



CLASS NOTES • CHAPTER 3

From Sample to Population

Probability, the normal curve, and how far an estimate can be trusted

Statistics & Data Analysis Fundamentals
Stat.BD

Mentor: Muhtasim Munif Fahim, MSc

CHAPTER 3

From Sample to Population

A class is a sample of a university. This chapter measures how far a sample's average can stray from the truth, and turns that measurement into a range of plausible values.

The simulated class of 120 students sleeps an average of 6.80 hours on a weeknight. Suppose another class of 120 at the same university filled in the same survey. Would it average exactly 6.80 hours? Almost certainly not: different students give different answers, and their average lands somewhere nearby. The useful question is how far away it could plausibly land. A tenth of an hour? Half an hour? A whole hour? Section 3.5 gives the answer. If the university's students vary as much as this class does, about 95 classes in 100 would average within 0.17 hours of the university's true average, which is about 10 minutes. That is a small number, and the reasons it is small are the subject of this chapter.

Chapter 1 separated the population a question is about from the sample that answers it, and named the gap between them sampling error. Chapter 2 described the sample. This chapter measures the gap. It needs three tools, and builds them in order. Probability is the language of chance, and the chapter starts there, reading probabilities off the class's own cross-tabulation. The normal curve is the shape that chance produces again and again, and the z -score is the ruler that goes with it. The standard error measures how much a sample's average moves from one sample to the next. The three combine into the confidence interval: an estimate with its uncertainty attached. The chapter ends by computing an interval that a published Bangladeshi survey did not print, and by listing what such an interval cannot see.

The data are the class survey described in Chapter 1. Until the real answers are coded, every figure drawn from it comes from a **simulated** stand-in of 120 students generated by a program from a fixed seed. They are not real students. The repeated samples in section 3 are drawn from that stand-in by the same program. A few examples are constructed on purpose, with made-up numbers, and are labelled as constructed where they appear.

What this chapter establishes

1. What a probability is, and how to read one from a table as a count over a count.
2. The complement and addition rules, and the one condition under which probabilities may be multiplied.
3. How “given that” changes the denominator, and why $P(A | B)$ is not $P(B | A)$.
4. Why coincidences are far more common than intuition expects.
5. What the normal curve is, and what the 68–95–99.7 rule says.
6. How to compute a z -score, compare values measured on different scales, and turn a z -score into a share of the curve.

7. How to check whether the normal curve fits, and what goes wrong when it does not.
8. Why sample means vary, and how the standard error measures by how much.
9. What the central limit theorem promises, and what it does not.
10. How to compute a 95% confidence interval for a mean and for a proportion by hand, and what “95% confident” means.
11. What sets the width of an interval, how to plan a sample size, and which errors an interval does not cover.

Notation

$P(A)$	the probability of event A
$P(\text{not } A)$	the probability that A does not happen: the complement
$P(A \text{ and } B)$	the probability that A and B both happen
$P(A \text{ or } B)$	the probability that A happens, or B , or both
$P(A B)$	the probability of A given that B has happened; the bar is read “given”
N, n	the size of a population, and of a sample
μ, σ	the population mean and SD: parameters, fixed and usually unknown
$\bar{x}, s$	the sample mean and SD: statistics, which change from sample to sample
$\pi, \hat{p}$	a population proportion, and a sample proportion (Chapter 1)
z	a z -score: a distance from the mean, counted in SDs
SE	a standard error: the SD of a statistic over repeated samples
t, df	a t multiplier, and its degrees of freedom
CI	a confidence interval
E	a margin of error: half the width of an interval

Contents

What this chapter establishes	1
Notation	2
1 Probability	5
1.1 A probability is a long-run share	5
1.2 The class as a bag of names	5
1.3 “And” and “or”	6
1.4 “Given that” changes the denominator	6
1.5 The swapped conditional	7
1.6 Independence	8
1.7 The coin has no memory	8
1.8 The birthday problem	9
2 The normal curve	9
2.1 A smooth curve over the class	9
2.2 The 68–95–99.7 rule	10
2.3 The z-score	11
2.4 The standard normal curve	11
2.5 Comparing across scales	11
2.6 Reading areas	12
2.7 Checking the curve against the class	13
2.8 When the normal curve is the wrong model	13
3 Samples vary	14
3.1 The class as a population	14
3.2 Five samples, five means	14
3.3 Building a sampling distribution by hand	14
3.4 A thousand samples	16
3.5 The standard error	17
3.6 The square-root law	17
3.7 SD or SE?	18
3.8 The central limit theorem	18

3.9	How unusual is a sample mean?	19
3.10	Scatter, not bias	19
4	Confidence intervals	20
4.1	Turning the picture round	20
4.2	An interval for the mean, by hand	21
4.3	The t distribution	21
4.4	What “95% confident” means	22
4.5	An interval is not a range of people	22
4.6	What sets the width	23
4.7	An interval for a proportion	23
4.8	The margin of error in a poll	24
4.9	How many to sample	25
4.10	Reporting an estimate	25
4.11	An interval that answers a question	26
5	Reading a published prevalence	26
5.1	Islam 2020: an interval the paper did not print	26
5.2	What the interval leaves out	26
6	Estimating a mean or a proportion: the procedure	27
7	Chapter summary	28
8	Key terms	29
9	Exercises	30
10	Solutions	32
11	Our arithmetic and constructed numbers	36
12	References and further reading	36
13	For students in the Stat.BD course	37

1 Probability

1.1 A probability is a long-run share

A fair coin lands heads with probability one half. That sentence cannot be a promise about the next toss, which will be heads or tails and never half of each. It is a statement about what happens over many tosses. Figure 1 tracks the share of heads in 400 tosses of a simulated fair coin. The first ten were H T T T H H T H H H, a share of 0.60. After 100 tosses the share was 0.55; after 400 it was 0.50. The line wanders early and settles late. Ten tosses are too few to see one half, and six heads in ten is not evidence that the coin is unfair. That settling is what a probability describes.

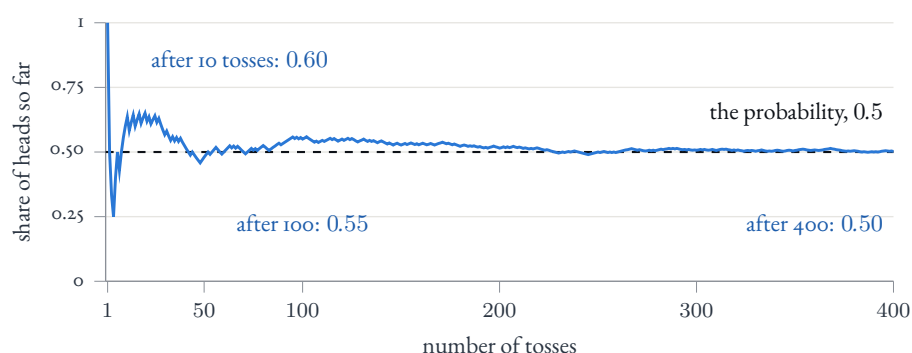


Figure 1: The running share of heads in 400 tosses of a fair coin. **Simulated** in R from a fixed seed; not the class data.

Definition 1.1 • Probability

The **probability** of an event is the share of times it happens in a very long run of repeats under the same conditions. It lies between 0, an event that never happens, and 1, an event that always happens.

The early wobble in Figure 1 is not a flaw in the coin. It is sampling variability, the same thing that makes one class's average differ from another's, and section 3 measures it.

1.2 The class as a bag of names

Chapter 2 cross-tabulated residence against tuition. Imagine the 120 names written on slips and put in a bag, and one slip drawn without looking. Every student has the same chance of being drawn, so the probability of any event is a count divided by a count: the students who fit, over the students who could have been drawn. Figure 2 marks both. The filled cell is the numerator; the boxed cell is the denominator. The same code is used for every probability in this section.

$$P(A) = \frac{\text{number of students for whom } A \text{ is true}}{\text{number of students who could be drawn}} \quad (1.1)$$

For living in a hall, $P(\text{hall}) = 47 \div 120 = 0.392$. The chance that the student does *not* live

	No tuition	Tuition	Total
Hall	30	17	47
Mess	11	32	43
Family home	20	10	30
Total	61	59	120

Figure 2: Residence by tuition in the **simulated** class, as a bag of names. Filled: the students wanted, those in a hall. Boxed: every student who could be drawn.

in a hall is everything else:

$$P(\text{not } A) = 1 - P(A), \quad P(\text{not hall}) = 1 - 0.392 = 0.608. \quad (1.2)$$

This complement rule is the most useful rule in probability, because “not” is often far easier to count than “is”. The birthday problem in section 1.8 runs on it entirely. Its basis is that the options cover every student exactly once, so their probabilities add to 1: $(47 + 43 + 30) \div 120 = 1$. This is Chapter 1’s rule of mutually exclusive and exhaustive options, now written as probabilities.

Complement rule

$P(\text{not } A) = 1 - P(A)$: the chance an event does not happen

1.3 “And” and “or”

“And” asks for both at once. A student who lives in a hall *and* gives tuition sits in one cell of the table, so $P(\text{hall and tuition}) = 17 \div 120 = 0.142$. “Or” asks for either, or both. The students who live in a hall or give tuition fill four cells: the whole hall row and the rest of the tuition column. Counting them once each gives 89 students, so $P(\text{hall or tuition}) = 89 \div 120 = 0.742$.

The addition rule gets the same answer from the three probabilities without redrawing the table:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \quad (1.3)$$

Here $0.392 + 0.492 - 0.142 = 0.742$. The subtraction matters. The 17 hall students who give tuition are in the hall row and in the tuition column, and adding the two probabilities counts them twice. When two events cannot happen together, such as living in a hall and living in a mess, the overlap is zero and the probabilities simply add. Such events are called mutually exclusive.

Addition rule

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

1.4 “Given that” changes the denominator

Suppose the student drawn is known to live in a mess. Only the mess row can have been drawn, and the bag shrinks to its 43 students. Figure 3, left, boxes that row. Of those students 32 give tuition, so

$$P(\text{tuition} \mid \text{mess}) = \frac{32}{43} = 0.744.$$

This is a conditional probability, and it is nothing new. Chapter 2 called it a row percentage and found 74.4%. The new thing is the notation, which makes the denominator explicit.

Conditional probability

the probability of A among only those cases where B is true, $P(A \mid B)$

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\text{number with } A \text{ and } B}{\text{number with } B} \quad (1.4)$$

$A | B$ “ A given B ”: B is known to be true, and the question is about A

$P(B)$ the new denominator: only the cases where B is true

The two forms of equation (1.4) agree exactly, because the 120 in each probability cancels. By hand, the count form is safer: the rounded probabilities can drift in the third decimal. Table 1 gives the tuition share within each residence group.

Table 1: The share giving tuition, overall and given each residence. **Simulated** class data; the probabilities are our arithmetic from the cross-tabulation.

Among	Give tuition	Students	Probability
Everyone	59	120	0.492
Hall residents	17	47	0.362
Mess residents	32	43	0.744
Family home	10	30	0.333

Tuition is far more common among mess residents in this class. Whether a difference like that could be produced by chance alone is the question of the chi-square test in Chapter 5.

	No tuition	Tuition	Total
Hall	30	17	47
Mess	11	32	43
Family home	20	10	30
Total	61	59	120

$P(\text{tuition} | \text{mess})$: the bag is the mess row

	No tuition	Tuition	Total
Hall	30	17	47
Mess	11	32	43
Family home	20	10	30
Total	61	59	120

$P(\text{mess} | \text{tuition})$: the bag is the tuition column

Figure 3: The same filled cell with two different denominators. **Simulated** class data.

1.5 The swapped conditional

Turn the question round. Of the students who give tuition, what share live in a mess? The numerator is the same 32 students, but the bag is now the tuition column of Figure 3, right:

$$P(\text{mess} | \text{tuition}) = \frac{32}{59} = 0.542, \quad P(\text{tuition} | \text{mess}) = \frac{32}{43} = 0.744.$$

Same numerator, different bag, different answer. Confusing the two is the most common probability mistake in newspapers, courtrooms and clinics.

Example 1.1 • Nervous before a viva

A **constructed** department holds a viva for 200 students. Nearly everyone is nervous beforehand: 180 of the 200. Twenty students fail, and 18 of those 20 were nervous. So

$$P(\text{nervous} | \text{failed}) = \frac{18}{20} = 0.90.$$

Example 1.1 • continued

“Nine in ten of those who failed were nervous” sounds like a warning about nerves. Now swap the condition. Of the 180 nervous students, 18 failed; of the 20 calm students, 2 failed:

$$P(\text{failed} \mid \text{nervous}) = \frac{18}{180} = 0.10, \quad P(\text{failed} \mid \text{calm}) = \frac{2}{20} = 0.10.$$

Nervous and calm students fail at exactly the same rate. The first figure was high only because nearly everyone is nervous, whether they pass or fail.

Check which way round

Before accepting any “most A were B ”, ask what the denominator is. “Most accidents happen near home” is true, and says little about the danger of home roads, because most driving happens near home.

1.6 Independence

Two events are independent when knowing one tells nothing about the other. In the language of this section, $P(A \mid B) = P(A)$. If residence told nothing about tuition, the tuition share would be 0.492 in every row of Table 1. It is not. For independent events, and only for them, the probability of both is the product:

$$P(A \text{ and } B) = P(A) \times P(B) \quad (\text{independent events only}). \quad (1.5)$$

If residence and tuition were independent, the probability of a hall student who gives tuition would be $0.392 \times 0.492 = 0.193$. Out of 120 students, that is about $47 \times 59 \div 120 = 23.1$ students. The table shows 17. Fewer hall students give tuition than independence predicts, and more mess students do.

The figure of 23.1 has a name, the expected count: what a cell would hold if the two variables were unrelated. Chapter 5’s chi-square test compares every cell of a table with its expected count, and asks whether the gaps are larger than chance would produce.

Multiplying when events are linked

Equation (1.5) is often used where it does not apply. The probability that a student both lives in a mess and gives tuition is not 0.358×0.492 , because in this class the two go together. Multiplying the probabilities of linked events can be wrong by a wide margin.

1.7 The coin has no memory

A student has guessed “false” on five true/false questions in a row, in a **constructed** quiz. Is the sixth answer “due” to be “true”? No. If the questions were set independently, the sixth is as likely to be “false” as it ever was. The belief that chance corrects itself is the gambler’s fallacy, and Figure 1 seems at first to support it: the line does settle at one half.

It settles for a different reason. An early surplus of heads is not cancelled by later tails; it is

Independent

knowing that one event happened does not change the probability of the other

Expected count

the count a cell of a table would hold if the two variables were independent: $\text{row total} \times \text{column total} \div \text{grand total}$

Gambler’s fallacy

the belief that a run of one outcome makes the other outcome “due”



diluted by later tosses. Two heads more than expected in 10 tosses give 7 of 10, a share of 0.70. The same two extra heads in 400 tosses give 202 of 400, a share of 0.505. The surplus never goes away. It stops mattering, because it is divided by an ever larger number. The same dilution is why larger samples give steadier averages in section 3.

1.8 The birthday problem

In a room of 23 people, what is the chance that at least two share a birthday? Most estimates are far too low. Counting all the ways two or more people can match is hard; counting the ways that nobody matches is one product, and the complement rule does the rest. Assume 365 equally likely birthdays, ignoring 29 February and twins. The first person can have any day; the second must avoid one day, the third two days, and so on:

$$P(\text{no shared birthday among 23}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{343}{365},$$

and the chance of at least one match is 1 minus that product. Table 2 gives the result for several room sizes.

Table 2: The chance that at least two people in a room share a birthday, assuming 365 equally likely days. Our arithmetic.

People	Pairs of people	Chance of a match
10	45	11.7%
23	253	50.7%
30	435	70.6%
50	1225	97.0%

The surprise comes from imagining someone matching *one's own* birthday. But 23 people form 253 pairs, and any pair can match. Real birthdays are not spread perfectly evenly over the year, and an uneven spread only makes a match more likely, so the table understates the chance slightly. The lesson reaches well beyond birthdays. Given enough comparisons, some coincidence is close to certain. Research that runs many comparisons will find some that look remarkable by chance alone, and Chapter 4 returns to this as the problem of multiple testing.

2 The normal curve

2.1 A smooth curve over the class

Figure 4 is Chapter 2's histogram of weeknight sleep, with a smooth curve laid over it. The curve has the same mean and SD as the class, 6.80 and 0.93 hours. It is not a perfect fit, and it does not need to be. It is a model: a smooth description of lumpy data, useful because so much follows from it.

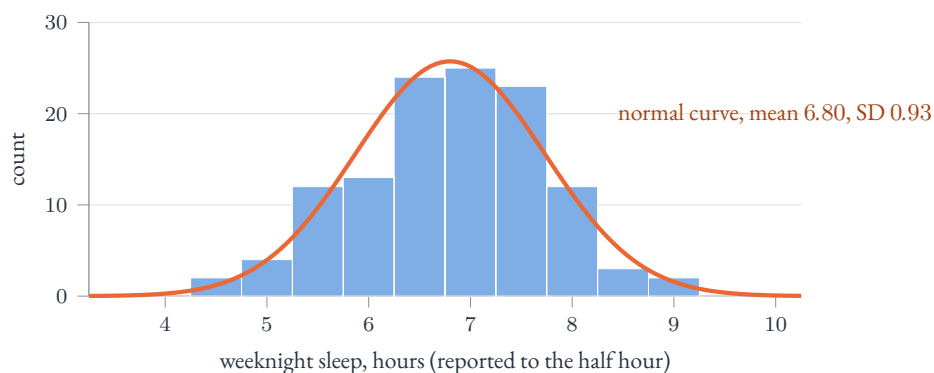


Figure 4: Weeknight sleep in the **simulated** class, with a normal curve of the same mean and SD. The curve is scaled to counts: $n \times \text{bar width} \times \text{the normal density}$.

Definition 2.1 • Normal distribution

A **normal distribution** is a symmetric, single-peaked, bell-shaped curve whose mean, median and mode coincide. Two numbers fix it completely: its mean, which sets where it sits, and its SD, which sets how wide it spreads.

Many measurements come out roughly normal: heights, measurement errors, scores built by adding many small items. Many do not, and section 2.8 shows what happens when the curve is used on one of those. The main reason the curve matters in statistics, though, is not that data follow it. It is that *averages* follow it, a fact that section 3.8 states as the central limit theorem.

2.2 The 68–95–99.7 rule

Chapter 2 gave a rule of thumb: for roughly symmetric data, most values lie within two SDs of the mean. The normal curve is where that rule comes from, and it makes it exact. For any normal curve, 68.3% of the area lies within one SD of the mean, 95.4% within two SDs, and 99.7% within three (Figure 5). This is the 68–95–99.7 rule, and it holds whatever the mean and SD.

68–95–99.7 rule
the shares of a normal curve within one, two and three SDs of its mean

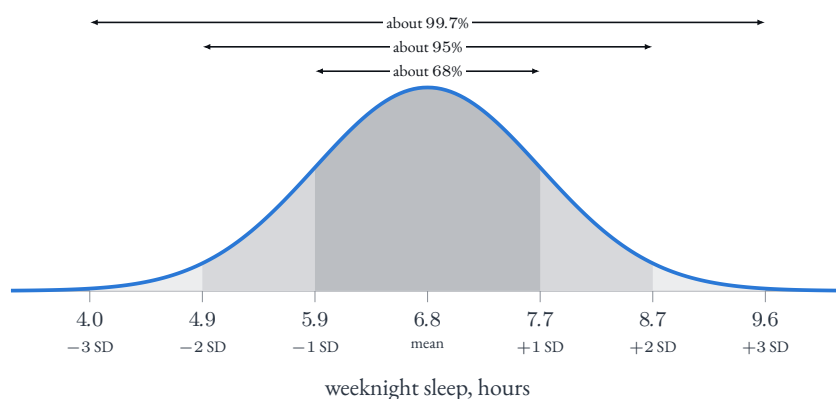


Figure 5: The normal curve with the mean and SD of the **simulated** class, and its one-, two- and three-SD bands. Band endpoints are computed from the rounded mean and SD.

For weeknight sleep, one SD either side of the mean runs from 5.87 to 7.73 hours, and two

SDs from 4.94 to 8.66. The class puts 70.8% of its students in the first band and 96.7% in the second, close to the curve's 68.3 and 95.4. Real data rarely match the rule exactly, and nobody expects them to. The question is whether they come close enough for the curve to be a useful description.

2.3 The z-score

A value's distance from the mean means little on its own. Two hours below the mean is enormous for weeknight sleep and trivial for monthly screen time. The z -score measures the distance in SDs instead:

$$z = \frac{x - \text{mean}}{\text{SD}} \quad (2.1)$$

x the value

mean, SD of the group the value is compared with

z its sign says which side of the mean; its size says how many SDs away

z -score

a value's distance from the mean, counted in SDs; negative below the mean, positive above

Example 2.1 • Two sleepers

A student who sleeps 4.5 hours has

$$z = \frac{4.5 - 6.80}{0.93} = -2.47.$$

That is nearly two and a half SDs below the mean, outside the two-SD band: an unusual value. A student who sleeps 8.5 hours has $z = (8.5 - 6.80) \div 0.93 = +1.83$, inside the two-SD band but near its edge.

2.4 The standard normal curve

A z -score has no units. Measure the short sleeper in minutes instead of hours: 4.5 hours is 270 minutes, and Chapter 2 converted the mean and SD to 408 and 55.8 minutes. Then $z = (270 - 408) \div 55.8 = -2.47$, the same as before. Minutes divided by minutes leaves a pure number, and a change of scale cancels out.

That has a large consequence. Converting every value of a normal variable to z turns its curve into one particular curve, with mean 0 and SD 1: the standard normal. Every normal variable, in any unit, becomes the same curve. One table of areas, or one software function, then serves every normal variable there is.

Standard normal

the normal curve with mean 0 and SD 1; the curve of z -scores

2.5 Comparing across scales

Which is more unusual: a student who sleeps 5 hours, or a student whose stress score is 22? Hours and score points cannot be compared directly. z -scores can. The stress score is the five-item total from Chapter 1, with a possible range of 5 to 25; the class mean is 15.04 and the SD is 4.57, from 118 students who answered every item.

$$\text{Sleep: } z = (5.0 - 6.80) \div 0.93 = -1.94,$$

$$\text{Stress: } z = (22 - 15.04) \div 4.57 = +1.52.$$

The short sleeper is further from typical. On the normal curve about 2.6% of students would sleep less than 5 hours, and about 6.4% would score above 22 (Figure 6).

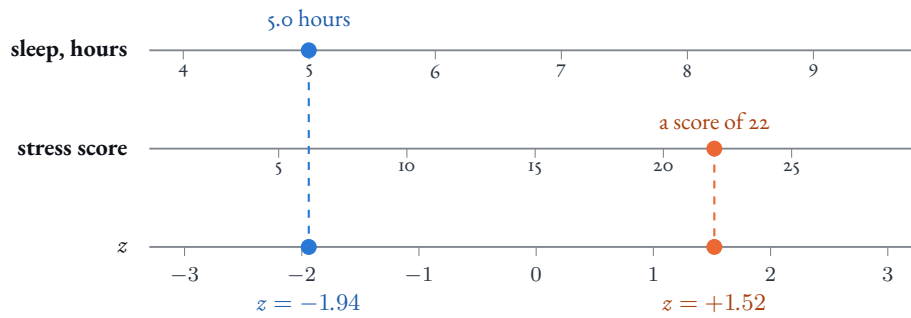


Figure 6: Five hours of sleep and a stress score of 22, each on its own normal curve. **Simulated** class data; the areas are our arithmetic.

The comparison can run the other way too. A stress score as far below the mean as 5 hours of sleep is would be $15.04 + (-1.94) \times 4.57 = 6.2$, and one as far above would be 23.9. Scores that far out, on either side, are as unusual as 5 hours of sleep.

2.6 Reading areas

On any curve of this kind, the area under the curve to the left of a value is the share of the population below it. For the standard normal, those areas come from a table, or from software. Figure 7 shades the area to the left of $z = 1$, and Table 3 lists the areas this chapter uses.

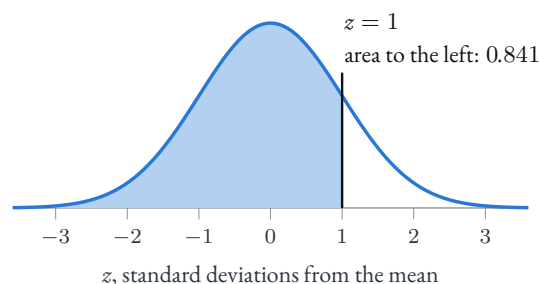


Figure 7: The standard normal curve, with the area to the left of $z = 1$ shaded.

Table 3: Areas to the left of selected z -scores on the standard normal curve, rounded to three decimals. The first two rows are the sleepers of sections 2.3 and 2.5.

z	Area to the left	Reading
-2.47	0.007	sleeping 4.5 hours
-1.94	0.026	sleeping 5 hours
-1.00	0.159	one SD below the mean
0.00	0.500	the mean: half on each side
1.00	0.841	one SD above the mean
1.645	0.950	leaves 5% above: used for 90% intervals
1.96	0.975	leaves 2.5% above: used for 95% intervals
2.576	0.995	leaves 0.5% above: used for 99% intervals

Two rules turn left areas into any other area. The area to the right of a value is 1 minus the area to its left: above $z = 1$ lies $1 - 0.841 = 0.159$. The area between two values is the

difference of their left areas: between $z = -1$ and $z = 1$ lies $0.841 - 0.159 = 0.682$, the 68 of the 68–95–99.7 rule, up to rounding. By symmetry, the area to the left of $-z$ equals the area to the right of $+z$.

In Excel, `NORM.S.DIST(z, TRUE)` gives the area to the left of z , and `NORM.S.INV(area)` goes back from an area to z . In R the same pair are `pnorm(z)` and `qnorm(area)`.

2.7 Checking the curve against the class

A model earns trust by agreeing with the data it describes. For a student sleeping 8.5 hours, $z = +1.83$ and the area to the left is 0.966: the curve says 96.6% of students sleep less. The class has 95.8% reporting less than 8.5 hours and 98.3% reporting 8.5 or less. The curve falls between the two, which is as close as it can come: it cannot see that every answer was given to the half hour.

At the other end, 4.5 hours has $z = -2.47$ and an area of 0.007. The curve says 0.68%, fewer than one student in 120. The class has 2 students at 4.5 hours or less. Real data often have fatter tails than the normal curve: more extreme values than it predicts. The curve describes the middle well and the edges less well, and conclusions that rest on the far tails deserve extra caution.

2.8 When the normal curve is the wrong model

Screen time is skewed to the right, as Chapter 2 found: most students cluster at a few hours, and a long tail stretches upward. Figure 8 lays a normal curve with the same mean and SD, 4.81 and 2.31 hours, over the histogram. The curve is symmetric and the data are not, so it misplaces the peak and spills below zero. For zero hours, $z = (0 - 4.81) \div 2.31 = -2.08$, and the area to the left is about 1.9%. The curve therefore expects about 2.2 of the 117 students who answered to spend *less than no time* on their phones.

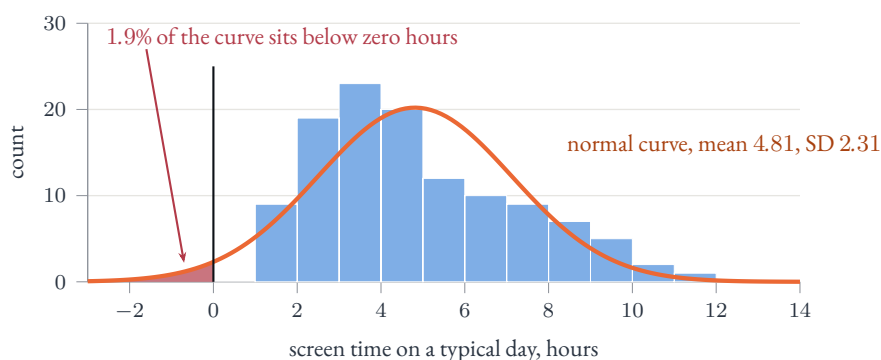


Figure 8: Daily screen time in the **simulated** class, with a normal curve of the same mean and SD. The curve is the wrong shape and extends below zero.

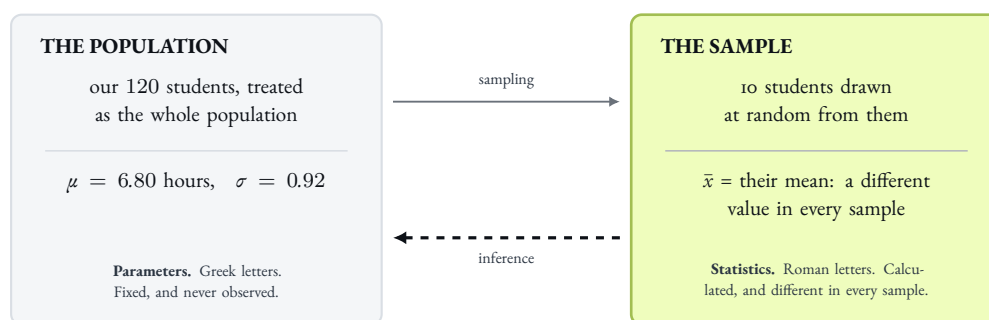
Look before using an area

Normal areas are only as good as the fit of the normal curve. Draw the histogram first. If the data are clearly skewed, a normal area will be wrong in both tails, and percentiles should be read from the data themselves. Chapter 4 meets methods built for such data.

3 Samples vary

3.1 The class as a population

Chapter 1 defined a *parameter* as a number describing a population, and a *statistic* as a number computed from a sample, and called the gap between them sampling error. Measuring that gap requires knowing the parameter, which in real research is never known. This section therefore makes one pretend. For its length, the 120 simulated students are the entire population. Their mean weeknight sleep is then a parameter, $\mu = 6.80$ hours, and samples can be drawn from them and compared with it (Figure 9).



Everything measured lives on the right. Everything the study is actually about lives on the left. The dashed arrow is the whole of statistical inference, and it is the only one of the two that can be wrong.

Figure 9: For this section only, the **simulated** class is the population, and samples of 10 are drawn from it.

One detail changes. Chapter 2 divided by $n - 1$ to find the SD of a sample, because a sample's own mean sits a little too close to its values. A population's SD, written σ , divides by N , the whole population. For the class that gives $\sigma = 0.92$ rather than the sample SD of 0.93. The difference is small here, and it vanishes as N grows.

3.2 Five samples, five means

Figure 10 shows five random samples of 10 students drawn from the population. Their means are 7.25, 6.20, 6.65, 7.05, 6.85 hours. Each is an honest random sample, and none is wrong. They differ because each drew different students. This is sampling variability, the scatter that Chapter 1 predicted and that Figure 1 showed in coin tosses.

Sampling variability
the variation of a statistic from one random sample to another

3.3 Building a sampling distribution by hand

The scatter can be seen by hand. Table 4 lists the weeknight sleep of all 120 simulated students, ten to a row. The row label plus the column label gives a student's id: row 50, column +7 is student 57.

1. Draw 10 ids from 1 to 120 with a random number tool, such as a phone app or a spreadsheet's `RANDBETWEEN(1, 120)`. If an id comes up twice, draw again.
2. Look up the 10 values and average them.

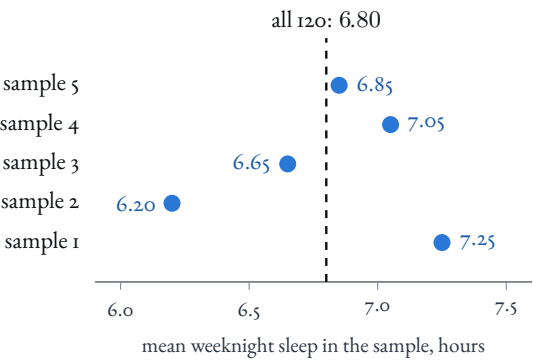


Figure 10: Five random samples of 10 from the **simulated** class, each with its mean. Drawn by data/class-data.R.

Table 4: Weeknight sleep in hours for every student in the **simulated** class, by id. The same grid is used in class, and in Exercises 21 and 22.

ids	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
0	7.0	7.5	6.5	8.0	7.5	7.5	8.0	8.0	7.0	7.5
10	7.5	8.5	5.5	6.5	5.5	7.0	6.5	9.0	5.5	5.5
20	7.0	6.0	5.0	7.0	8.0	5.5	6.5	5.5	7.5	7.0
30	7.0	6.5	7.5	7.0	7.0	8.5	7.0	6.0	6.5	6.5
40	7.0	7.0	6.5	7.5	7.0	7.5	6.0	7.5	7.5	6.0
50	6.5	4.5	7.0	6.5	8.0	6.5	7.5	9.0	6.5	8.0
60	6.0	6.5	7.0	6.5	6.0	5.0	7.5	7.0	8.0	5.5
70	6.5	6.0	5.5	7.0	6.5	7.5	6.5	6.0	7.0	7.5
80	6.5	5.5	7.5	6.0	8.0	7.0	6.0	6.5	7.5	8.0
90	5.5	6.5	6.0	6.5	7.5	7.5	6.0	6.5	8.0	5.5
100	6.5	8.5	5.5	7.0	5.0	7.5	7.5	7.5	5.0	7.0
110	4.5	7.0	7.0	7.0	6.5	8.0	7.5	8.0	7.0	6.0

3. Mark the mean as a dot on a number line from 5 to 9 hours.
4. Repeat. A group of students, each drawing a few samples, soon has a pile of dots.

Two questions to ask of the pile: where is it centred, and how wide is it? The hand-drawn samples are taken without replacement, so no student appears twice in one sample. The computer simulations below draw with replacement, which lets a student appear twice. With 10 drawn from 120, the difference is small. The formulas of this section assume the population is much larger than the sample, which is the usual situation in research.

3.4 A thousand samples

A computer can draw far more samples than a classroom. Figure 11 shows the means of 1,000 random samples of 10, and of 1,000 random samples of 30. Both piles centre on the true mean: the average of the 1,000 means is 6.81 for samples of 10 and 6.80 for samples of 30, against $\mu = 6.80$. The samples of 30 crowd closer to it. The SD of their means is 0.17, against 0.28 for samples of 10. Samples of 10 produced means as low as 5.95 and as high as 7.65.

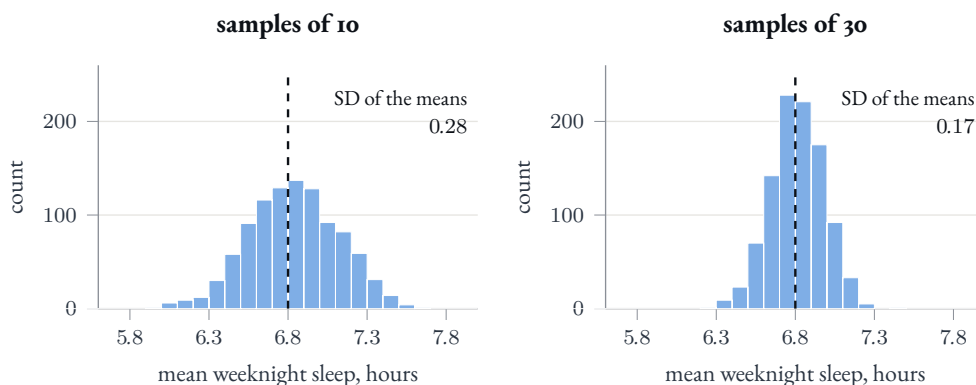


Figure 11: The means of 1,000 random samples of 10 (left) and of 30 (right) from the **simulated** class, drawn with replacement.

Definition 3.1 • Sampling distribution

The **sampling distribution** of a statistic is the distribution of its values over all possible random samples of the same size from the same population. The piles in Figure 11 are sampling distributions of the mean, drawn from 1,000 samples rather than all of them.

Three distributions are now in play, and keeping them apart is half the battle (Table 5).

Inference lives in the third row. A real study draws one sample and never sees the sampling distribution. Mathematics, however, says what its centre, spread and shape must be, and that is why one sample can be enough.

Table 5: Three distributions not to confuse.

	What it holds	Centre, spread	Is it seen?
The population	every member's value	μ, σ	almost never
One sample	the n values collected	$\bar{x}, s$	yes: it is the data
The sampling distribution	the means of all possible samples	μ, SE	never: it is imagined

3.5 The standard error

The spread of the sampling distribution has its own name. The standard error of the mean is the SD of the sample means, and a formula gives it without drawing a single sample:

$$\text{SE} = \frac{\sigma}{\sqrt{n}} \quad (3.1)$$

σ the population SD: how much individuals differ

n the sample size

SE how much the sample mean would differ from one sample to the next

Standard error

the SD of a statistic over repeated random samples; for a mean, $\sigma/\sqrt{n}$

Table 6 checks the formula against the 1,000 simulated samples of Figure 11. It agrees to within simulation noise. One line of algebra replaces a thousand samples.

Table 6: The standard error by equation (3.1), and the SD of 1,000 simulated sample means.

Simulated class data, $\sigma = 0.92$.

n	By the formula	From 1,000 samples
10	$0.92 \div \sqrt{10} = 0.29$	0.28
30	$0.92 \div \sqrt{30} = 0.17$	0.17

Example 3.1 • The answer to the opening question

Another class of 120 at the university fills in the survey. If the university's students vary as much as this class does, σ is about 0.92 hours, and

$$\text{SE} = \frac{0.92}{\sqrt{120}} = 0.084 \text{ hours.}$$

Sample means pile up in a roughly normal shape (section 3.8), so about 95 classes in 100 would average within about two standard errors of the true mean: $2 \times 0.084 \approx 0.17$ hours, which is about 10 minutes. Not half an hour, and not an hour: a class of 120 is a large sample, and its average is steady.

3.6 The square-root law

The $\sqrt{n}$ in equation (3.1) has a hard consequence. To halve the standard error, the sample must be four times as large; to cut it to a third, nine times. Figure 12 plots the SE of mean weeknight sleep against the sample size. It falls steeply at first and then flattens: each extra student buys less precision than the one before.

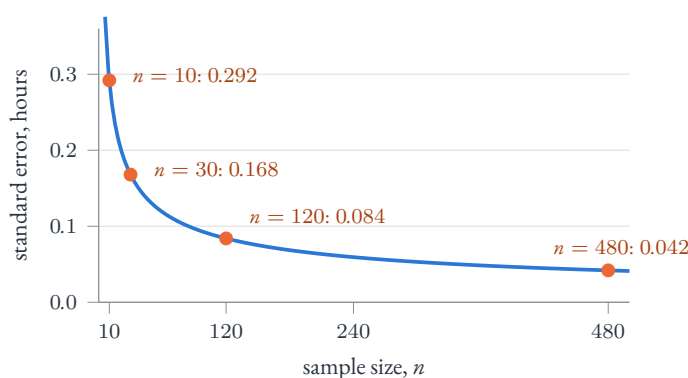


Figure 12: The standard error of the mean against the sample size, $\sigma = 0.92$ from the **simulated** class. The SE values are our arithmetic.

From 120 students to 480 is four times the forms to print, collect and enter, for half the error. Whether that is worth it depends on how precise the answer needs to be, and section 4.9 turns the question round to plan a sample size.

3.7 SD or SE?

The two measure different things, and papers mix them up. The SD, 0.93 hours for weeknight sleep, says how much *students* differ from each other. It does not shrink when more students are surveyed; a bigger sample only estimates it better. It is the number used to *describe*. The SE says how much the *mean* would differ between samples. Computed from the sample SD, it is $0.93 \div \sqrt{120} = 0.085$ hours, smaller by a factor of $\sqrt{120}$, and it keeps shrinking as n grows. It is the number used to *estimate*.

Mean $\pm$ what?

“Sleep 6.80 ± 0.085 hours” looks far more consistent than “ 6.80 ± 0.93 hours”, although both describe the same data. A table that reports an SE where a reader expects an SD makes the data look far tighter than they are. Every “ $\pm$ ” needs a label saying what follows it.

3.8 The central limit theorem

Weeknight sleep is roughly symmetric, so it is no surprise that its sample means pile up symmetrically. Screen time is skewed. Figure 13 draws 2,000 values of three kinds from the simulated screen-time column: single students, means of 5 students, and means of 30. The single values keep the long right tail, with a skewness of 0.69. Means of 5 are more symmetric, at 0.34, and means of 30 are close to a normal curve, at 0.13. Averaging cancels the extremes: an unusually high value in a sample is usually diluted by ordinary ones.

Definition 3.2 • The central limit theorem

For random samples of size n from a population with mean μ and SD σ , the sampling distribution of the mean is centred on μ , has standard error $\sigma/\sqrt{n}$, and comes closer to a normal curve as n grows, whatever the shape of the population.

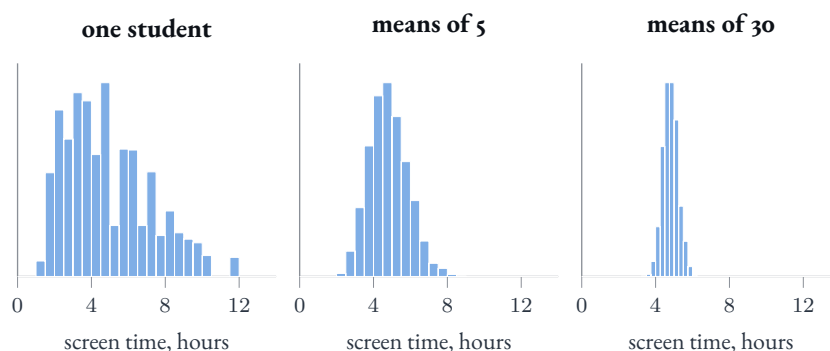


Figure 13: Screen time for single students, and means of 5 and of 30 students. 2,000 random draws of each kind from the **simulated** class.

The theorem is the reason the normal curve sits at the centre of statistics. It is also widely misread, and the limits matter as much as the promise.

What the theorem does not promise

1. It does not make the *data* normal. Screen time stays skewed however many students are surveyed; only the means of samples become normal.
2. It does not repair a biased sample. The means pile up around what the sampling method measures, which is the truth only if the sampling is fair.
3. It does not set a magic n . “Thirty is enough” is a rule of thumb. For data close to normal, far fewer will do; for very skewed data or data with extreme outliers, more are needed. The honest check is to look at the shape.

Central limit theorem

sample means from large enough random samples follow a normal curve centred on μ with spread $\sigma/\sqrt{n}$, whatever the data’s shape

3.9 How unusual is a sample mean?

Suppose a sample averages 6.40 hours, 0.40 hours below the true 6.80. Is that unusual? The answer depends on the size of the sample, because the SE does. A sample mean has a z -score like any other value, measured in standard errors rather than SDs:

$$z = \frac{\bar{x} - \mu}{SE} \quad (3.2)$$

For samples of 10, $z = (6.40 - 6.80) \div 0.29 = -1.38$. The normal area to the left is 8.39%: about one sample of 10 in 12 averages this low or lower. The 1,000 simulated samples agree, with 8.1% at or below 6.40. For samples of 120, $z = (6.40 - 6.80) \div 0.084 = -4.76$, which is about one sample in 1,000,000 (Figure 14).

The same gap is ordinary for a small sample and all but impossible for a large one. Turned into a question about a real study, this is the logic of every test in Chapter 4: if the truth were such-and-such, how unusual would a result like ours be?

3.10 Scatter, not bias

Chapter 1 drew the difference between sampling error and bias as four targets. The standard error measures the size of the cloud of shots. It says nothing about where the cloud is centred. Every calculation in this section assumed random samples, so the cloud was centred on the

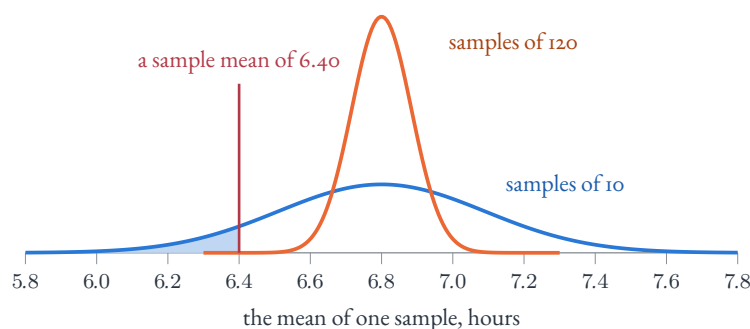


Figure 14: The sampling distributions of the mean for samples of 10 and of 120, and where a mean of 6.40 hours falls on each. **Simulated** class data; the areas are our arithmetic.

truth. A biased method moves the whole cloud, and the SE cannot detect it. A snowball survey of 5,000 people has a tiny standard error, and can still be centred on the wrong answer, a point section 5 makes with a real study.

4 Confidence intervals

4.1 Turning the picture round

Section 3 started from the truth and asked where samples land. Real research runs the other way: it has one sample and asks where the truth is. The link is a symmetry (Figure 15). In a normal sampling distribution, 95% of sample means land within 1.96 standard errors of μ . Whenever a sample mean is within 1.96 SE of μ , μ is within 1.96 SE of that sample mean. So for 95% of samples, the range “sample mean $\pm$ 1.96 SE” contains μ .

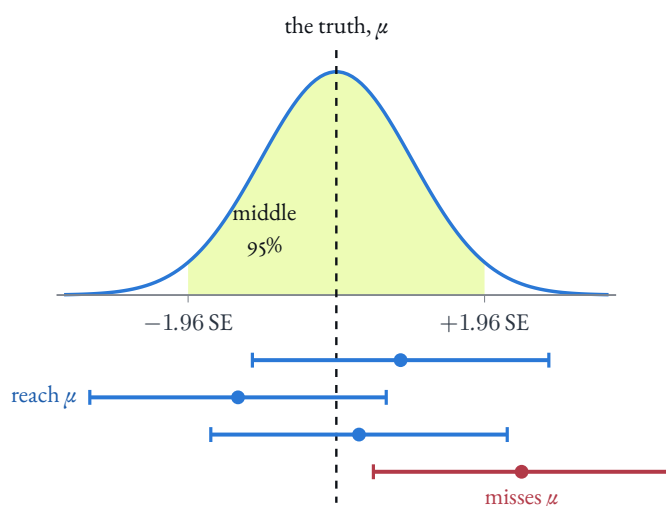


Figure 15: If a sample mean lands within 1.96 SE of the truth, the truth lies within 1.96 SE of the sample mean. Positions are illustrative (**constructed**), in SE units.

Definition 4.1 • Confidence interval

A **95% confidence interval** is a range computed from a sample by a method that, over repeated random samples, contains the parameter 95% of the time. For a mean or a proportion it takes the form

$$\text{estimate} \pm \text{multiplier} \times \text{SE},$$

and the product of the multiplier and the SE is the **margin of error**.

4.2 An interval for the mean, by hand

In practice σ is unknown, so the sample SD s stands in for it, and the multiplier becomes a value from the t distribution of section 4.3:

$$\bar{x} \pm t \times \frac{s}{\sqrt{n}} \quad (4.1)$$

$\bar{x}, s, n$ the sample mean, SD and size

t the 95% multiplier for $n - 1$ degrees of freedom (Figure 16)

$s/\sqrt{n}$ the estimated standard error

Example 4.1 • Mean weeknight sleep

The class, now treated again as a sample of the university, has mean 6.80 hours, SD 0.93 and $n = 120$.

Step	What	Working
1	standard error	$0.93 \div \sqrt{120} = 0.085$
2	multiplier	t for 119 df = 1.980
3	margin	$1.980 \times 0.085 = 0.17$
4	interval	6.80 ± 0.17 , from 6.63 to 6.97 hours

In words: the average weeknight sleep of students like these is estimated at 6.80 hours, and values from 6.63 to 6.97 are plausible, at 95% confidence. Chapter 1 showed this interval and promised this calculation.

4.3 The t distribution

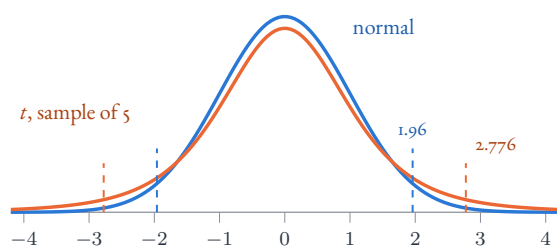
Why 1.980 and not 1.96? Replacing the unknown σ with s , which is itself estimated from the sample, adds uncertainty. The t distribution accounts for it. It looks like the standard normal curve with fatter tails (Figure 16), and its exact shape depends on the *degrees of freedom*, $n - 1$ for a single mean, the same $n - 1$ as in Chapter 2's SD. Fatter tails mean a larger multiplier and a wider interval.

For small samples the difference is large: with 5 observations the multiplier is 2.776, not 1.96. By 120 it has almost vanished. The distribution was published in 1908 by William Sealy Gosset, a chemist at the Guinness brewery in Dublin who worked with small samples and wrote under the pen name “Student” (reference 3). Software uses t automatically, and so should a hand calculation.

 t distribution

a bell-shaped curve with fatter tails than the normal, used when s stands in for σ ; its shape depends on the degrees of freedom





n	df	95% multiplier
5	4	2.776
10	9	2.262
30	29	2.045
120	119	1.980
very large		1.960

Figure 16: The t distribution with 4 degrees of freedom against the standard normal, with their 95% multipliers; and the multiplier for several sample sizes. Quantiles rounded to three decimals.

4.4 What “95% confident” means

The 95% is a property of the method, not of any single interval. Figure 17 makes the point by simulation. It draws 100 random samples of 30 from the class, treated once more as the population, and computes a 95% interval from each. 97 of the 100 intervals reach the true $\mu = 6.80$; 3 miss it. Over many more repeats the share reaching μ settles at 95%. Any one interval either contains μ or does not, and a real study, with one interval and no knowledge of μ , never learns which.

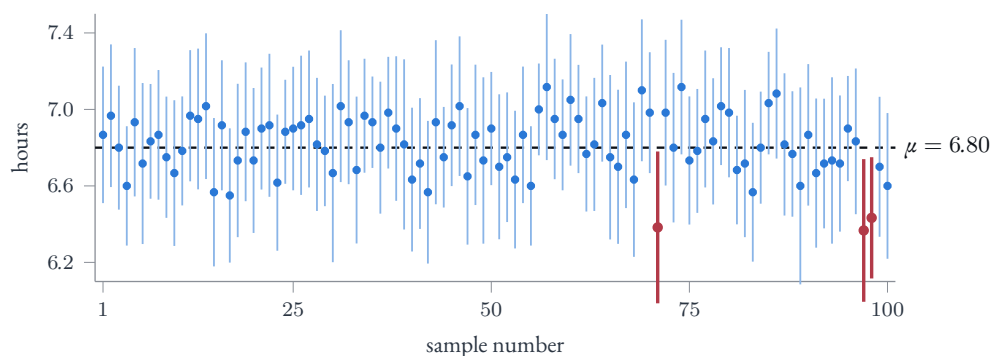


Figure 17: One hundred random samples of 30 from the **simulated** class, each with its 95% t interval. Intervals that miss the true mean are drawn in a second colour.

Table 7 sorts the common readings. The last row is the everyday way of speaking, and it does little harm in conversation. In writing, “we are 95% confident that the mean lies between” is the standard wording, and it means the first row.

4.5 An interval is not a range of people

The third reading in Table 7 is the most tempting, and the class data show how wrong it is. Not one of the 120 students reports a weeknight sleep between 6.63 and 6.97 hours: every answer was given to the half hour, and the nearest values are 6.5 and 7.0. The interval locates the *average*. Where most *students* are is a different question, answered by Chapter 2’s mean plus or minus two SDs: about 4.9 to 8.7 hours. With more students the confidence interval narrows towards a point, while the spread of people stays as wide as it is.

Table 7: Readings of the interval 6.63 to 6.97 hours.

Statement	Verdict
The method captures the true mean in 95% of samples.	Right.
The true mean is plausibly between 6.63 and 6.97 hours.	Right.
95% of students sleep between 6.63 and 6.97 hours.	Wrong. The interval is about the average, not about students.
A new sample's mean has a 95% chance of falling inside.	Wrong. The interval targets μ , not the next sample.
There is a 95% chance that μ is in <i>this</i> interval.	Loose. The 95% belongs to the method, not to one interval.

4.6 What sets the width

Equation (4.1) shows three things that set the margin: the multiplier, the spread s , and the sample size n . The spread belongs to the population and cannot be chosen. The other two can, and Figure 18 shows both.

More confidence costs width. With the same data and the same SE, a 90% interval uses the multiplier 1.658 and a 99% interval uses 2.618, against 1.980 for 95%. The margins are ± 0.14 , ± 0.17 and ± 0.22 hours. A 100% interval would run from 0 to 24 hours: certain, and useless. Ninety-five per cent is a convention, not a law; it is the one most journals expect.

More data buys narrowness. With the same mean and SD, a hypothetical class of 30 would give a margin of ± 0.35 hours, and one of 480 a margin of ± 0.08 . Each fourfold increase in n roughly halves the margin, the square-root law of section 3.6 again, with a slight extra gain because the t multiplier also shrinks.

4.7 An interval for a proportion

Chapter 2 showed that a proportion is the mean of a 0/1 column, so the same machinery applies. The SE of a sample proportion has its own formula, and the multiplier is 1.96:

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (4.2)$$

$\hat{p}$ the sample proportion, between 0 and 1
 $\sqrt{\hat{p}(1-\hat{p})/n}$ its standard error

Example 4.2 • The share giving tuition

In the class, 59 of 120 give tuition, so $\hat{p} = 0.4917$ and $1 - \hat{p} = 0.5083$.

$$SE = \sqrt{\frac{0.4917 \times 0.5083}{120}} = 0.0456, \quad \text{margin} = 1.96 \times 0.0456 = 0.0894.$$

The interval is 0.4917 ± 0.0894 , from 40.2% to 58.1%. Between two and three students in five, among students like these, give tuition, at 95% confidence.

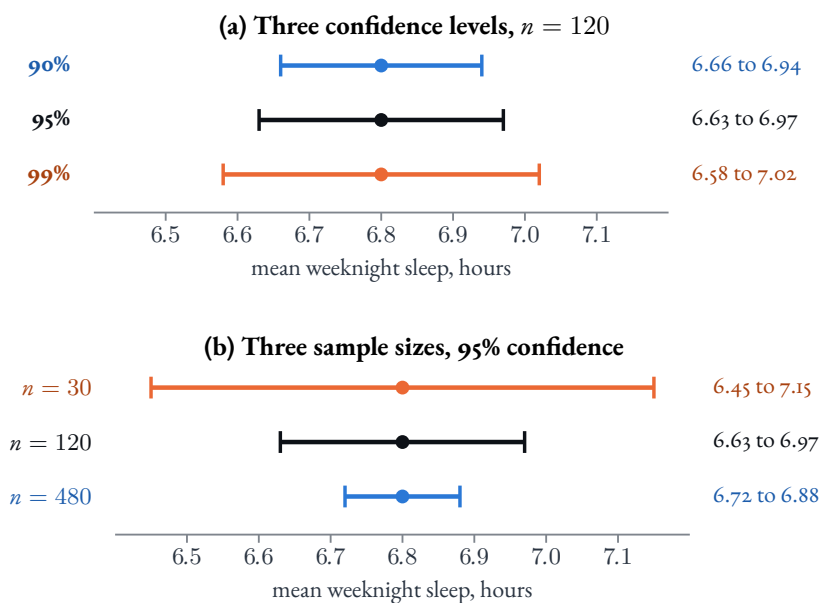


Figure 18: What sets the width of an interval. (a) The **simulated** class at three confidence levels. (b) Hypothetical classes of 30 and 480 with the same mean and SD as the **simulated** class. All intervals are our arithmetic.

Equation (4.2) is an approximation that works when both counts, the yeses and the noes, are at least about 10. Here they are 59 and 61. For small samples, or for shares close to 0 or 1, software offers better intervals, of which the Wilson interval is the most common. They give almost the same answer as equation (4.2) whenever the at-least-10 rule is met.

The quantity $\hat{p}(1 - \hat{p})$ controls the width, and it is largest at $\hat{p} = 0.5$ (Figure 19). A share near one half gives the widest interval a sample of a given size can give. The class's tuition share, 0.492, sits almost at the peak. That is why planners who do not yet know p assume 0.5: the margin can then only turn out smaller than planned.

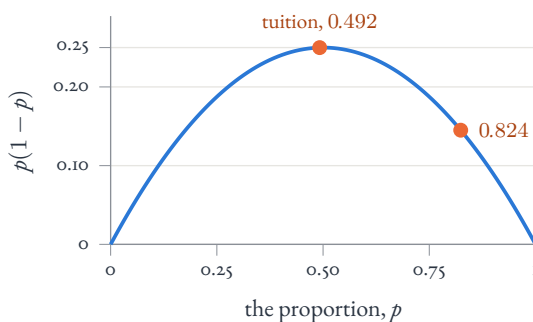


Figure 19: $p(1 - p)$ against p . The two marked points are the **simulated** class's tuition share and the share 392/476 from Islam 2020 (section 5).

4.8 The margin of error in a poll

News reports of opinion polls often end with “the margin of error is plus or minus three points”. The number comes from equation (4.2) with the worst case $p = 0.5$ and a typical

poll of 1,000 people:

$$1.96 \times \sqrt{\frac{0.5 \times 0.5}{1000}} = 0.031,$$

about ± 3.1 percentage points. A quick version is $1/\sqrt{n}$, which gives 3.2 points for 1,000.

The surprising part is what is missing. The size of the population appears nowhere in the formula. A random sample of 1,000 is as informative about Bangladesh as about Rajshahi, just as a spoonful of well-stirred soup tells the cook as much about a large pot as about a small one. The stirring is what matters: the sample must be random. The population size matters only when the sample is a large fraction of it, when a small adjustment called the finite population correction narrows the interval.

4.9 How many to sample

The margin formulas can be turned round to plan a study. Choose the margin E that the answer needs; the sample size follows.

$$n = \left(\frac{1.96 \times \sigma}{E} \right)^2 \text{ for a mean,} \quad n = \frac{1.96^2 \times 0.25}{E^2} \text{ for a proportion} \quad (4.3)$$

E the margin wanted, in the units of the answer

σ a planning guess at the SD, from a pilot study or a published one

0.25 $p(1-p)$ at its worst case, $p = 0.5$

Example 4.3 • Two survey plans

These plans are **constructed**; the planning SD comes from the **simulated** class.

Sleep to within ± 0.1 hours. With $\sigma \approx 0.93$, $n = (1.96 \times 0.93 \div 0.1)^2 = 332.3$. Round up: 333 students.

A percentage to within ± 5 points. $n = 1.96^2 \times 0.25 \div 0.05^2 = 384.2$, so 385. Within ± 3 points it is 1068. If only 70% of those invited are expected to reply, invite $385 \div 0.70 = 550$ to end with 385.

Two habits go with every plan. Always round *up*, because rounding down misses the target. And plan for non-response, because the formula counts answers, not invitations.

4.10 Reporting an estimate

Say which interval, give n , round sensibly

- “Mean weeknight sleep was 6.80 hours (95% CI 6.63 to 6.97; $n = 120$).”
- “49.2% gave tuition (95% CI 40.2 to 58.1%; $n = 120$).”

Not “sleep was 6.80 ± 0.17 ”, which leaves the reader to guess whether 0.17 is an SD, an SE or a margin. And not “sleep was 6.8000 hours”: four decimals claim a precision that answers given to the half hour cannot support.

4.11 An interval that answers a question

Adults are advised to sleep seven hours or more a night (Watson 2015, reference 2). Is the average for students like these below seven, or could the class's 6.80 be the luck of the draw? The interval already answers. Seven lies outside 6.63 to 6.97, so seven is not a plausible value for the average at the 95% level (Figure 20). The interval also says how far below: between about 2 and 22 minutes a night, a gap that is real but modest.

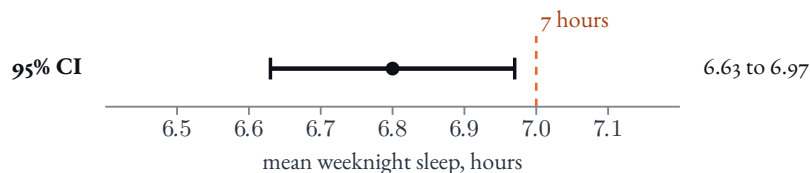


Figure 20: The 95% interval for mean weeknight sleep in the **simulated** class, against the recommended seven hours.

Chapter 4 asks the same question as a *test*, and gets a *p*-value. Interval and test will agree, because they are built from the same standard error. The interval carries more information: a test says only whether seven is plausible, while the interval shows every value that is.

5 Reading a published prevalence

5.1 Islam 2020: an interval the paper did not print

ISLAM 2020: A PREVALENCE WITHOUT AN INTERVAL

Islam and colleagues surveyed 476 university students in Bangladesh online in May 2020, during the first COVID-19 lockdown (reference 1). Chapter 1 met the study. By the PHQ-9 questionnaire, 392 students had at least mild depressive symptoms, which the paper reports as 82.4%. It prints no confidence interval for that figure.

Equation (4.2) supplies one. The working is ours, not the paper's.

$$\begin{aligned}\hat{p} &= 392 \div 476 = 0.8235, \\ \text{SE} &= \sqrt{\frac{0.8235 \times 0.1765}{476}} = 0.0175, \\ \text{margin} &= 1.96 \times 0.0175 = 0.0343.\end{aligned}$$

The interval is 0.8235 ± 0.0343 , from 78.9% to 85.8%. Its margin, ± 3.4 points, is much narrower than the tuition interval's ± 8.9 , for two reasons: 476 students is far more than 120, and a share of 0.82 lies far from the worst case of 0.5 (Figure 19).

5.2 What the interval leaves out

Equation (4.2) assumes a random sample, and the study did not have one. As Chapter 1 described, the questionnaire spread through Facebook and Messenger from friend to friend. The authors state the choice plainly: “the snowball sampling strategy was chosen instead



of random samples” (reference 1). The interval 78.9 to 85.8% measures the luck of which students happened to answer, and nothing else. It cannot see who the link never reached, who saw it and chose not to answer, or the gap between a screening score and a diagnosis. Table 8 lists the kinds of error a survey can contain.

Table 8: The errors a survey can contain, and which of them a confidence interval covers.

Error	Covered?	What it is
Sampling	Yes	The luck of which people were drawn.
Coverage	No	People the method could never reach: here, students offline or outside the authors’ networks.
Non-response	No	People reached who did not answer.
Measurement	No	Question wording: a screening score read as a diagnosis.
Processing	No	Mistakes in entering or printing: Chapter 2’s 25-hour typo, and the 87.7% in Islam 2020 that the paper’s own counts put at 81.7% (Chapter 1).

One sentence to keep

A margin of error is a floor on the uncertainty, not a ceiling. The true uncertainty in the 82.4% is larger than ± 3.4 points, by an amount no formula can give.

6 Estimating a mean or a proportion: the procedure

1. **Ask how the sample was drawn.** Every formula in this chapter assumes a random sample. If the sample was not random, say so, and treat the interval as a floor on the uncertainty.
2. **Look at the data.** A histogram for a numerical variable; the counts of yes and no for a proportion. Check for impossible values, as in Chapter 2.
3. **Compute the estimate,** $\bar{x}$ or $\hat{p}$.
4. **Compute the standard error:** $s/\sqrt{n}$ for a mean, $\sqrt{\hat{p}(1-\hat{p})/n}$ for a proportion.
5. **Choose the multiplier:** t with $n-1$ df for a mean (Figure 16); 1.96 for a proportion when both counts are at least about 10.
6. **Compute the margin and the interval,** and round both to what the data can bear.
7. **Report** the estimate, the interval, its level and n , in one sentence.

7 Chapter summary

1. A probability is a long-run share, between 0 and 1. From a table it is a count over a count: the cases wanted over the cases that could occur.
2. The complement rule, $P(\text{not } A) = 1 - P(A)$, and the addition rule, which subtracts the overlap, hold always. Probabilities multiply only for independent events.
3. A conditional probability restricts the denominator. $P(A \mid B)$ and $P(B \mid A)$ share a numerator and can differ greatly.
4. Chance has no memory: an early run is diluted, not corrected. With many comparisons, some coincidence is close to certain.
5. The normal curve is fixed by its mean and SD, and holds 68.3%, 95.4% and 99.7% of its area within one, two and three SDs.
6. A z-score counts SDs from the mean. It has no units, so it compares values on different scales, and a table or software turns it into an area.
7. The normal curve is a model. Check it against the data; for skewed data its areas are wrong in both tails.
8. Sample means vary. Their SD is the standard error, $\sigma/\sqrt{n}$, which falls with the square root of the sample size.
9. The central limit theorem makes sample means close to normal for large samples. It does not make the data normal or fix a biased sample.
10. A 95% confidence interval is the estimate plus or minus a multiplier times the SE. The 95% belongs to the method, and the interval locates an average, not people.
11. The interval covers sampling error only. Coverage, non-response, measurement and processing errors lie outside it.

8 Key terms

Addition rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$, equation (1.3).

Central limit theorem

Sample means from large enough random samples are close to normal, centred on μ , with spread $\sigma/\sqrt{n}$.

Complement

The event that A does not happen; $P(\text{not } A) = 1 - P(A)$.

Conditional probability

The probability of A among the cases where B is true, equation (1.4).

Confidence interval

A range from a method that captures the parameter in a stated share of repeated samples, equations (4.1) and (4.2).

Degrees of freedom

For a single mean, $n - 1$; it sets the shape of the t distribution.

Expected count

What a cell of a table would hold if the two variables were independent: row total times column total over grand total.

Gambler's fallacy

The belief that a run of one outcome makes the other “due”.

Independent events

Events for which $P(A | B) = P(A)$; only for them does $P(A \text{ and } B) = P(A) \times P(B)$.

Margin of error

The multiplier times the SE: half the width of the interval.

Mutually exclusive

Unable to happen together; the overlap is zero.

Normal distribution

A symmetric bell-shaped curve, fixed by its mean and SD.

Probability

The share of times an event happens in a long run of repeats.

Sampling distribution

The distribution of a statistic over all possible random samples of one size.

Sampling variability

The variation of a statistic from one sample to another.

Standard error

The SD of a statistic over repeated samples; for a mean, $\sigma/\sqrt{n}$, estimated by $s/\sqrt{n}$.

Standard normal

The normal curve with mean 0 and SD 1.

t distribution

A bell curve with fatter tails than the normal, used when s replaces σ .

z -score

A distance from the mean counted in SDs, equation (2.1).

9 Exercises

Exercises in Part A need no calculation. Parts B to E need a calculator, and a few need Table 3 or the software functions of section 2.6. Numbers not attributed to a paper or to the class data are constructed for teaching.

Part A. Ideas.

1. A forecast gives a 70% chance of rain tomorrow, and it stays dry. (a) Was the forecast wrong? (b) Over a year of forecasts, what evidence would show that this forecaster is good at their job?
2. An invigilator's report says that 90% of the students caught cheating in an exam hall had a mobile phone with them. The invigilator concludes that a student carrying a phone is very likely to be cheating. (a) Write the report's figure and the invigilator's conclusion as conditional probabilities. (b) What is wrong with the conclusion? (c) What further information would be needed to judge the risk?
3. A paper describes 400 students as aged " 21.4 ± 0.1 years", without saying what follows the $\pm$. (a) Is 0.1 more likely to be an SD or an SE? Explain. (b) If it is an SE, what is the SD?
4. For the interval 6.63 to 6.97 hours of section 4.2, say whether each statement is right or wrong, and why. (a) Most students sleep between 6.63 and 6.97 hours. (b) If the survey were repeated many times, about 95% of intervals made this way would contain the true mean. (c) A second class's mean would fall inside this interval 95% of the time. (d) An average of 7 hours is not a plausible value for students like these.
5. A student writes: "The class sample is bigger than 30, so by the central limit theorem screen time is normally distributed, and the normal curve gives the share who use their phone less than one hour a day." Correct the reasoning.
6. A news website runs an online poll that 5,000 readers choose to answer, and reports the result with a margin of error of ± 1.4 points. (a) Where does 1.4 come from? (b) Is the result accurate to within 1.4 points? Explain.

Part B. Calculation.

7. From the residence-by-tuition table (Figure 2), find $P(\text{mess or tuition})$ (a) by counting the students and (b) by the addition rule.
8. (a) Find $P(\text{hall} \mid \text{tuition})$, $P(\text{mess} \mid \text{tuition})$ and $P(\text{home} \mid \text{tuition})$. (b) What should the three add to, and why? Do they? (c) Find $P(\text{home})$ and $P(\text{not home})$.
9. (a) If residence and tuition were independent, what would $P(\text{mess and tuition})$ be? (b) How many mess residents giving tuition would that mean among 120 students, and how many does the table show? (c) Repeat (b) for family home.
10. Five friends sit together. Assuming 365 equally likely birthdays, find the probability that at least two share a birthday.

- 11.** Use the class mean and SD of weeknight sleep. (a) Find the z -scores for 7.5, 6 and 8 hours, and say in words what the first one means. (b) Find the areas to the left of 6 and of 8 hours. (c) What share of students does the normal curve put between 6 and 8 hours? (d) The class has 97 of 120 students between 6 and 8 hours inclusive. Compare.
- 12.** Use the class mean and SD of the stress score. (a) Find the z -score of a score of 10. (b) What share of students does the normal curve put at or below 10? (c) In the class, 21 of 118 students score 10 or less. Compare.
- 13.** Weekend sleep in the class has mean 7.87 hours, SD 1.05 and $n = 119$. The t multiplier for 118 df is 1.980. (a) Find the 95% confidence interval for mean weekend sleep. (b) Is 8 hours a plausible value for the average?
- 14.** 43 of the 120 students live in a mess. Find a 95% confidence interval for the share of students like these who live in a mess, and check the at-least-10 rule.
- 15.** (a) How many students would be needed to estimate mean weeknight sleep to within ± 0.25 hours, taking $\sigma \approx 0.93$? (b) How many to estimate a percentage to within ± 4 points? (c) If 60% of those invited reply, how many should be invited for (b)?
- 16.** (a) Using Figure 12's $\sigma = 0.92$, find the SE for samples of 10 and of 40, and the ratio of the second to the first. (b) How large a sample would give one third of the SE of a sample of 30?

Part C. Interpretation.

- 17.** The weekend interval of Exercise 13 and the weeknight interval of section 4.2 do not overlap. (a) What does that suggest? (b) The same students answered both questions. Why does that matter for the right way to compare them?
- 18.** Explain, without calculating, why the interval for Islam 2020's 82.4% is narrower than the interval for the class's tuition share. Then explain why it may still understate the real uncertainty.
- 19.** A draft report says "Stress: 15.04 ± 4.57 ". Rewrite it so that a reader knows exactly what is reported.
- 20.** Suppose another class had the same SD and size as this one but averaged 6.95 hours. Using the same margin as section 4.2, find its interval, and say what it implies about the seven-hour recommendation.

Part D. The class data.

- 21.** A group drew ids 7, 19, 23, 38, 44, 58, 71, 86, 99, 112 from Table 4. (a) Look up the ten values and find their mean. (b) Find the SD by Chapter 2's table method. (c) Find the SE, and a 95% confidence interval, using $t = 2.262$ for 9 df. (d) Does the interval contain $\mu = 6.80$? (e) Find the z -score of this sample mean, taking μ and σ as known.

22. Draw a sample of your own: ten ids at random from Table 4. Find its mean, SD and 95% confidence interval as in Exercise 21. Does your interval contain $\mu = 6.80$? About how wide should you expect the margin to be?

Part E. Challenge.

23. In a room of 23 people, find the chance that at least one of the other 22 shares *your* birthday. Compare it with Table 2, and explain the difference.

24. A **constructed** sample of four students sleeps a mean of 6.5 hours with SD 1.0. The t multiplier for 3 df is 3.182. (a) Find the 95% interval. (b) Find the interval that 1.96 would have given. (c) Why does the difference matter for small samples?

25. A **constructed** poll of 1,000 voters finds candidate A at 52% and candidate B at 48%. (a) What is the usual margin for each share? (b) For the *lead* of one candidate over the other, when both shares come from the same poll, the margin is $1.96 \times \sqrt{(p_A + p_B - (p_A - p_B)^2)/n}$. Compute it. (c) Is the lead clear?

26. Compute a 99% interval for Islam 2020's 392 of 476, using the multiplier 2.576 and the SE of section 5.1.

27. The other 84 of the 476 students in Islam 2020 had none or minimal depressive symptoms. Find their percentage and a 95% interval. Why is the SE the same as in section 5.1?

10 Solutions

Part A.

1. (a) Not necessarily. A 70% chance means that on days like this it rains about 7 times in 10, so it stays dry about 3 times in 10. One dry day cannot show the forecast was wrong. (b) Collect all the days given 70%, and check that it rained on about 70% of them; do the same for 10%, 50% and so on. A forecaster whose stated chances match the long-run shares is well calibrated. (Exercise 1.)

2. (a) The report gives $P(\text{phone} \mid \text{cheating}) = 0.90$; the conclusion claims $P(\text{cheating} \mid \text{phone})$ is high. (b) It is the swapped conditional of section 1.5. If nearly every student carries a phone, nearly every cheater will too, and so will nearly every honest student. (c) The share of all students who carry phones, and the number who cheat. If phones are almost universal, $P(\text{cheating} \mid \text{phone})$ is close to the overall cheating rate. (Exercise 2.)

3. (a) An SE. Students' ages spread over several years, so an SD of 0.1 years, about five weeks, is implausible for 400 students. (b) $SD = SE \times \sqrt{n} = 0.1 \times \sqrt{400} = 0.1 \times 20 = 2.0$ years, which is plausible. (Exercise 3.)

4. (a) Wrong: the interval is about the average, and in fact 0 students report a value inside it (section 4.5). (b) Right: this is what 95% confidence means. (c) Wrong: the interval targets μ , not another sample's mean, which has its own sampling error. (d) Right: 7 lies outside the interval (section 4.11). (Exercise 4.)

5. The theorem is about the *means* of samples, not the data. Screen time stays skewed whatever the sample size, and section 2.8 shows the normal curve misplacing its lower tail badly. The share under one hour should be counted from the data. (Exercise 5.)

6. (a) $1.96 \times \sqrt{0.25/5000} = 0.014$, the worst-case margin for a random sample of 5,000. (b) No. The readers chose themselves, so the sample is not random, and the margin covers sampling error only. Those who choose to answer an online poll can differ from everyone else in exactly the way the poll measures. (Exercise 6.)

Part B.

7. (a) The mess row and the tuition column together: $11 + 32 + 17 + 10 = 70$ students, so $P = 70 \div 120 = 0.583$. (b) $P(\text{mess}) + P(\text{tuition}) - P(\text{mess and tuition}) = 0.358 + 0.492 - 0.267 = 0.583$. (Exercise 7.)

8. (a) The bag is the tuition column of 59: $17 \div 59 = 0.288$, $32 \div 59 = 0.542$ and $10 \div 59 = 0.169$. (b) To 1, since every tuition-giver lives in exactly one of the three places. They add to 0.999 because each was rounded. (c) $P(\text{home}) = 30 \div 120 = 0.250$, and $P(\text{not home}) = 1 - 0.250 = 0.750$. (Exercise 8.)

9. (a) $0.358 \times 0.492 = 0.176$. (b) $43 \times 59 \div 120 = 21.1$ expected; the table shows 32. (c) $30 \times 59 \div 120 = 14.8$ expected; the table shows 10. Mess residents give tuition more often than independence predicts, and home residents less often. (Exercise 9.)

10. $P(\text{no match}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \frac{361}{365} = 0.973$, so $P(\text{at least one match}) = 1 - 0.973 = 0.027$, about 2.7%. (Exercise 10.)

11. (a) $z = (7.5 - 6.80) \div 0.93 = 0.75$: 7.5 hours is three quarters of an SD above the mean. For 6 hours $z = -0.86$, and for 8 hours $z = 1.29$. (b) 0.195 and 0.901. (c) $0.901 - 0.195 = 0.706$. (d) The class has $97 \div 120 = 80.8\%$, more than the curve's share. Answers given to the half hour pile onto 6.0 and 8.0 exactly, and "inclusive" counts them; the curve spreads the same students smoothly. (Exercise 11.)

12. (a) $z = (10 - 15.04) \div 4.57 = -1.10$. (b) The area to the left is 0.136. (c) The class has $21 \div 118 = 17.8\%$, more than the curve predicts. A total of whole-number items takes only whole-number values, and "at or below 10" includes everyone who scored exactly 10. (Exercise 12.)

13. (a) $SE = 1.05 \div \sqrt{119} = 0.096$; margin $= 1.980 \times 0.096 = 0.19$; interval 7.87 ± 0.19 , from 7.68 to 8.06 hours. (b) Yes: 8 lies inside the interval. (Exercise 13.)

14. $\hat{p} = 43 \div 120 = 0.3583$ and $1 - \hat{p} = 0.6417$. $SE = \sqrt{0.3583 \times 0.6417 \div 120} = 0.0438$; margin $= 1.96 \times 0.0438 = 0.0858$. The interval is 27.3% to 44.4%. Both counts, 43 in a mess and 120 minus that elsewhere, are well above 10. (Exercise 14.)

15. (a) $(1.96 \times 0.93 \div 0.25)^2 = 53.2$; round up to 54. (b) $1.96^2 \times 0.25 \div 0.04^2 = 0.9604 \div 0.0016 = 600.25$; round up to 601. (c) $601 \div 0.60$, rounded up: 1002. (Exercise 15.)

16. (a) $0.92 \div \sqrt{10} = 0.292$ and $0.92 \div \sqrt{40} = 0.146$, a ratio of 0.50: four times the sample, half the SE. (b) One third of the SE needs nine times the sample: $9 \times 30 = 270$ students. (Exercise 16.)

Part C.

17. (a) That average sleep is longer at weekends: the lowest plausible weekend average, 7.68, is above the highest plausible weeknight average, 6.97. (b) The two intervals treat the answers as two separate groups of students. They are one group answering twice, so the natural analysis looks at each student's own difference, weekend minus weeknight. That is the paired comparison of Chapter 4, and it is more precise, because it removes the differences between students. Note also that the reverse rule does not work: overlapping intervals do not show that there is no difference. (Exercise 17.)

18. The SE is $\sqrt{p(1-p)/n}$. Islam 2020 has a larger n , 476 against 120, and a share further from 0.5, where $p(1-p)$ is smaller (Figure 19). Both shrink the SE. But the interval assumes a random sample, and the study used a snowball sample through Facebook and Messenger. Coverage and non-response errors are outside the interval (Table 8). (Exercise 18.)

19. "Stress score (range 5 to 25): mean 15.04 (SD 4.57), $n = 118$." The label says the $\pm$ is an SD, and n tells the reader that some students did not answer every item. (Exercise 19.)

20. 6.95 ± 0.17 gives 6.78 to 7.12 hours. Seven lies inside, so such a class could not rule out an average of seven hours: the data would be consistent with the recommendation being met. (Exercise 20.)

Part D.

21. (a) The values are 8.0, 5.5, 5.0, 6.0, 7.5, 9.0, 6.5, 7.0, 8.0, 7.0. Their sum is 69.5, and the mean is $69.5 \div 10 = 6.95$ hours.

(b)

id	x	$x - \bar{x}$	$(x - \bar{x})^2$
1st	8.0	+1.05	1.1025
2nd	5.5	-1.45	2.1025
3rd	5.0	-1.95	3.8025
4th	6.0	-0.95	0.9025
5th	7.5	+0.55	0.3025
6th	9.0	+2.05	4.2025
7th	6.5	-0.45	0.2025
8th	7.0	+0.05	0.0025
9th	8.0	+1.05	1.1025
10th	7.0	+0.05	0.0025
Sum	69.5	0	13.725

The variance is $13.725 \div 9$, and the SD is its square root, $s = 1.23$ hours.

(c) $SE = 1.23 \div \sqrt{10} = 0.389$; margin = $2.262 \times 0.389 = 0.88$; interval 6.95 ± 0.88 ,



from 6.07 to 7.83 hours. (d) Yes: 6.80 lies inside. The sample mean missed μ , as nearly every sample mean does, but the interval reached it. (e) $z = (6.95 - 6.80) \div 0.29 = 0.52$: an ordinary sample. (Exercise 21.)

22. Answers differ from sample to sample, which is the point. With an SD near 0.93, the margin should be around $2.262 \times 0.93 \div \sqrt{10} \approx 0.67$ hours, larger or smaller as the sample's own SD comes out larger or smaller. About 95 samples in 100 give an interval that contains 6.80; a group whose interval misses has not made a mistake. (Exercise 22.)

Part E.

23. $1 - (364/365)^{22} = 5.9\%$, against 50.7% for any match among 23 people. A match with one particular person involves only 22 pairs; a match between anyone involves 253. (Exercise 23.)

24. (a) $SE = 1.0 \div \sqrt{4} = 0.5$; margin $= 3.182 \times 0.5 = 1.59$; interval 4.91 to 8.09 hours. (b) $1.96 \times 0.5 = 0.98$; interval 5.52 to 7.48 hours. (c) With four observations, s is a rough estimate of σ . Using 1.96 ignores that, and the interval would capture the truth far less often than 95% of the time. (Exercise 24.)

25. (a) About ± 3.1 points each, as in section 4.8. (b) The lead is $0.52 - 0.48 = 0.04$, and

$$1.96 \times \sqrt{\frac{0.52 + 0.48 - 0.04^2}{1000}} = 1.96 \times \sqrt{\frac{0.9984}{1000}},$$

which is ± 6.2 points. (c) No. The lead is 4 points and its margin is ± 6.2 , so a tie, or a lead for B, is plausible. The margin on a lead is roughly twice the margin on each share, a point news reports often miss. (Exercise 25.)

26. Margin $= 2.576 \times 0.0175 = 0.0451$; interval 0.8235 ± 0.0451 , from 77.8% to 86.9%. It is wider than the 95% interval, 78.9 to 85.8%, as more confidence always costs width. (Exercise 26.)

27. $\hat{p} = 84 \div 476 = 0.1765$, or 17.6%. The SE uses $\hat{p}(1 - \hat{p}) = 0.1765 \times 0.8235$, the same product as before with the factors swapped, so the SE is again 0.0175 and the margin 0.0343. The interval is 14.2% to 21.1%: the mirror image of the first, as it must be, since the two shares add to 1. (Exercise 27.)

II Our arithmetic and constructed numbers

Every number below was computed for these notes. None is printed by the papers cited, although the inputs from real papers are.

Value	How	Inputs from
0.8235; SE 0.0175; 78.9 to 85.8%	95% interval for 392 of 476, equation (4.2)	Islam 2020
77.8 to 86.9%	99% interval, Exercise 26	Islam 2020
17.6%; 14.2 to 21.1%	the remaining 476 – 392 students, Exercise 27	Islam 2020
2 to 22 minutes	how far the interval lies below seven hours	Watson 2015; the simulated class
Normal areas, t multipliers	standard tables, rounded to three decimals	none

Simulated, not real students: every figure from the class data, including the cross-tabulation, the sleep grid and every repeated sample drawn from it: the five samples, the 1,000 samples of 10 and of 30, the 2,000 screen-time draws and the 100 intervals. The 400 coin tosses are simulated too. All are generated by `data/class-data.R` from a fixed seed, and the class figures will be replaced by the real survey.

Constructed for teaching, and not from any study: the true/false quiz; the viva of Example 1.1; the invigilator, the forecast, the ages and the online poll of Exercises 1 to 6; the survey plans of Example 4.3; the positions in Figure 15; the classes of 30 and 480 in Figure 18; the class averaging 6.95 in Exercise 20; and the samples and poll of Exercises 24 and 25.

12 References and further reading

- Islam MA, Barna SD, Raihan H, Khan MNA, Hossain MT. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: a web-based cross-sectional survey. *PLoS One* 2020;15(8):e0238162. PMID 32845928. PMC7449469. doi:10.1371/journal.pone.0238162. Licensed CC BY 4.0; free to read and share.
- Watson NF, Badr MS, Belenky G, et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep* 2015;38(6):843–844. PMID 26039963. PMC4434546. doi:10.5665/sleep.4716. The PubMed abstract does not state the seven-hour figure, and the full text was not available to us; the figure is as summarised by the US Centers for Disease Control and Prevention.
- Student [Gosset WS]. The probable error of a mean. *Biometrika* 1908;6(1):1–25. doi:10.1093/biomet/6.1.1. The paper that introduced the t distribution. Subscription journal; it has no PMID.
- Class data: a **simulated** stand-in for the Stat.BD class survey, generated by `data/class-data.R` from the fixed seed 20261005. Not real students. No DOI or PMID.



13 For students in the Stat.BD course

Before the next session. Read this chapter, and work the Part A exercises (Exercises 1 to 6). Exercise 22 repeats the class activity with a sample of your own, and is worth doing even if the activity was finished in class. The rest are for practice at any point in the course. Bring a calculator.

What the next session does with this chapter. Session 4 turns intervals into tests. In the class, students who give tuition score 15.78 on the stress scale on average, and those who do not score 14.33, a gap of 1.44 points. Could a gap like that come from the luck of which students were surveyed? Session 4 answers with the standard error of this chapter, applied to a difference between two means, and with the logic of section 3.9.

Questions this chapter left open, and where each is settled.

Question	Where it is settled
Is average sleep below seven hours, asked as a test?	Chapter 4, the one-sample t -test
Is the tuition gap in stress more than chance?	Chapter 4, the two-sample t -test
How to compare weekend and weeknight sleep for the same students?	Chapter 4, the paired t -test
Are residence and tuition independent? Is 17 far enough from 23.1?	Chapter 5, chi-square
What to do with skewed data such as screen time?	Chapter 4, rank-based tests
Why many comparisons produce false alarms	Chapter 4, multiple testing
Intervals for a correlation or a regression slope	Chapters 5 and 6
Producing these intervals in software	Session 7 onward