean is 12.00, close to the median.

WORKED EXAMPLE 2 · A PARAMETER AND A STATISTIC SIDE BY SIDE

A tiny population has the values 4, 8, 10, 14. Find μ and σ. Then take the sample 4, 14 and find $\bar{x}$ and s.

STEP 1 · POPULATION MEAN

$\mu = 36 \div 4 =$ **9**.

STEP 2 · SQUARED DEVIATIONS

Deviations: $-5, -1, 1, 5$. Squares: 25, 1, 1, 25, which sum to 52.

STEP 3 · DIVIDE BY N

$\sigma^2 = 52 \div 4 = 13$, so $\sigma = \sqrt{13} \approx$ **3.61**.

STEP 4 · THE SAMPLE

$\bar{x} = (4 + 14) \div 2 =$ **9**. Deviations from 9 are -5 and 5; squares 25 and 25, sum 50.

STEP 5 · DIVIDE BY N – 1

$s^2 = 50 \div 1 = 50$, so $s = \sqrt{50} \approx$ **7.07**.

WHAT IT SHOWS

The sample mean happened to land on $\mu = 9$, but $s = 7.07$ is far from $\sigma = 3.61$. Two values is a tiny sample, and statistics from tiny samples wobble a lot. That wobble is sampling error.

WORKED EXAMPLE 3 · RELATIVE RISK, ABSOLUTE RISK AND NNT

An ad says a supplement "cuts your risk by 40%." In the trial the condition struck 10 of 1,000 people on placebo and 6 of 1,000 on the supplement. Find the relative risk reduction, the absolute risk reduction and the number needed to treat.

STEP 1 · THE TWO RISKS

Control: $10 \div 1{,}000 = 0.010$. Treated: $6 \div 1{,}000 = 0.006$.

STEP 2 · RELATIVE REDUCTION

$(0.010 - 0.006) \div 0.010 = 0.004 \div 0.010 =$ **40%**. The ad's arithmetic is honest.

STEP 3 · ABSOLUTE REDUCTION

$0.010 - 0.006 =$ **0.004**, which is 0.4 percentage points.

STEP 4 · NUMBER NEEDED TO TREAT

NNT $= 1 \div 0.004 =$ **250**. Two hundred fifty people take it for one to avoid the condition.

READ IT FAIRLY

"Cuts risk by 40%" and "249 of every 250 people get no benefit" describe the same trial. A fair report gives both numbers.

WORKED EXAMPLE 4 · WHY RATIOS FAIL ON AN INTERVAL SCALE

A friend says 30°C is "twice as hot" as 15°C. Test the claim by converting to Fahrenheit and kelvin.

STEP 1 · THE CELSIUS RATIO

30 ÷ 15 = 2, which looks like "twice."

STEP 2 · FAHRENHEIT

15°C = (9 ÷ 5)(15) + 32 = 59°F and 30°C = (9 ÷ 5)(30) + 32 = 86°F. Ratio: 86 ÷ 59 = **1.46**.

STEP 3 · KELVIN

15°C = 288.15 K and 30°C = 303.15 K. Ratio: 303.15 ÷ 288.15 = **1.052**.

STEP 4 · CONCLUDE

The same two days give ratios of 2, 1.46 and 1.052. A "multiple" that changes with the unit is not real. The difference is real: 15 Celsius degrees is 15 kelvins. Celsius is interval; kelvin, with a true zero, is ratio.

TRAPS PEOPLE FALL INTO

- **Calling a sample summary inferential.** "These 40 patients averaged 128" is descriptive. It turns inferential only when it claims something about patients not measured.
- **Mixing up the symbols.** μ and σ are population parameters; $\bar{x}$ and s are sample statistics. A question that gives $\bar{x}$ is giving you a statistic.
- **Thinking a huge sample cures bias.** Size shrinks sampling error only. A self-selected sample of 500,000 still estimates the wrong group.
- **Averaging labels.** A spreadsheet will average ZIP codes or 1 = single, 2 = married. The result means nothing.
- **Taking ratios on an interval scale.** 20°C is not twice 10°C, and the year 2000 is not twice the year 1000.
- **Reporting only the mean for skewed data.** One extreme value can move the mean far from a typical case. Give the median too.
- **Reading relative risk alone.** "Halves your risk" may mean 2 in 1,000 down to 1 in 1,000. Ask for the baseline and the NNT.
- **Misreading the p-value.** $p = 0.03$ is not a 3% chance there is no effect, and $p = 0.40$ does not prove there is no effect.

Chapter 1 self-test · ten items

1. A school reports that the 28 students in one class averaged 81 on a test. Which kind of statement is this?

A. Descriptive, because it only summarizes the 28 students measured
B. Inferential, because it uses an average of several scores
C. Inferential, because the class is a sample of the school
D. Descriptive only if the class is chosen at random

A. Descriptive, because it only summarizes the 28 students measured It makes no claim beyond the 28 students, so it is descriptive. It would become inferential if it said something about students who were not tested.

2. Which symbol names a population parameter?

A. s
B. σ
C. $\bar{x}$
D. $\hat{p}$

B. σ Greek letters go with populations. σ is the population standard deviation; s, $\bar{x}$ and $\hat{p}$ are computed from a sample.

3. Five daily orders are 6, 9, 7, 30 and 8. What are the mean and median?

A. Mean 8, median 12.0
B. Mean 7.5, median 8
C. Mean 12.0, median 8
D. Mean 12.0, median 7

C. Mean 12.0, median 8 Sum = 60, so $\bar{x} = 60 \div 5 = 12.0$. Sorted: 6, 7, 8, 9, 30, so the median is the 3rd value, 8. The 30 pulls the mean above the median.

4. A population has the values 2, 6 and 10. What is σ?

A. About 4.00
B. Exactly 6.00
C. About 10.67
D. About 3.27

D. About 3.27 $\mu = 6$. Squared deviations 16, 0 and 16 sum to 32. For a population divide by $N = 3$: $\sigma^2 = 10.67$ and $\sigma \approx 3.27$. Dividing by $n - 1$ gives the sample value 4.00.

5. A radio station gets 60,000 call-in votes on a new law. What is the main problem with using them to describe all residents?

A. Sampling error, which is too large with 60,000 people
B. Voluntary response bias, which a large count does not fix
C. The votes are quantitative, so a mean is needed
D. The count is too small to be a valid sample

B. Voluntary response bias, which a large count does not fix Callers chose themselves, and people with strong views are far more likely to call. That is a systematic bias, and more callers only make the wrong answer look precise.

6. Which variable is quantitative and continuous?

A. Number of pets at home
B. Hospital patient ID number
C. Time to run a mile
D. Favorite music genre

C. Time to run a mile Time is measured and can take any value in an interval. Pets are counted (discrete), and an ID number and a genre are labels (qualitative).

7. Finishing place in a race (1st, 2nd, 3rd) is measured at which level?

A. Ordinal
B. Nominal
C. Interval
D. Ratio

A. Ordinal The places have a clear order, but the gaps between them are not equal: the winner may lead by a tenth of a second and second place by five minutes.

8. Why is it wrong to say 40°F is twice as warm as 20°F?

A. Fahrenheit is ordinal, so differences have no meaning
B. Temperatures are discrete, so they cannot be divided
C. Fahrenheit has no true zero, so ratios change with the unit
D. The two values are too close to compare as a ratio

C. Fahrenheit has no true zero, so ratios change with the unit Fahrenheit is an interval scale. Its zero is a chosen landmark, not the absence of heat, so the ratio depends on the scale used. Differences are still meaningful.

9. In a trial, a condition affects 8 of 1,000 untreated and 4 of 1,000 treated people. What is the number needed to treat?

A. 50
B. 250
C. 125
D. 2

B. 250 The absolute risk reduction is $0.008 - 0.004 = 0.004$. NNT = $1 \div 0.004 = 250$. The relative reduction is 50%, which sounds far bigger.

10. Ice cream sales and drownings rise and fall together over a year. What is the best explanation?

A. Eating ice cream causes people to drown
B. Drownings cause people to buy more ice cream
C. The data must be wrong because the link is odd
D. A confounding variable, hot weather, drives both

D. A confounding variable, hot weather, drives both Hot weather raises both ice cream sales and swimming. The two series are correlated, but neither causes the other.

CHAPTER 2

Sampling and Data Collection

Read it in full: **Chapter 2 in the textbook**. Practice it in the lab: **Five Sampling Methods** and **Watch Bias Appear**.

What to know

Almost every number in a report came from a sample. A sample earns its authority from how it was chosen, not from how big it is. Random selection makes every characteristic, measured or not, show up in the sample in about its population share on average. Randomness does not promise a perfect sample every time. It promises that the misses are unbiased and that their size can be calculated.

That size is the **standard error**. For a sample mean, $SE(\bar{x}) = \sigma \div \sqrt{n}$. Because n sits under a square root, you must quadruple the sample to halve the standard error: with $\sigma = 12$ minutes, $n = 36$ gives 2.0, $n = 144$ gives 1.0 and $n = 576$ gives 0.5. The formula assumes independent observations, which in practice means random selection and a sample no bigger than about 10% of the population.

Population size N does not appear in $\sigma \div \sqrt{n}$. A random sample of 1,000 is about as precise for a nation of 300 million as for a town of 30,000. What you need is a big enough spoonful from a well

stirred pot. The Literary Digest's 2.4 million ballots in 1936 lost to George Gallup's much smaller, deliberately chosen sample.

Every method starts from a **sampling frame**, the actual list you draw from. Anyone missing from the frame is invisible. A **probability sample** gives every individual a known, nonzero chance. A **simple random sample** (SRS) makes every possible group of n equally likely, which is stronger than "everyone has an equal chance."

Stratified sampling splits the frame into similar groups (strata) and samples randomly inside every one; it is usually more precise when the strata differ. With proportional allocation each stratum gets $n_h = n \times (N_h \div N)$. **Cluster** sampling randomly picks whole natural groups (blocks, clinics, classrooms) and measures everyone inside; it is cheaper but less precise. **Systematic** sampling takes every k-th unit after a random start, with $k = N \div n$, and fails if the list repeats with a cycle that matches k. Stratified takes some from every group; cluster takes everyone from some groups.

Convenience and **voluntary response** samples are not probability methods. No one's chance of selection is known, so no standard error or margin of error means anything. They can suggest ideas or pilot a questionnaire, nothing more.

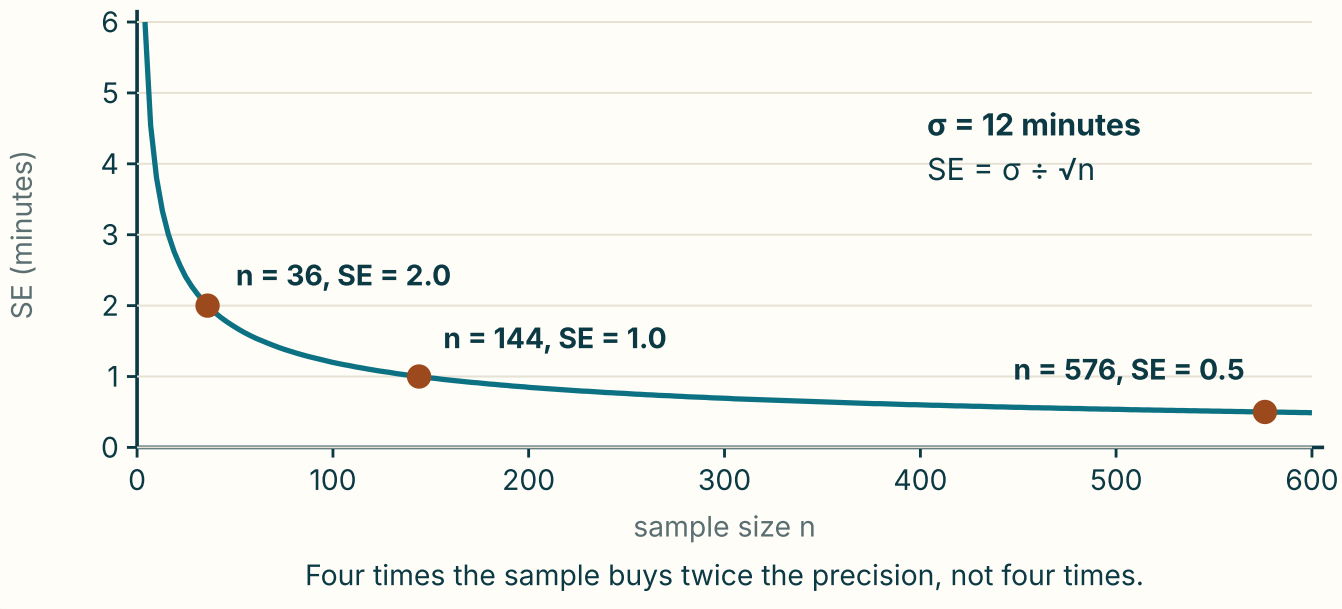
Sampling error is random, centered on zero and shrinks with $1 \div \sqrt{n}$. **Bias** is systematic and does not shrink. The named biases are undercoverage (part of the population cannot be in the frame), voluntary response, nonresponse (selected people do not answer and differ from those who do) and response bias (wording, social desirability, interviewer effects, recall). These plus processing errors are **nonsampling error**, which can strike even a census and usually exceeds sampling error. A margin of error counts chance only, never bias.

In an **observational study** the researcher records exposures people already have; in an **experiment** the researcher assigns the treatment. Observational designs include cross-sectional, prospective cohort and retrospective case-control studies. A **confounder** is linked to both the explanatory variable and the response, so observation alone cannot prove cause. Random sampling buys generalization to the population; random assignment buys cause. They are separate guarantees.

A good experiment uses **control** (a comparison group and constant conditions), **randomization**, **replication** (enough units per treatment) and, when useful, **blocking** (randomizing within groups of similar units). A **placebo** equalizes the belief of being treated. In a single-blind study subjects do not know their group; in a double-blind study neither subjects nor the people who assess them know. Two factors at two levels each make 2 × 2 = 4 treatments.

LOCK THESE IN

- $SE(\bar{x}) = \sigma \div \sqrt{n}$. Four times the sample gives half the standard error.
- Population size does not appear in the SE formula, as long as the sample is under about 10% of it.
- SRS: every group of size n equally likely. Systematic: every k-th after a random start, $k = N \div n$.
- Stratified takes some from every group; cluster takes everyone from some groups.
- Proportional allocation: $n_h = n \times N_h \div N$.
- Bias is systematic and never shrinks with n. A margin of error covers chance only.
- Random sampling licenses generalization; random assignment licenses cause.
- Experiment pillars: control, randomization, replication, blocking; add a placebo and double blinding.



Quadruple n to halve the standard error. With σ = 12 minutes, the standard error falls from 2.0 at n = 36 to 1.0 at n = 144 and 0.5 at n = 576. Population size never appears in the formula.

Formulas and when to use them

Quantity	Formula	When to use it
Standard error of $\bar{x}$	$SE = \sigma \div \sqrt{n}$	Typical miss of a sample mean; random, independent sample
Halving the SE	Multiply n by 4	Planning a sample size
10% condition	$n \le 0.10N$	Checking that draws are close to independent
Proportional allocation	$n_h = n \times (N_h \div N)$	Stratified sample sized to each stratum
Systematic step	$k = N \div n$; picks are start + $k(j - 1)$	Every k-th unit from a list or stream
Sample proportion	$\hat{p}$ = successes ÷ n	Share of a sample with a trait
Margin of error (proportion)	$1.96 \times \sqrt{[\hat{p}(1 - \hat{p}) \div n]}$	95% chance error only; needs 10 successes and 10 failures
Number of treatments	levels of factor 1 × levels of factor 2	Counting arms in a factorial experiment

Methods, biases and designs

Item	What it is	Key risk or rule
Simple random	Every group of n equally likely	Needs a clean, complete frame
Stratified	Random sample inside every stratum	Strata similar inside, different from each other
Cluster	Random whole groups, everyone inside	Cheaper but less precise; clusters should be diverse
Systematic	Every k-th unit after a random start	Periodicity in the list
Convenience	Whoever is easy to reach	No valid inference at any size
Undercoverage	Part of the population is missing from the frame	Systematic; more n does not help
Nonresponse	Selected people do not answer	Systematic; low response rate is a warning
Response bias	Answers are inaccurate as given	Wording, social desirability, recall
Observational study	Exposure observed, not assigned	Association only; confounding
Experiment	Researcher assigns treatment at random	Can show cause

Worked examples

WORKED EXAMPLE 1 · STANDARD ERROR AT THREE SAMPLE SIZES

Systolic blood pressure in a large adult population has $\sigma = 15$ mmHg. Find the standard error of $\bar{x}$ for n = 25, 100 and 400.

STEP 1 · N = 25

SE = 15 ÷ √25 = 15 ÷ 5 = **3.0 mmHg**.

STEP 2 · N = 100

SE = 15 ÷ √100 = 15 ÷ 10 = **1.5 mmHg**.

STEP 3 · N = 400

SE = 15 ÷ √400 = 15 ÷ 20 = **0.75 mmHg**.

STEP 4 · READ THE PATTERN

Each step multiplies n by 4 and halves the SE. Sixteen times the data (25 to 400) buys four times the precision.

CHECK THE CONDITION

The population is large, so 400 is far less than 10% of it, and the draws can be treated as independent.

WORKED EXAMPLE 2 · PROPORTIONAL ALLOCATION AND A SYSTEMATIC PLAN

A university of N = 8,000 students has 3,200 in Engineering, 2,400 in Business, 1,600 in Arts, 800 in Nursing. Plan a stratified sample of n = 200 with proportional allocation, then a systematic sample of the same size.

STEP 1 · THE SAMPLING FRACTION

n ÷ N = 200 ÷ 8,000 = 1/40.

STEP 2 · ALLOCATE

Engineering: 3,200 ÷ 40 = 80; Business: 2,400 ÷ 40 = 60; Arts: 1,600 ÷ 40 = 40; Nursing: 800 ÷ 40 = 20.

STEP 3 · CHECK

80 + 60 + 40 + 20 = **200**. Every student has the same 1/40 chance.

STEP 4 · SYSTEMATIC STEP

k = 8,000 ÷ 200 = 40. Pick a random start from 1 to 40, say 23, then take 23, 63, 103 and so on.

STEP 5 · LAST PICK

The j-th pick is 23 + 40(j − 1), so the 200th is 23 + 40(199) = **7,983**, inside the list.

WORKED EXAMPLE 3 · NONRESPONSE BIAS AGAINST THE MARGIN OF ERROR

A researcher randomly selects 800 adults and asks if they volunteer monthly. Only 320 respond and 208 say yes. A follow-up finds 30% of the nonrespondents volunteer. Compare the bias with the margin of error.

STEP 1 · THE ESTIMATE

$\hat{p}$ = 208 ÷ 320 = **0.65**.

STEP 2 · CHECK THE COUNTS

208 yes and 112 no are both at least 10, so the formula may be used.

STEP 3 · MARGIN OF ERROR

1.96 × √(0.65 × 0.35 ÷ 320) = 1.96 × √0.000711 = 1.96 × 0.0267 ≈ **0.052**, about ±5.2 points.

STEP 4 · THE TRUTH

Nonrespondents: 800 − 320 = 480, and 0.30 × 480 = 144. True share: (208 + 144) ÷ 800 = 352 ÷ 800 = **0.44**.

STEP 5 · COMPARE

Bias = 0.65 − 0.44 = 0.21, or 21 points, about 4.0 times the margin of error. The margin of error never saw it.

WORKED EXAMPLE 4 · A CONFOUNDER BUILDS A FAKE LINK

Among 1,000 people who carry a lighter, 700 smoke; among 1,000 who do not, 100 smoke. Lung disease strikes 8% of smokers and 1% of nonsmokers, and carrying a lighter does nothing. Compare the disease rates.

STEP 1 · LIGHTER CARRIERS

(700 × 0.08) + (300 × 0.01) = 56 + 3 = 59 cases, a rate of **5.9%**.

STEP 2 · NON-CARRIERS

(100 × 0.08) + (900 × 0.01) = 8 + 9 = 17 cases, a rate of **1.7%**.

STEP 3 · THE RATIO

5.9% ÷ 1.7% ≈ 3.47. Carriers look about 3.5 times as likely to get sick.

STEP 4 · NAME THE CAUSE

Smoking is linked to both carrying a lighter and to disease. It is the confounder. More data would only repeat the same tangle; random assignment is what breaks it.

TRAPS PEOPLE FALL INTO

- **Thinking precision grows with n.** It grows with √n. Doubling n cuts the SE by only about 29%; you need four times the n to halve it.
- **Sizing a sample as a share of the population.** N is not in σ ÷ √n. A random 1,000 works for a country as well as a town.
- **Calling systematic sampling an SRS.** Everyone has an equal chance, but most groups of size n can never be chosen, so it is not an SRS.
- **Swapping stratified and cluster.** Stratified: some from every group. Cluster: all from some groups.
- **Trusting the margin of error to cover bias.** It counts chance only. Nonresponse and undercoverage can be many times larger.
- **Using causal verbs for observational findings.** Rewrite "cuts" or "boosts" as "is associated with" unless the treatment was randomly assigned.
- **Mixing up the two randomizations.** Random sampling gives generalization. Random assignment gives cause.
- **Thinking randomization is about fairness.** Its job is to balance every variable, including ones nobody measured.

Chapter 2 self-test · ten items

1. Heights in a large population have σ = 8 cm. What is the standard error of $\bar{x}$ for n = 64?

A. 0.125 cm
B. 8.0 cm
C. 1.0 cm
D. 2.0 cm

C. 1.0 cm SE = σ ÷ √n = 8 ÷ √64 = 8 ÷ 8 = 1.0 cm.

2. A survey team wants to cut its standard error in half. By what factor must it multiply the sample size?

A. 2
B. √2
C. 4
D. 8

C. 4 SE = σ ÷ √n, so halving the SE needs √n to double, which means n must be four times as large.

3. A town of 20,000 and a nation of 200 million are each surveyed with a random sample of 1,000. How do the standard errors compare?

A. The town sample is more precise, as a bigger share
B. About the same, because N is not in the formula
C. The nation sample is more precise, with more people
D. They cannot be compared, since the sizes differ

B. About the same, because N is not in the formula As long as the sample is a small share of the population, precision depends on n alone. Both samples are under 10% of their populations.

4. A researcher randomly picks 12 of a county's 150 schools and surveys every teacher in them. What design is this?

A. Stratified sampling
B. Systematic sampling
C. Cluster sampling
D. Convenience sampling

C. Cluster sampling Whole natural groups are chosen at random and everyone inside is measured. That is cluster sampling.

5. A clinic has 600 adults, 300 teens and 100 children and wants a proportional stratified sample of 50. How many teens?

A. 15
B. 17
C. 30
D. 10

A. 15 The fraction is 50 ÷ 1,000 = 1/20, so teens get 300 ÷ 20 = 15. Adults get 30 and children 5, for 50 in all.

6. A poll uses only landline numbers when many adults have only cell phones. Which bias is this?

A. Response bias
B. Voluntary response
C. Sampling error
D. Undercoverage

D. Undercoverage Cell-only adults cannot appear in the frame at all. Missing part of the population from the frame is undercoverage.

7. Which statement about bias is correct?

A. It shrinks in proportion to $1 \div \sqrt{n}$
B. It does not shrink as n grows
C. It is included in the margin of error
D. It averages out to zero over many samples

B. It does not shrink as n grows Bias is systematic. It points the same way every time, so a larger sample only gives a tighter estimate of the wrong value.

8. Researchers record that people who walk daily have lower blood pressure. Why can they not conclude walking lowers blood pressure?

A. The sample size of an observational study is always too small
B. Blood pressure is an ordinal variable, so a mean is not allowed
C. Walking was measured, so it cannot be an explanatory variable
D. Walkers may differ in diet, age or health, which could confound it

D. Walkers may differ in diet, age or health, which could confound it The researchers did not assign walking. People who walk may differ in many ways that also affect blood pressure, so the link may be confounded.

9. An experiment tests 3 doses of a drug at 2 times of day. How many treatments are there?

A. 5
B. 6
C. 3
D. 2

B. 6 Treatments are combinations of factor levels: $3 \times 2 = 6$.

10. In a double-blind trial, who does not know which treatment each subject received?

A. The subjects and the people who assess their outcomes
B. Only the subjects who are receiving the new treatment
C. Only the statistician who later analyzes the data
D. Only the doctors who write the drug prescriptions

A. The subjects and the people who assess their outcomes Double blinding hides the assignment from both subjects and the staff who interact with them and score the response, so expectation cannot tilt the results.

ONE PAGE

Cram sheet

Everything from chapters 1 and 2 in one place. Cover the right column and see if you can rebuild each row.

Formulas

Name	Formula	Note
Mean	$\bar{x} = \Sigma x \div n$; $\mu = \Sigma x \div N$	Pulled by outliers
Median	Middle sorted value	Average the two middle values when n is even
Range	Max − min	Set by the two extremes
Variance	$\sigma^2 = \Sigma(x - \mu)^2 \div N$; $s^2 = \Sigma(x - \bar{x})^2 \div (n - 1)$	SD is the square root
Risk	$RRR = (r_c - r_t) \div r_c$; $ARR = r_c - r_t$	NNT = 1 ÷ ARR
Standard error	$SE(\bar{x}) = \sigma \div \sqrt{n}$	4 × n halves it
Allocation	$n_h = n \times N_h \div N$	Proportional stratified sample
Systematic	$k = N \div n$	Random start from 1 to k
Margin of error	$1.96 \times \sqrt{[\hat{p}(1 - \hat{p}) \div n]}$	Chance error only

Rules that are not formulas

Topic	The rule
Symbols	Greek for populations (μ, σ, p, N); Roman for samples ($\bar{x}$, s, $\hat{p}$, n).
Error versus bias	Sampling error shrinks with n; bias needs better design.
Data types	Qualitative or quantitative; quantitative is discrete (counted) or continuous (measured).
Levels	Nominal (mode), ordinal (median), interval (mean, no ratios), ratio (all, ratios valid).
Correlation	Four explanations: cause, reverse cause, confounder, coincidence.
Sampling methods	Stratified: some from every group. Cluster: all from some groups.
No inference	Convenience and voluntary response samples.
Two randomizations	Random sampling generalizes; random assignment shows cause.
Experiment	Control, randomization, replication, blocking, placebo, double blinding.
P-value and CI	Neither is the probability that a hypothesis or parameter value is true.

Dr. Garg Learning · CC BY-SA 4.0