

From Sample to Population

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

Muhtasim Munif Fahim, MSc

Stat.BD

Data Science Research Lab, Department of Statistics & Data Science,
University of Rajshahi

Session 3 of 14 • Theory

Last session described 120 students. This one reaches past them

Session 2 gave us

- ▶ tables for every category
- ▶ a centre: mean, median, mode
- ▶ a spread: SD, IQR, and the values that stand out
- ▶ charts, and how charts mislead

Today, in five acts

- | | |
|-----|-------------------------------------|
| I | Probability, the language of chance |
| II | The normal curve, and z-scores |
| III | Samples vary, and by how much |
| IV | The confidence interval |
| V | An interval for a published figure |

Every one of those numbers was about the 120 people who answered.

The rule for the whole session

An estimate without a margin of error is only half an answer.

Class data throughout: **simulated** stand-in for the class survey (data/class-data.R, fixed seed). Not real students.



What you will be able to do by the end of today

The test of this session

Given a sample mean or a percentage, compute a 95% confidence interval, and say in one plain sentence what it tells you and what it does not.

- ▶ Explain why different samples from the same population give different results
- ▶ Use z-scores to compare values measured on different scales
- ▶ Describe the central limit theorem in plain words
- ▶ Interpret a 95% confidence interval correctly, and name two ways to misread one



6.80

hours of weeknight sleep: the average of our 120

Another class of 120 at this university fills in the same survey.

Simulated class data, $n = 120$.

Write down your guesses

1. Would they average exactly 6.80 hours?
2. If not, how far off could they plausibly be: 0.1 hours? 0.5? 1?
3. What would make the gap bigger or smaller?

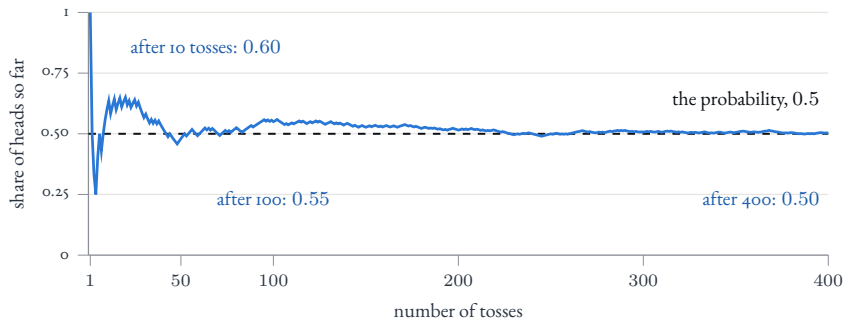


Act I

Probability, the language of chance



A probability is what happens in the long run



The first ten tosses: H T T T H H T H H H. Six heads is not “wrong” for a fair coin; ten tosses is simply too few to see one half.

Probability: the share of times an event happens in a very long run of repeats. Always between 0 and 1.

Simulated: 400 tosses of a fair coin in R (data/class-data.R, fixed seed).



Our table as a bag of names: pick one student at random

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

Filled cell: the students we want. Boxed cell: all the students we could draw.

Simulated class data: residence by tuition, the cross-tab from Session 2.

$$P(\text{hall}) = \frac{47}{120} = 0.392$$

Complement: the chance it does *not* happen.

$$P(\text{not hall}) = 1 - 0.392 = 0.608$$

All the options add to 1:

$$\frac{47 + 43 + 30}{120} = 1$$



“And” and “or”: count each student once

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

Filled: in a hall, *or* giving tuition, or both.

Simulated class data. Probabilities are our arithmetic from the table.

And: both at once, one cell.

$$P(\text{hall and tuition}) = \frac{17}{120} = 0.142$$

Or: either one, or both.

$$P(\text{hall or tuition}) = \frac{89}{120} = 0.742$$

The addition rule gives the same answer:

$$0.392 + 0.492 - 0.142 = 0.742$$



“Given that” changes the denominator

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

Given “mess”, only the boxed row can be drawn.

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

Read the bar as “given”.

Among	share giving tuition
everyone	0.492
hall	0.362
mess	0.744
family home	0.333

Simulated class data. Conditional probabilities are our arithmetic from the table.



$P(A \mid B)$ is not $P(B \mid A)$

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

$$P(\text{mess} \mid \text{tuition}) = \frac{32}{59} = 0.542$$

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

Same numerator. Different bag.

Given “tuition”, the boxed column is the bag.

The swapped conditional

“Most students who failed the viva were nervous” does not mean “most nervous students fail the viva”. Nearly everyone is nervous before a viva.

Simulated class data. The viva example is **constructed**.



Independent: knowing one tells you nothing about the other

If residence and tuition were independent

$$\begin{aligned}P(\text{hall and tuition}) &= P(\text{hall}) \times P(\text{tuition}) \\&= 0.392 \times 0.492 = 0.193\end{aligned}$$

That is about 23.1 of our 120 students.

What the table shows

- ▶ 17 students in a hall give tuition, not about 23.1
- ▶ $P(\text{tuition} \mid \text{mess}) = 0.744$,
not 0.492

Knowing where a student lives changes the odds that they give tuition.

Hold on to the 23.1

“What the table would show if the two were unrelated” is the expected count. It is the engine of Session 5’s chi-square test.

Simulated class data. Expected count $47 \times 59 \div 120$ is our arithmetic.



The gambler's fallacy

A student has guessed “false” on five true/false questions in a row. The sixth is not “due” to be “true”. Each question is new.

Then why does the coin's line settle at one half?

Not because tails catch up. An early run of extra heads is simply *diluted*. Two heads more than expected in 10 tosses gives 7 of 10, a share of 0.70.

The same two extra in 400 tosses gives 202 of 400, a share of 0.505.

The long run is patient. It does not correct the past; it outnumbers it.

The true/false example is **constructed**. Shares are our arithmetic.



Why coincidences are common: the birthday problem

In a room of 23 people, what is the chance that two share a birthday?

Easier to count the opposite: *nobody* shares.

$$\frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{343}{365}$$

Then use the complement rule.

People in the room	Chance of a shared birthday
10	11.7%
23	50.7%
30	70.6%
50	97.0%

Why it surprises

We imagine someone matching *our* birthday. But 23 people make 253 pairs, and any one pair can match.

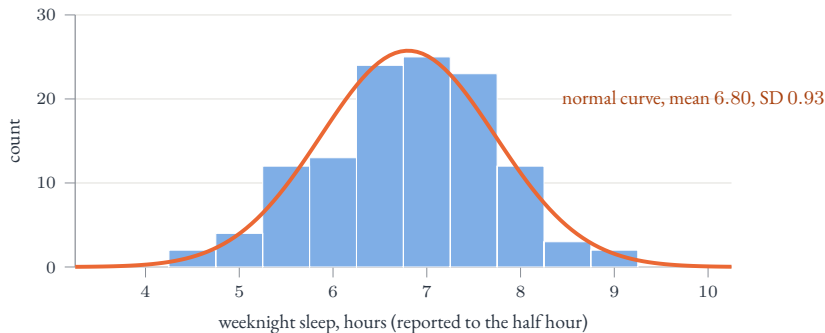
Our arithmetic, assuming 365 equally likely birthdays and ignoring 29 February and twins.



Act II

The normal curve, and z-scores

Our sleep data, with a smooth curve laid over them



The bars are our 120 students. The curve is a **normal distribution** with the same mean and SD. It is not a perfect fit, and it does not need to be: it is a model, a smooth description of a lumpy reality.

Simulated class data. Curve scaled to counts: $n \times \text{bar width} \times \text{normal density}$.



The normal curve, and the 68–95–99.7 rule

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

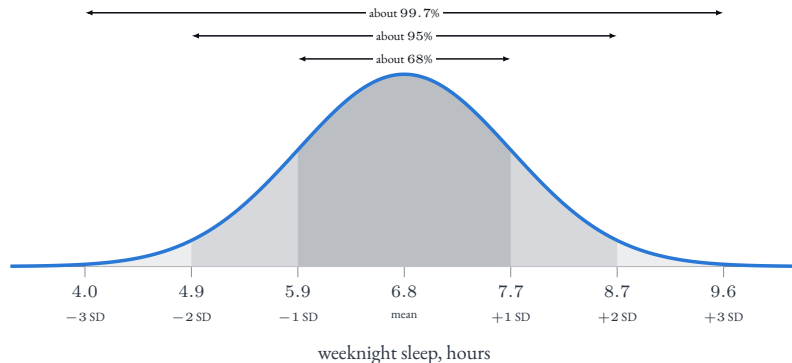
Samples vary

Confidence intervals

A published figure

Close

References



Symmetric, one peak, mean = median = mode. **Two numbers fix it:** the mean and the SD.

Our class: 70.8% within 1 SD, 96.7% within 2 SD.
Session 2's rule of thumb, now with its name.

Simulated class data. Band endpoints computed from the rounded mean and SD.



A z-score counts standard deviations from the mean

$$z = \frac{x - \text{mean}}{\text{SD}}$$

The sign says which side of the mean. The size says how far, in SDs.

A ruler with SDs for inches

$z = 0$ is exactly average. $z = -2$ is two SDs below it.

A student who sleeps 4.5 hours

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

Outside the 95% band: unusual.

A student who sleeps 8.5 hours

$$z = \frac{8.5 - 6.80}{0.93} = +1.83$$

Inside the band, near its edge.

Simulated class data. z-scores from the rounded mean and SD, our arithmetic.



A z-score has no units

In hours

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

In minutes

$$z = \frac{270 - 408}{55.8} = -2.47$$

Hours divided by hours: the unit cancels. So does any change of scale.

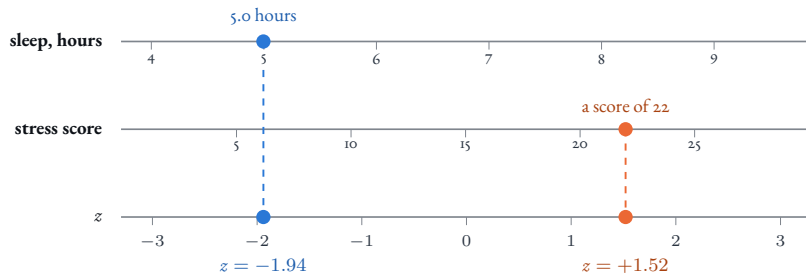
Why that matters

Every normal curve, in any unit, becomes one curve with mean 0 and SD 1: the **standard normal**. One table of areas then serves every variable.

Simulated class data. The minute conversion is Session 2's; the z-scores are our arithmetic.



Which is more unusual: 5 hours of sleep, or a stress score of 22?



Sleep: $(5.0 - 6.80) \div 0.93 = -1.94$.

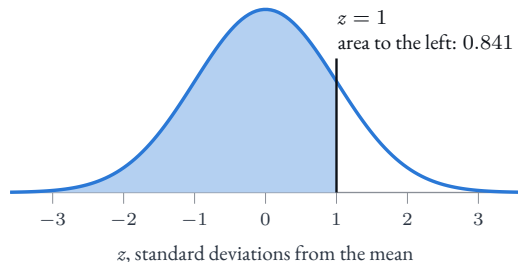
Stress: $(22 - 15.04) \div 4.57 = +1.52$.

On the normal curve, about 2.6% sleep less than 5 hours, and about 6.4% score above 22. **The short sleeper is further from typical.**

Simulated class data. Stress: the five-item score from Session 1, range 5 to 25. Areas are our arithmetic.



Reading areas from the standard normal curve



z	area to the left
-2.47	0.007
-1.94	0.026
-1.00	0.159
0.00	0.500
1.00	0.841
1.645	0.950
1.96	0.975
2.576	0.995

Excel: `NORM.S.DIST(z, TRUE)`. R: `pnorm(z)`.

Area to the right = $1 - \text{area to the left}$.

Area between two values = the difference of their areas.

Standard normal areas, rounded to three decimals.



From a z-score to a percentile, checked against the class

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

Samples vary

Confidence intervals

A published figure

Close

References

8.5 hours

$z = +1.83$, area to the left 0.966.

The curve says: 96.6% sleep less.

The class: 95.8% reported less than 8.5, and 98.3% reported 8.5 or less.

4.5 hours

$z = -2.47$, area to the left 0.007.

The curve says: 0.68%, under one student in 120.

The class: 2 students reported 4.5 hours or less.

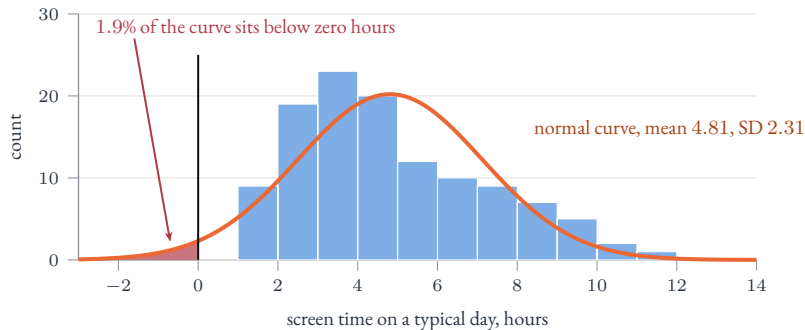
What the check shows

The curve cannot see that answers come in half hours, so it splits the difference. And real data often have fatter tails than the curve: more extreme values than it predicts.

Simulated class data. Normal areas and the comparison are our arithmetic.



When the normal curve is the wrong model



z for zero hours: $(0 - 4.81) \div 2.31 = -2.08$. The curve puts about 2.2 of our 117 students below zero hours of screen time.

Look before you use an area

The curve is symmetric; screen time is skewed.
Its areas are wrong in both tails.

Simulated class data. Normal areas are our arithmetic.



Act III

Samples vary, and by how much



Stat.BD

For the next hour, our 120 are the population

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

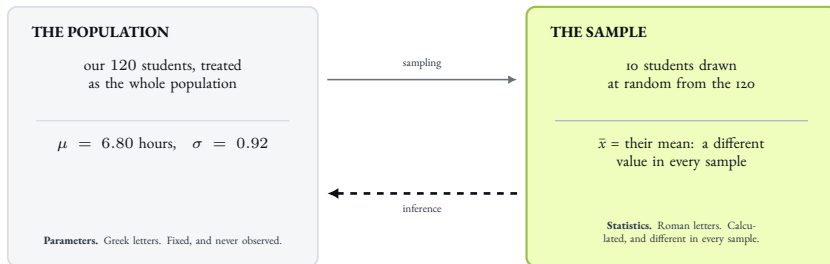
Samples vary

Confidence intervals

A published figure

Close

References

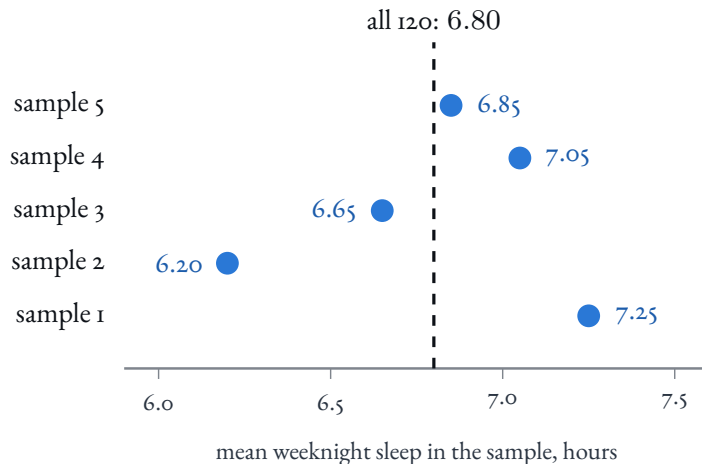


Everything measured lives on the right. Everything the study is actually about lives on the left. The dashed arrow is the whole of statistical inference, and it is the only one of the two that can be wrong.

Simulated class data. σ uses the population divisor N , so it is 0.92 rather than the sample SD 0.93.



Five samples of 10, five different means



Each is an honest random sample. None is wrong. They differ because each drew different students: this is **sampling variability**.

Simulated: five random samples of 10 from the simulated class (data/class-data.R).



Activity: build a sampling distribution together

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

Row plus column gives the id: row 50, column +7 is student 57.

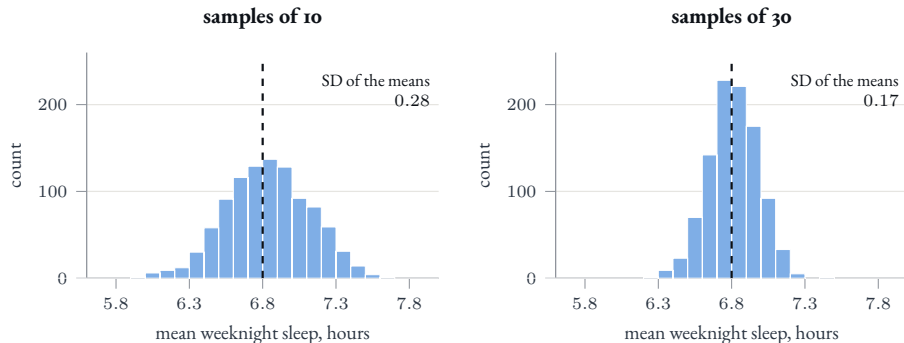
1. Each group draws 10 ids from 1 to 120, using a phone's random number tool. An id drawn twice: draw again.
2. Look up the 10 values. Average them.
3. Put a dot for your mean on the board's number line.
4. Repeat, as time allows.

Then: where is the pile centred? How wide is it?

Simulated weeknight sleep for all 120 students. The same grid is in the class notes.



A thousand samples draw the pattern



Each bar counts sample means. Both piles centre on the true 6.80. Samples of 30 crowd closer to it: the SD of their means is 0.17 against 0.28. The pile is the **sampling distribution of the mean**.

Simulated: 1,000 random samples of each size from the simulated class, drawn with replacement.



Three distributions not to confuse

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

Where inference lives

Inference lives in the third row. We never see it in a real study, but mathematics tells us its centre, its spread and its shape. That is why one sample can be enough.

Definitions.



The standard error: the SD of the sample means

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

One formula replaces a thousand samples.

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

The answer to the question we left on

For a class of 120:

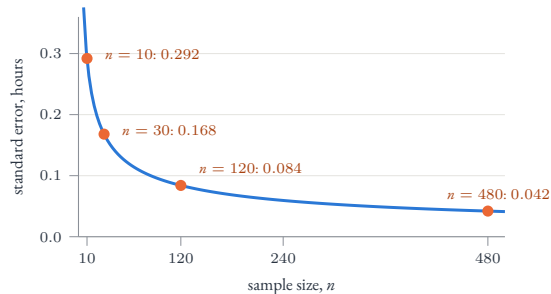
$$SE = 0.92 \div \sqrt{120} = 0.084 \text{ hours}$$

About 95% of such classes would average within $2 \times 0.084 \approx 0.17$ hours of the true mean. Not half an hour: about 10 minutes.

Simulated class data. SE by formula and by simulation, both ours.



The square-root law: four times the data, half the error



n	SE, hours
10	0.292
30	0.168
120	0.084
480	0.042

From 120 to 480 students: four times the forms to collect, for half the error.

Simulated class data, $\sigma = 0.92$. SE values are our arithmetic.



SD or SE? Two numbers that papers mix up

Standard deviation

How much *students* differ from each other.

Ours: 0.93 hours.

Does not shrink with a bigger sample.

Use it to **describe**.

Standard error

How much the *mean* would differ between samples.

Ours: 0.085 hours.

Shrinks as n grows.

Use it to **estimate**.

Mean $\pm$ what?

“ 6.80 ± 0.085 ” looks far more consistent than “ 6.80 ± 0.93 ”. A table that reports the SE where a reader expects the SD makes the data look tighter than they are. Always check which one it says.

Simulated class data. $SE = 0.93 \div \sqrt{120}$, using the sample SD.



The central limit theorem: averages become normal

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

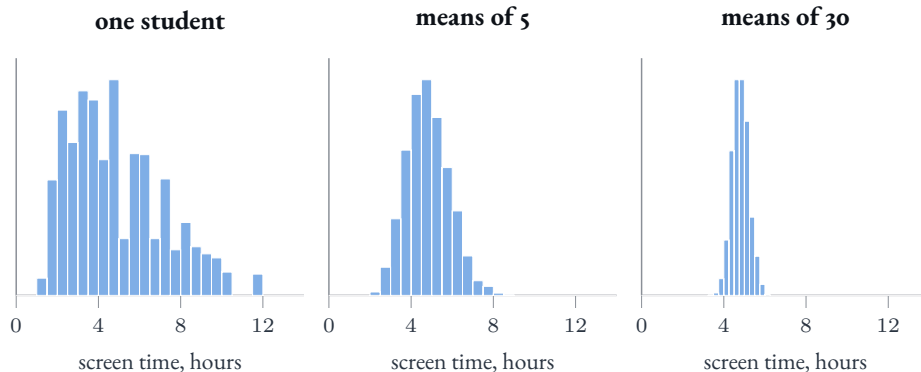
Samples vary

Confidence intervals

A published figure

Close

References



Screen time is skewed: skewness 0.69 for single students, 0.34 for means of 5, 0.13 for means of 30. Averaging cancels the extremes.

In plain words: whatever the shape of the data, the means of large enough samples pile up in a normal curve, centred on the population mean, with spread $\sigma/\sqrt{n}$.

Simulated: 2,000 random draws of each kind from the simulated screen-time column.



What the theorem promises, and what it does not

It does promise

- ▶ sample means close to normal for large n
- ▶ centred on the population mean
- ▶ spread $\sigma/\sqrt{n}$

It does not

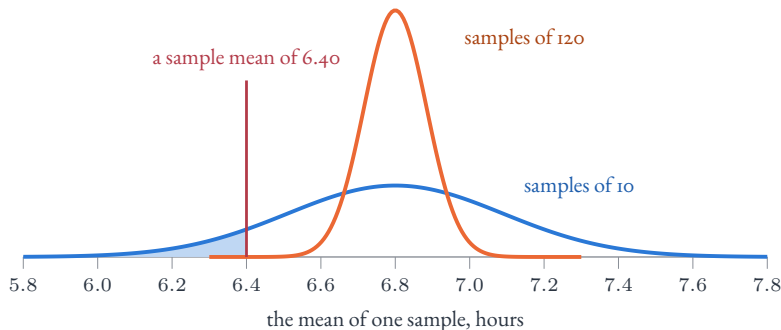
- ▶ make the *data* normal: screen time stays skewed
- ▶ fix a biased sample
- ▶ set a magic $n = 30$: heavy skew or outliers need more

“ $n \geq 30$ ” is a rule of thumb, not a law. For data close to normal, far fewer will do. For very skewed data, look at the shape of the means, not a rule.

Statement of the theorem.



How unusual is a sample averaging 6.40 hours?



Samples of 10:

$z = (6.40 - 6.80) \div 0.29 = -1.38$. About 8.39% of samples do this well or worse. (1,000 simulated: 8.1%.)

The same 0.40-hour gap is ordinary for a small sample and nearly impossible for a large one. **This is the logic of every test in Session 4.**

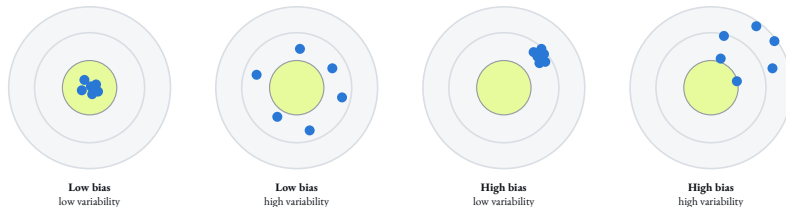
Simulated class data. z and normal areas are our arithmetic.

Samples of 120:

$z = (6.40 - 6.80) \div 0.084 = -4.76$. About one sample in 1, 000, 000.



The standard error measures scatter, not bias



The centre is the truth. Each dot is the estimate from one sample. A bigger sample pulls the dots closer together, and does nothing to where the cloud is centred.

The SE is the size of the cloud. It says nothing about where the cloud is centred. A snowball survey of 5,000 has a tiny SE and can still be centred on the wrong answer.

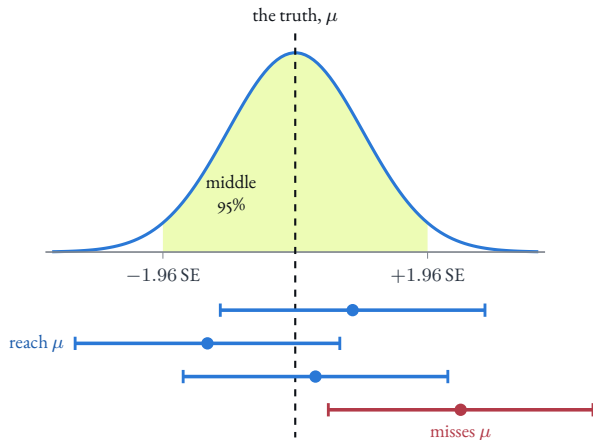
Dot positions illustrative (**constructed**).



Act IV

The confidence interval

Turning the picture round



95% of sample means land within 1.96 SE of the truth.

So, for 95% of samples, the truth lies within 1.96 SE of the sample mean.

A 95% confidence interval

$$\text{estimate} \pm 1.96 \times \text{SE}$$

Positions illustrative (**constructed**), in SE units.



Our 95% interval for weeknight sleep, by hand

- 1 mean 6.80, SD 0.93, $n = 120$
- 2 $SE = 0.93 \div \sqrt{120} = 0.085$
- 3 multiplier: 1.980 (next frame)
- 4 margin = $1.980 \times 0.085 = 0.17$
- 5 6.80 ± 0.17

6.63 to 6.97

hours: the 95% confidence interval for the mean

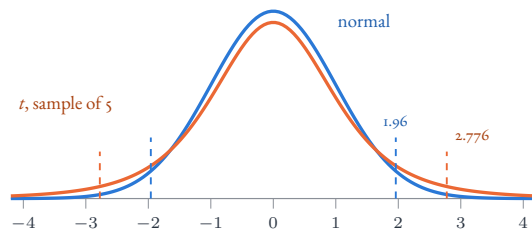
The sentence

We are 95% confident that the average weeknight sleep of students like these lies between 6.63 and 6.97 hours.

Simulated class data. The interval is ours; Session 1 showed it, and promised this frame.



Why 1.980 and not 1.96: the t distribution



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

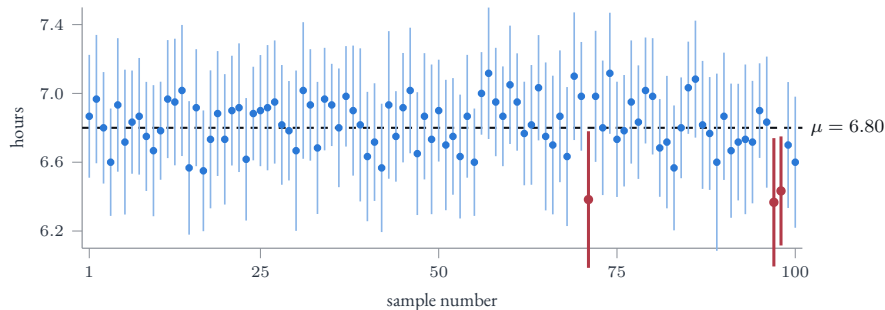
df, degrees of freedom: $n - 1$ here.

We used s , estimated from the sample, in place of the unknown σ . That adds uncertainty, and the t distribution pays for it with fatter tails and a wider multiplier.

t and normal quantiles, rounded to three decimals.



A hundred samples, a hundred intervals



Each line is one sample of 30 from our 120, with its 95% interval. 97 reach the true 6.80; 3 miss it. In the long run, 95 in 100 reach it.

Any one interval either contains μ or does not. In a real study, we never learn which.

Simulated: 100 random samples of 30 from the simulated class, each with a t interval.



What “95% confident” means, and what it does not

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.

Interpretation of a frequentist confidence interval.



Not one student sleeps inside the interval

0 of 120

students report a weeknight sleep between 6.63 and 6.97 hours

Every answer was to the half hour, so the nearest values are 6.5 and 7.0.

Where the average is	6.63 to 6.97
-----------------------------	--------------

Where most students are	about 4.9 to 8.7
--------------------------------	------------------

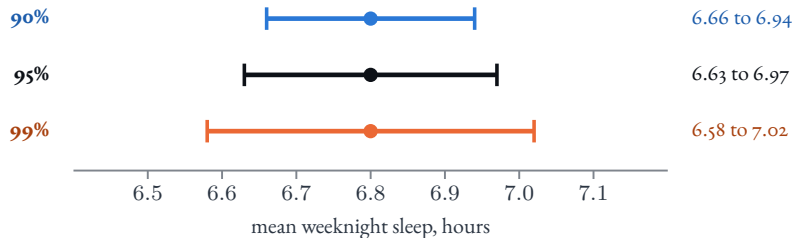
The interval is not a range of people

A confidence interval narrows as the sample grows. The spread of people does not.

Simulated class data. “About 4.9 to 8.7” is mean ± 2 SD, Session 2’s band; our arithmetic.



More confidence costs width



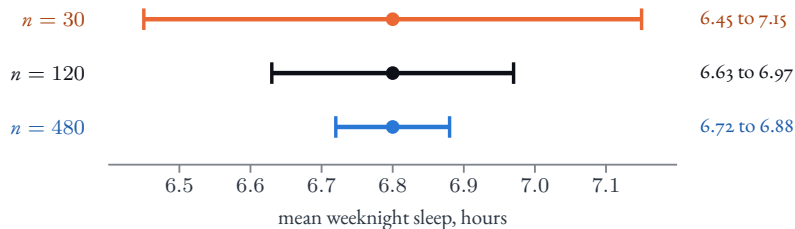
Multipliers: 1.658, 1.980, 2.618. Same data, same SE; only the multiplier changes.

A 100% interval would run from zero to 24 hours: certain, and useless.

Simulated class data. Intervals are our arithmetic.



More students, narrower interval



Margins: ± 0.35 , ± 0.17 , ± 0.08 . Each fourfold increase roughly halves the margin; the multiplier shrinks a little too.

Hypothetical: the same mean and SD as our **simulated** class, with 30 or 480 students. Our arithmetic.



An interval for a percentage

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

Samples vary

Confidence intervals

A published figure

Close

References

59 of 120 give tuition: $p = 0.4917$.

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4917 \times 0.5083}{120}} = 0.0456$$

margin = $1.96 \times 0.0456 = 0.0894$

0.4917 ± 0.0894 , which is 40.2% to 58.1%.

Session 2: a proportion is the mean of a 0/1 column. The same machinery applies.

40.2 **to** 58.1%

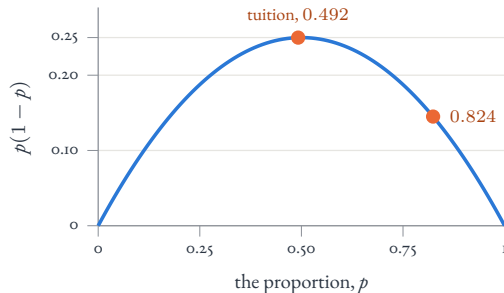
of students like these give tuition: 95% confidence interval

Works when both counts, yes and no, are at least about 10. Here 59 and 61.

Simulated class data. The interval is our arithmetic.



Why 50% is the worst case



The SE of a percentage depends on $p(1-p)$, and that peaks at $p = 0.5$.

Our tuition share, 0.492, sits almost at the peak: the widest interval 120 people can give.

When planning a survey before you know p , assume 0.5. Your margin can then only turn out smaller.

$p(1-p)$ is algebra. The two marked points: our **simulated** tuition share, and Islam 2020's 392/476.



Where the “ ± 3 points” in a news poll comes from

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

Samples vary

Confidence intervals

A published figure

Close

References

A poll of 1,000 people, planning with $p = 0.5$:

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

So about ± 3.1 percentage points.

A quick version: $1/\sqrt{n} = 1/\sqrt{1000} \approx 3.2$ points.

Like every interval, it covers sampling error only. Act V shows what it leaves out.

The surprising part

The population's size is not in the formula.
A random sample of 1,000 tells you as much
about Bangladesh as about Rajshahi.

Our arithmetic. The population-size point holds while the sample is a small fraction of the population.



How many students would we need?

For a mean, within $\pm E$:

$$n = \left(\frac{1.96 \times \sigma}{E} \right)^2$$

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

For a percentage, within $\pm E$:

$$n = \frac{1.96^2 \times 0.25}{E^2}$$

Within ± 5 points: 384.2, so 385.
Within ± 3 points: 1068.

Two habits

Always round *up*. Then allow for people who will not answer: expecting 70% to reply, invite $385 \div 0.70 = 550$.

Our arithmetic. The planning SD comes from the **simulated** class; a real plan would take it from a pilot or a published study.



How to report an estimate

Write it like this

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).

Not like this

“Sleep was 6.80 ± 0.17 .” An SD? An SE? A margin?

“Sleep was 6.8000 hours.” Four decimals the data cannot support.

Say which interval, give n , and round to what the data can bear.

Simulated class data.



An interval already answers a question for Session 4

From Sample to
Population

Muhtasim Munif Fahim,
MSc

Orientation

Probability

The normal curve

Samples vary

Confidence intervals

A published figure

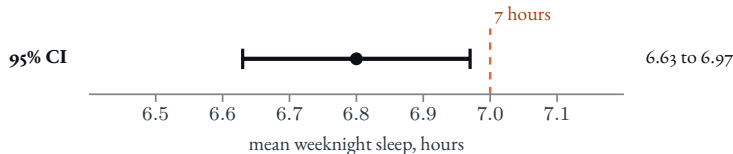
Close

References

Adults are advised to sleep 7 hours or more a night (Watson 2015).

7 lies outside 6.63 to 6.97. So 7 is not a plausible value for the average, at the 95% level.

Is our population's average below 7, or could our 6.80 be luck?



Session 4 asks the same question as a **test** and gets a p -value. Interval and test will agree.

Simulated class data. Recommendation: Watson NF et al., *Sleep* 2015; PMID 26039963.



Act V

An interval for a published figure

A published prevalence, and the interval it left out

Islam and colleagues, 2020. 476 university students in Bangladesh, online, May 2020.

392 had mild to severe depressive symptoms on the PHQ-9: 82.4%. **No interval is printed.**

78.9 to 85.8%

95% interval: ours, not the paper's

$$p = 392 \div 476 = 0.8235$$

$$SE = \sqrt{\frac{0.8235 \times 0.1765}{476}} = 0.0175$$

$$\text{margin} = 1.96 \times 0.0175 = 0.0343$$

0.8235 ± 0.0343 , which is 78.9% to 85.8%.

Islam MA et al. *PLoS One* 2020;15(8):e0238162, Results and Table 1. PMID 32845928. CC BY 4.0. The interval is our arithmetic.



The interval covers sampling error, and nothing else

The formula assumes a **random sample**. The authors, in their own words:

“the snowball sampling strategy was chosen instead of random samples”

The survey spread through Facebook and Messenger, from friend to friend.

The true uncertainty is larger than seven points, by an amount no formula can give.

What 78.9 to 85.8% does not cover

- ▶ who could be reached: students online, in the authors' networks
- ▶ who chose to answer
- ▶ the PHQ-9 is a screening score, not a diagnosis

Islam MA et al. *PLoS One* 2020; quotation from its Strengths and limitations section. PMID 32845928.



What a margin of error leaves out

Error	Covered by the interval?
Sampling	Yes. The luck of which people were drawn.
Coverage	No. People the method could never reach.
Non-response	No. People reached who did not answer.
Measurement	No. Question wording; a screening tool read as a diagnosis.
Processing	No. The 25-hour typo; the 87.7% that should be 81.7%.

One sentence to keep

A margin of error is a floor on the uncertainty, not a ceiling.

The two processing errors are Sessions 1 and 2's: our simulated typo, and Islam 2020's printed percentage.



What today established

1. A probability is a long-run share. “Given that” changes the denominator.
2. A z-score counts SDs from the mean, and compares values across scales. Check the shape before using a normal area.
3. Sample means vary. Their SD is the standard error, $\sigma / \sqrt{n}$, and their shape tends to normal: the central limit theorem.
4. A 95% interval is estimate $\pm$ about 2 SE. The 95% belongs to the method.
5. The interval covers sampling error only. Bias is outside it.



Session 4: Testing hypotheses and comparing groups

The null hypothesis, the p -value and how it is misread, t -tests and ANOVA: deciding whether a difference could be luck.

Before then

Read Chapter 3 of the class notes. Try the Part A exercises. Bring a calculator.

The tuition and stress comparison uses the **simulated** class data.

One question to think about

Students who give tuition score higher on our stress scale than those who do not. How big would the gap need to be before you stopped calling it luck?



- ▶ 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. CC BY 4.0.
- ▶ 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. doi:10.5665/sleep.4716.
- ▶ Student [Gosset WS]. The probable error of a mean. *Biometrika* 1908;6(1):1–25. doi:10.1093/biomet/6.1.1.
- ▶ Class data: **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from a fixed seed. Not real students. Every figure computed from it, including all simulations of repeated sampling, is ours.
- ▶ Constructed examples, made up for teaching: the viva and true/false examples, the interval positions in the “turning the picture round” figure, and the sample-size plans.

