

Describing Data

Tables, averages, spread and charts, before any test

Muhtasim Munif Fahim, MSc

Stat.BD

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

Session 2 of 14 • Theory

Session 1 gave us

- ▶ a question, and the variables it needs
- ▶ each variable's type and scale
- ▶ a codebook that turns answers into numbers
- ▶ 120 rows, one per student

Today, in five acts

-
- | | |
|-----|---|
| I | Tables for categories |
| II | The centre: mean, median, mode |
| III | The spread, and the values that stand out |
| IV | Charts, honest and otherwise |
| V | A table in a published paper |
-

The rule for the whole session

Before any test, look. Every mistake in Sessions 3 to 6 is easier to make if you skipped today.

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



The test of this session

Given any column of a dataset, choose the right summary and the right chart, compute it, and say in one plain sentence what it shows.

- ▶ Build a frequency table and a cross-tabulation, and pick the right percentage
- ▶ Compute a mean, median and standard deviation by hand, and say which to report
- ▶ Read a box plot, and decide what to do about a value that stands out
- ▶ Choose a chart for the data, and spot a chart built to mislead



The question we left on: one student typed 25 hours

25.0

hours of self-study on a typical class day, as typed in row 37 of 120

A day has 24 hours. The student meant 2.5 and hit the wrong key.

Write down your guesses

1. How far does the class **average** move?
2. How far does the **middle value** move?
3. Would you notice from the average alone?

Simulated class data. The error was planted in the simulated raw file on purpose, to be found today.



Act I

A table for every category



A frequency table counts how often each answer occurs

Faculty	n	%
Arts & Humanities	29	24.2
Business Studies	23	19.2
Science & Engineering	33	27.5
Social Sciences & Law	35	29.2
Total	120	100.1

Percentage = $\text{count} \div \text{total} \times 100$. For Arts:
 $29 \div 120 \times 100 = 24.2$.

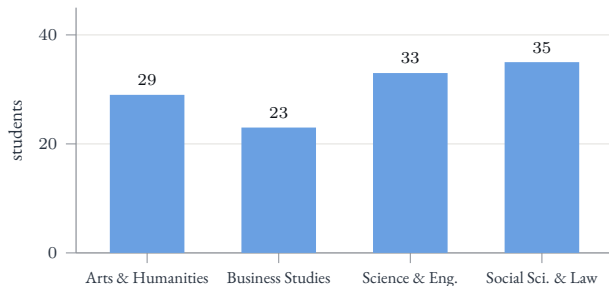
Simulated class data, $n = 120$. The faculty grouping is ours, from the codebook.

Why the total is 100.1

Each percentage was rounded to one decimal. Rounded parts need not add to exactly 100. Leave the total as it is and say so in a footnote. Never nudge a number to make it add.



A bar chart draws the same table



Simulated class data, $n = 120$.

Three rules

1. Gaps between bars: the categories are separate.
2. The axis starts at zero, because the bar's *length* is the message.
3. For nominal data, any order will do; sorting by size helps the eye.



Three percentage columns: of everyone, of those who answered, and running

CGPA band	<i>n</i>	% of all	Valid %	Cumulative %
below 2.50	9	7.5	8.6	8.6
2.50–2.99	39	32.5	37.1	45.7
3.00–3.49	46	38.3	43.8	89.5
3.50–4.00	11	9.2	10.5	100.0
no result yet	15	12.5		
Total	120	100.0	100.0	

% of all divides by 120, everyone.

Valid % divides by 105, those with a result.

Cumulative % adds the valid column down: 45.7% of those with a result are below 3.00.

Simulated class data. SPSS prints exactly these columns, as Session 7 will show.



A running total only makes sense when the categories have an order

Ordinal: cumulative works

CGPA bands run low to high, so “below 3.00” means something: 45.7%.

Nominal: cumulative is noise

Faculties have no order. “Arts plus Business” is 43.4% of the class, and that number answers no question anyone asked.

This is Session 1’s scale of measurement paying off. The scale decides which columns of a table mean anything.

Simulated class data. $43.4 = 24.2 + 19.2$, our arithmetic.



Numbers can be grouped into classes and counted the same way

Self-study, hours	n	%	Cum. %
0 to under 1	2	1.7	1.7
1 to under 2	27	22.5	24.2
2 to under 3	45	37.5	61.7
3 to under 4	23	19.2	80.8
4 to under 5	16	13.3	94.2
5 to under 6	6	5.0	99.2
6 to under 7	1	0.8	100.0
Total	120		

Rules for the classes

- ▶ Equal widths, so the counts compare fairly.
- ▶ “2 to under 3”: a student at exactly 3 goes in one class only.
- ▶ Enough classes to show a shape; five to ten is usual.

Read it: 61.7% study under three hours a day.

Simulated class data, corrected file, $n = 120$.



A cross-tabulation counts two variables at once

Lives in	Tuition or a job		Total
	No	Yes	
Hall	30	17	47
Mess	11	32	43
Family home	20	10	30
Total	61	59	120

Each student sits in exactly one of the six cells.

Simulated class data, $n = 120$.

How to read it

The **margins** are the two single-variable tables. The **inside** is new: it shows how the two variables go together.



Row, column and total percentages answer three different questions

"Yes" cell	Row % of that residence	Column % of all who say yes	Total % of all 120
Hall	36.2	28.8	14.2
Mess	74.4	54.2	26.7
Family home	33.3	16.9	8.3

Row: $17 \div 47$.

Of hall students, how many work?

Column: $17 \div 59$.

Of those who work, how many live in a hall?

Total: $17 \div 120$.

Of everyone, how many are working hall students?

Simulated class data. Same 17 students in every column; three denominators.



Pick the percentage from the question you are asking

Question	Use	Answer
Are mess students more likely to do tuition than hall students?	row %	74.4% against 36.2%
Where do the students who do tuition live?	column %	54.2% in a mess
What share of the class is a mess student doing tuition?	total %	26.7%

The rule

Compare groups by dividing *within* each group. The variable that defines the groups you compare, here residence, is the one whose totals go in the denominator.

Simulated class data. Session 5 asks whether a gap this size could be chance.



Act II

Where the middle is

The mean: add them up, divide by how many

Self-study hours, rows 1 to 7 of the class data:

2.5, 2.0, 3.0, 3.5, 3.0, 2.5, 2.0

$$\bar{x} = \frac{\sum x}{n} = \frac{18.5}{7} = 2.64$$

Simulated class data, first seven rows.

In words

If the total study time were shared out equally, each of the seven would get 2.64 hours.

$\bar{x}$ is “x-bar”, the sample mean. $\sum$ means “add them all”.



The median: the middle value once they are sorted

Seven values, sorted

2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5

The 4th of 7 is the middle: median = 2.50.

Add an eighth student, who studies 6.5

2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5, 6.5

Two middles: median = $(2.5 + 3.0)/2 = 2.75$.

What moved

Mean: 2.64 $\rightarrow$ 3.125

Median: 2.50 $\rightarrow$ 2.75

One large value pulled the mean almost twice as far as the median.

Simulated rows 1 to 7. The eighth student is added by us for the example.



The mode: the most common value

Rows 1 to 7

2.0, 2.5 and 3.0 each appear twice. Three modes.
The mode need not be unique, and in a small sample it is often useless.

All 120 students

Most common self-study time: 2.0 hours.

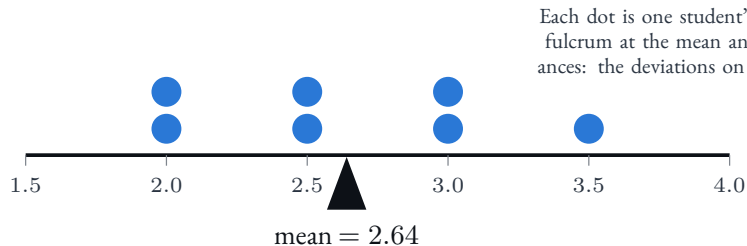
Simulated class data.

Where the mode earns its place

For nominal data it is the *only* centre. The modal faculty is Social Sciences & Law, with 35 students. There is no “mean faculty”.



The mean is the balance point of the data



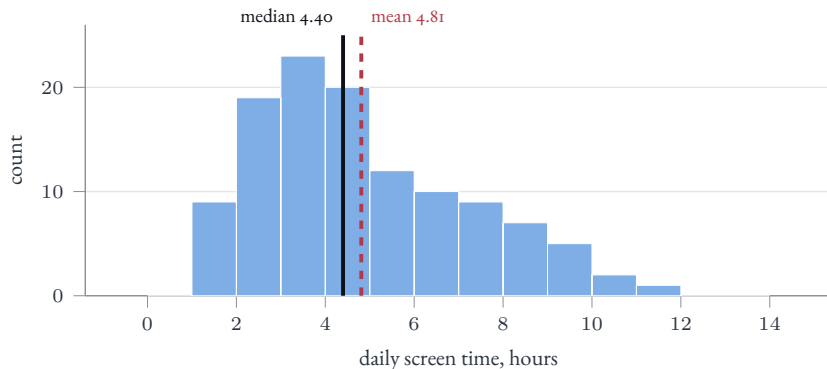
Each dot is one student's hours. Put the fulcrum at the mean and the beam balances: the deviations on each side cancel.

Deviations from the mean: $-0.14, -0.64, 0.36, 0.86, 0.36, -0.14, -0.64$. Before rounding they sum to exactly zero. They always do.

Simulated rows 1 to 7.



When one tail is long, the mean follows it

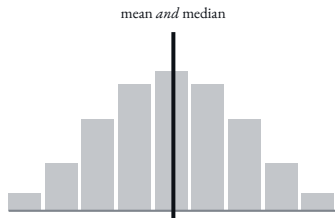


Most students cluster between 2 and 7 hours; a few run to 10 and beyond. Those few pull the mean, 4.81, to the right of the median, 4.40.

Simulated class data, $n = 117$ (3 answered “don’t know”).

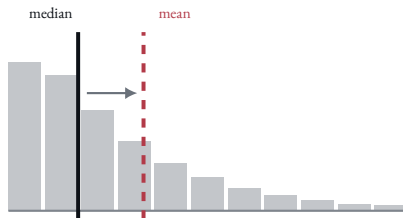


Symmetric or skewed decides how the mean and median relate



Symmetric

The two coincide. Either one describes the centre, and the mean is the more useful because later methods are built on it.



Skewed to the right

The long tail drags the mean away from the bulk of the data. The median stays where most of the observations are.

Both distributions are illustrative. The rule they carry is not: **the further apart the mean and the median, the more skewed the column**, and the median is the summary that survives the skew.



Which centre to report

The variable	Centre	Why
Nominal: faculty, residence	mode	No order, so no middle; no numbers, so no mean.
Ordinal: CGPA band, a Likert answer	median	Order, so a middle exists. The gaps are not equal, so a mean is shaky.
Numerical, roughly symmetric: weeknight sleep	mean	Uses every value, and later methods build on it.
Numerical, skewed or with outliers: screen time, income	median	Resists the tail.

The median CGPA band is 3.00–3.49: the cumulative percent first passes 50 there (45.7, then 89.5).

Simulated class data.



The mean of a 0/1 column is a proportion

Tuition or a job, coded No = 0, Yes = 1:

0, 1, 1, 0, 1, 0, ...

Adding the column counts the 1s: 59.

Dividing by 120:

$$\bar{x} = \frac{59}{120} = 0.492$$

Simulated class data, $n = 120$.

Why it matters

A proportion is a mean in disguise. Everything we learn about means, including the margin of error in Session 3, carries over to proportions.

So 0.492 is 49.2%, the same number as in the frequency table.



The average of three group means is not the class mean

Lives in	<i>n</i>	Mean sleep
Hall	47	6.74
Mess	43	6.53
Family home	30	7.27

Average of the three means: 6.85

The class mean: 6.80

Weight by size

$$\frac{47(6.74) + 43(6.53) + 30(7.27)}{120} = 6.80$$

Each group counts in proportion to its size.

Simulated class data, weeknight sleep in hours.



Your CGPA is already a weighted mean

Course	Credits	Grade point
Course A	3.0	3.75
Course B	3.0	3.25
Course C	2.0	4.00
Lab	1.5	2.50

Plain mean of the grade points: $13.50 \div 4 = 3.375$

Weighted by credits

Multiply each grade point by its credits and add:

$$11.25 + 9.75 + 8.00 + 3.75 = 32.75.$$

Divide by the total credits, 9.5:

$$\frac{32.75}{9.5} = 3.45$$

The low lab grade carries only 1.5 credits, so it pulls less.

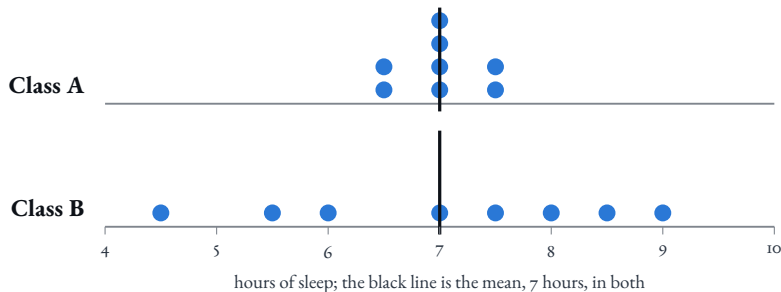
Constructed example; the grade points are made up. Your department's grading rules are the authority.



Act III

How far apart the values are

Same average, different classes



Both classes average 7 hours of sleep. In Class A nearly everyone is close to 7. In Class B the mean describes almost nobody. A centre without a spread is half a summary.

Constructed: two made-up classes of eight, drawn to share a mean.



The range uses two values and ignores the rest

range = largest – smallest

Self-study	Min	Max	Range
corrected	0.5	6.5	6.0
as typed	0.5	25.0	24.5

Fragile by design

One typing error quadrupled the range. The range is built from the two most extreme values, which are exactly the ones most likely to be mistakes.

Still worth printing: the minimum and maximum are the fastest check for impossible values.

Simulated class data. The ranges are our subtraction.



x	$x - \bar{x}$	$(x - \bar{x})^2$
2.5	-0.14	0.02
2.0	-0.64	0.41
3.0	0.36	0.13
3.5	0.86	0.73
3.0	0.36	0.13
2.5	-0.14	0.02
2.0	-0.64	0.41
sum		1.86

Four steps

1. Subtract the mean, 2.64, from each value.
2. Square each deviation, so negatives stop cancelling.
3. Add the squares, and divide by $n - 1 = 6$:

$$s^2 = \frac{1.86}{6} = 0.31$$

4. Take the square root, back to hours:

$$s = \sqrt{0.31} = 0.56 \text{ hours}$$

Simulated rows 1 to 7. Squares shown rounded; the sum uses unrounded values.



Why divide by $n - 1$ and not by n

The short reason

The deviations are measured from the *sample* mean, which sits in the middle of these seven values by construction. Around the true population mean they would be a little larger. Dividing by $n - 1$ makes up the shortfall, on average.

Divide by 7: $s = 0.52$. Divide by 6: $s = 0.56$. With 120 students the difference is under half of one percent.

Divide-by-7 figure: our arithmetic on the **simulated** rows 1 to 7.

Degrees of freedom

The seven deviations must sum to zero. Know any six and the seventh is forced. Only $n - 1 = 6$ of them are free to vary.



In a mound-shaped distribution, most values sit within two SDs

Weeknight sleep: mean 6.80, SD 0.93

Band	Hours	Inside
mean $\pm$ 1 SD	5.87 to 7.73	70.8%
mean $\pm$ 2 SD	4.94 to 8.66	96.7%

The rough rule

For a roughly symmetric, single-humped distribution: about two-thirds of values within 1 SD of the mean, about 95% within 2 SD.

Screen time is skewed. Its shares still land near the rule, but all 5 students beyond 2 SD sit on the high side, none on the low.

Simulated class data. The band limits are our arithmetic from the rounded mean and SD.



Percentiles and quartiles cut the sorted data into parts

Name	Below it	Screen, hours
10th percentile	10%	2.20
Q_1 , 25th	25%	3.00
median, 50th	50%	4.40
Q_3 , 75th	75%	6.30
90th percentile	90%	8.12

Interquartile range: $Q_3 - Q_1 = 3.30$ hours

What the IQR says

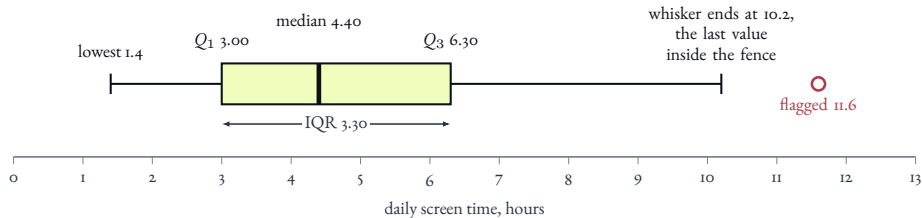
The middle half of the class uses its phone between 3.00 and 6.30 hours a day. The extremes cannot move it.

SPSS and R use slightly different rules for quartiles. With small samples they can disagree a little. Neither is wrong.

Simulated class data, $n = 117$. Quartiles by R's default rule.



A box plot draws five numbers

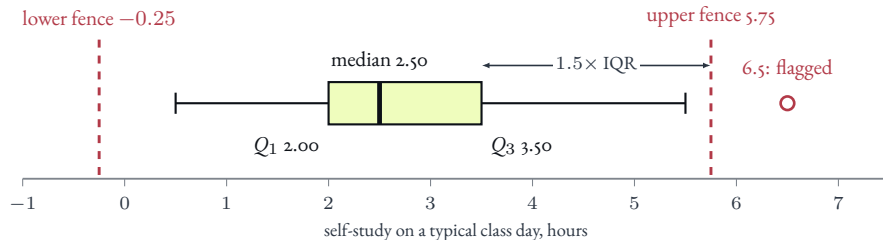


The box is the middle half. The thick line is the median. The whiskers reach to the last values inside the fences; anything beyond is drawn on its own.

Simulated class data, $n = 117$.



A fence flags a value. It does not convict it



Upper fence:

$$Q_3 + 1.5 \times IQR = 3.50 + 1.5 \times 1.50 = 5.75$$

In the corrected data, one student reports 6.5 hours: flagged.

Flagged is not wrong

Six and a half hours of study before an exam is entirely possible. A rule found an unusual value; only you can decide if it is an error.

Simulated corrected file. Fences by Tukey's rule (Tukey 1977); computed by us.



The answer to the 25-hour question



	as typed	corrected
Mean	2.81	2.62
Median	2.50	2.50
SD	2.38	1.21
IQR	1.50	1.50
Range	24.5	6.0

What one wrong key did

The mean moved by $22.5 \div 120 \approx 0.19$ hours, easy to miss. The SD almost doubled. The median and IQR did not move at all. And the codebook's allowed range, 0 to 16, would have caught it before any of this.

Simulated: raw and corrected files. The shift $22.5 \div 120$ and the IQR and ranges are our arithmetic.



What to do with a value that stands out

1. **Is it possible?** 25 hours in a day is not. Go back to the form. Correct it if you can; code it missing if you cannot. Write down what you did.
2. **Is it plausible?** 6.5 hours of study is. Keep it.
3. **Does it change the answer?** Analyse with and without it, and report both if they differ.
4. **Report it.** “One value was corrected from 25 to 2.5 after checking the paper form.”

Never

Never delete a value because it is inconvenient, or because it makes a result “significant”. That is how honest data become dishonest results.



Change the units, and the summaries follow predictably

Multiply by 60, hours to minutes

Mean 6.80 $\rightarrow$ 408 min.

SD 0.93 $\rightarrow$ 55.8 min.

Both are multiplied.

Add half an hour to everyone

Mean 6.80 $\rightarrow$ 7.30.

SD stays 0.93.

Shifting everyone moves the centre, not the spread.

Simulated weeknight sleep. The converted values are our arithmetic.

Why it matters

Spread is about distances between students. Adding the same amount to everyone changes no distance. Multiplying stretches every distance by the same factor.



The coefficient of variation compares spread across scales

$$CV = \frac{SD}{\text{mean}} \times 100$$

Variable	Mean	SD	CV %
Weeknight sleep	6.80	0.93	13.6
Self-study	2.62	1.21	46.3
Screen time	4.81	2.31	48.1

Simulated class data.

Read it

Students differ far more, relative to the average, in how much they study than in how much they sleep.

Ratio scales only

The CV needs a true zero. The stress score runs from 5 to 25; its CV means nothing.



Skewness puts a number on the lean

Variable	Skewness	Mean – median
Self-study	0.74	+0.12
Screen time	0.67	+0.41
Stress score	0.09	+0.04
Weeknight sleep	-0.17	-0.20

A rule of thumb, not a law. Look at the histogram; the number only summarises it.

Simulated class data. Mean minus median: our subtraction.

A common rule of thumb

Between -0.5 and $+0.5$: roughly symmetric.
Beyond ± 1 : strongly skewed.
Positive leans right; negative leans left.



Report mean (SD) for symmetric data, median (IQR) for skewed

Variable	<i>n</i>	Summary
Weeknight sleep, hours	120	mean 6.80 (SD 0.93)
Weekend sleep, hours	119	mean 7.87 (SD 1.05)
Screen time, hours	117	median 4.40 (IQR 3.00 to 6.30)
Self-study, hours	120	median 2.50 (IQR 2.00 to 3.50)
Tuition or a job	120	59 (49.2%)
Lives in a mess	120	43 (35.8%)

This is “Table 1”

Almost every paper opens its results with a table like this. You can now read one, and build one.

Simulated class data. Each row's *n* is the number who answered that question.



Act IV

Charts, honest and otherwise



Choose the chart from the variables

What you have	Draw	Avoid
One categorical	bar chart	a pie with many slices
One numerical	histogram, box plot	a bar for each person
Two categorical	bars of row percentages	two pies side by side
One numerical, across groups	box plots side by side	bars of the means alone
Two numerical	scatter plot	two separate histograms
One numerical, over time	line chart	bars with a cut axis

The type of each variable, from Session 1, decides the chart. The rest of this act draws each row.



A histogram shows the shape of one numerical variable

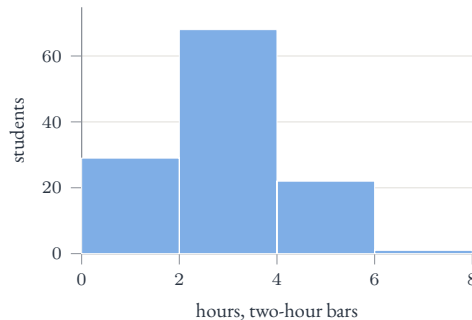
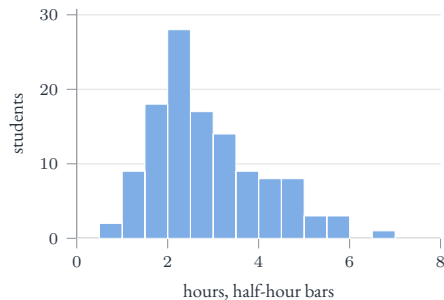


No gaps between bars: the scale is continuous. The bulk sits at 1.5 to 3 hours and the right side trails off. Right-skewed, as the numbers said.

Simulated class data, corrected file, $n = 120$. Half-hour bars.



The width of the bars can change the story

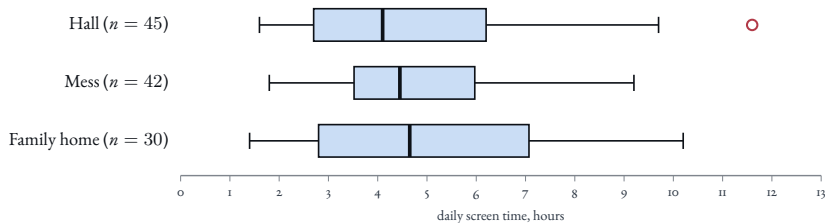


Same 120 students. With two-hour bars the tail vanishes into one bar, and the skew is hard to see. Too narrow, and every bar is noise.

Simulated class data. Always try more than one bar width before choosing.



Box plots side by side compare a number across groups

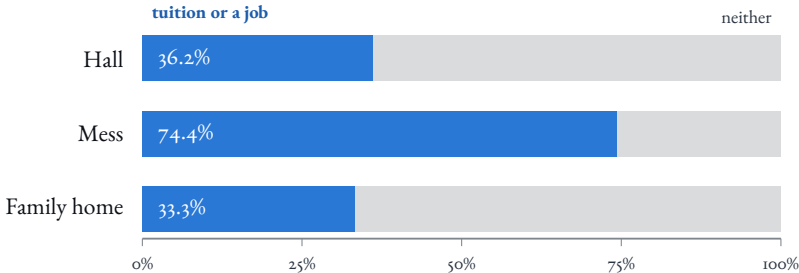


Medians sit close together, and the boxes overlap heavily. Residence makes little difference to screen time here. The hall group has the one flagged value.

Simulated class data. Group sizes are those who answered the screen-time question.



Stacked 100% bars compare the make-up of groups

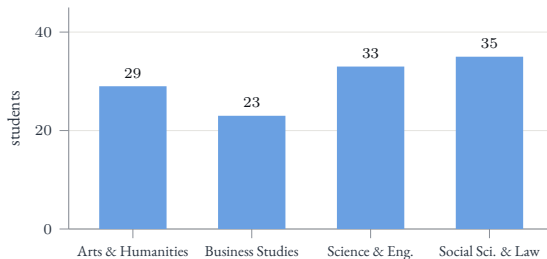
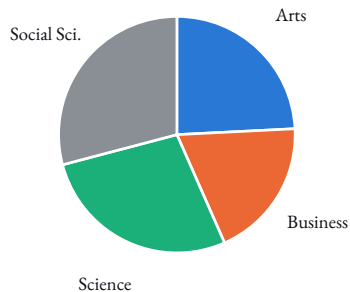


Each bar is one residence group, stretched to 100%. The blue part is the **row percentage** from Act I. The group sizes are hidden: 47, 43 and 30. Print them beside the bars.

Simulated class data.



A pie chart asks the eye to compare angles, which it does badly

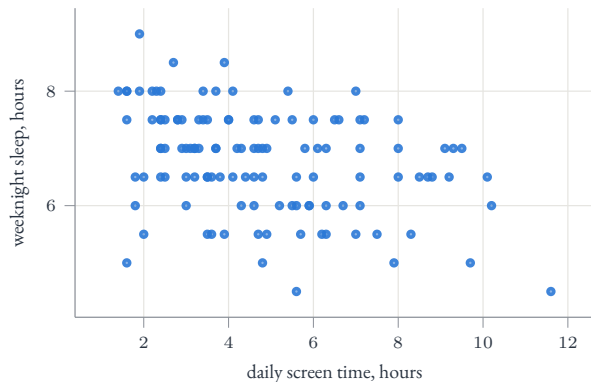


Which faculty is largest? In the pie, Science and Social Sciences look alike. In the bars, 33 against 35 is obvious.

Simulated class data, the same four numbers in both charts.



A scatter plot shows two numbers together



Simulated class data, $n = 117$ with both answers.

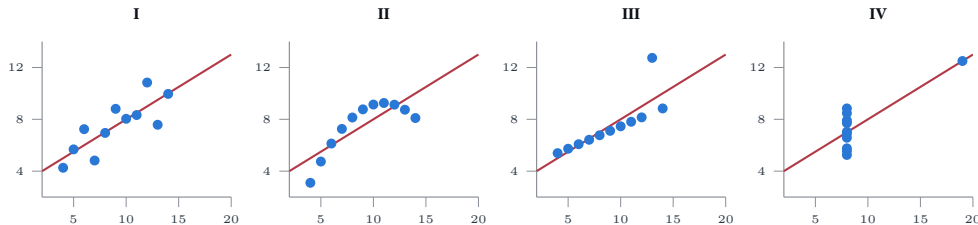
One dot per student.

More screen time goes with a little less sleep, with plenty of scatter around the trend.

Describe three things: direction, strength, and any points that sit apart. Session 5 puts a number on the strength.



Four datasets, identical summaries, nothing alike



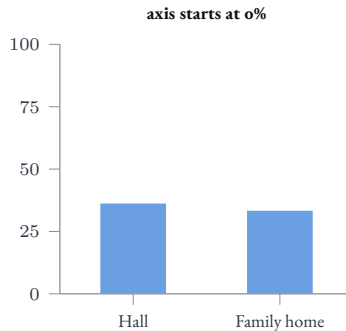
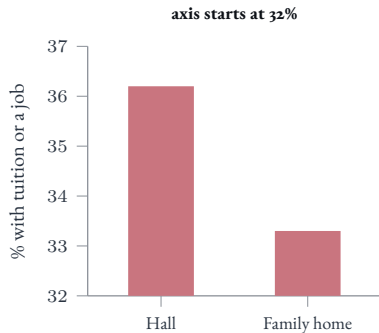
All four: mean of $x = 9.00$, mean of $y = 7.50$, correlation ≈ 0.82 , fitted line $y = 3.00 + 0.500x$.

I: a loose line. II: a curve. III: a tight line and one stray point. IV: one point makes the whole trend.

Anscombe 1973, *The American Statistician*, via R's `anscombe` data. Summaries computed by us. To three places the correlations are 0.816 for sets I to III and 0.817 for set IV.



A bar chart that does not start at zero exaggerates



Hall 36.2%, family home 33.3%. On the left the hall bar is more than three times as tall; the real gap is under three points.

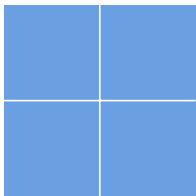
Simulated class data. Same two numbers in both panels; only the axis differs.



A pictogram that doubles the side quadruples the area



last year: 10



this year: 20

The value doubled, so the designer doubled the side. The area, which is what the eye reads, went up four times: the white lines show four copies of last year's square.

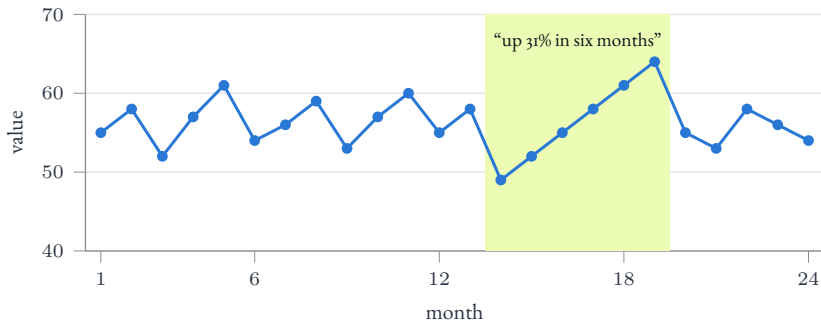
The fix

Scale one dimension only, as in a bar, or repeat the icon: two small squares for 20.

Constructed example.



Choose the window, and you choose the story



Over 24 months the series goes nowhere. Months 14 to 19 alone run from 49 to 64. A press release calls that “up 31% in six months”. It is true, and it is misleading.

Constructed series. $64 \div 49 = 1.31$, our arithmetic.



Print the count next to every percentage

Honest

74.4% of mess students do tuition or a job (32 of 43).

Technically true

“67% of Masters students in our department agree” (2 of 3).

With 43 students, one person moves the percentage by 2.3 points. With 3, one person moves it by 33.

Left: **simulated** class data. Right: **constructed**. Points per person, $100 \div n$, our arithmetic.



Activity: put a real chart through six questions

The checklist

1. Is the chart type right for the variables?
2. Does a bar axis start at zero?
3. Are areas or icons scaled honestly?
4. How many people is each percentage based on?
5. Why does the time window start and end where it does?
6. Who made it, and from what data?

In pairs, 5 minutes

Find one chart on your phone: a news site, a Facebook page, a company advert. Put it through the six questions. Report the worst problem you found.



Activity: describe one variable, by hand

In groups of four, 8 minutes

Use the 20-row extract of the class data in the class notes. Your group gets one variable.

Produce:

1. the right table or summary numbers
2. the right chart, sketched on paper
3. one plain sentence a friend would understand

Group	Variable
1	faculty
2	residence
3	CGPA band
4	self-study hours
5	weekend sleep
6	screen time

The extract is 20 rows of the **simulated** class data.



Act V

Reading a table in a published paper



Stat.BD

Islam 2020 prints its cross-tab as percentages of all 476

Tuition	Depression, PHQ-9 category				
	none-minimal	mild	moderate	mod. severe	severe
No	44 (9.2)	71 (14.9)	59 (12.4)	37 (7.8)	29 (6.1)
Yes	44 (8.4)	65 (13.7)	74 (15.5)	35 (7.4)	22 (4.6)

$71 \div 476 = 14.9\%$: each percentage is of *all* 476 students, a total percentage. It cannot compare the two tuition groups.

Row percentages, ours

Moderate or worse: No 125 of 240 (52.1%); Yes 131 of 236 (55.5%).

Counts and printed percentages: Islam 2020, Table 2, PMID 32845928. Group sizes 240 and 236: Table 1. The grouping “moderate or worse” and its percentages are ours.



Three checks catch one misprint

The cell: tuition Yes, none—minimal, printed “44 (8.4)”

1. **The row.** $44 + 65 + 74 + 35 + 22 = 240$. Table 1 says 236 students do tuition.
2. **The column.** $44 + 44 = 88$ with none—minimal. Table 1 says 84.
3. **The percentage.** $44 \div 476 = 9.2\%$, not 8.4%. And 8.4% of 476 is 40.

All three say 40

With 40, the row makes 236, the column makes 84, and the percentage is 8.4. The count was mistyped; the percentage was right.

The conclusion hardly changes. The habit is the point.

Printed values: Islam 2020, Tables 1 and 2, PMID 32845928. All three checks and the corrected value 40 are ours.



What today established

1. Categories get counts and percentages; the question picks the denominator.
2. The mean uses every value and follows a long tail; the median resists it.
3. The SD is the typical distance from the mean; the IQR is the spread of the middle half. Report mean (SD) or median (IQR) by the shape.
4. A fence flags a value; only checking decides if it is wrong. Check ranges first.
5. The variables decide the chart. Bars start at zero, areas scale honestly, and every percentage carries its n .



Session 3: From sample to population

Probability, the normal curve, why sample means behave, and the confidence interval: how to say something about students we did not ask.

Before then

Read Chapter 2 of the class notes. Try the Part A exercises. Bring a calculator.

One question to think about

Our class averages 6.80 hours of sleep. If another class of 120 at this university filled in the survey, would they average exactly 6.80 too?



- ▶ 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.
- ▶ Anscombe FJ. Graphs in statistical analysis. *The American Statistician* 1973;27(1):17–21. doi:10.1080/00031305.1973.10478966.
- ▶ Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
- ▶ 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 is ours.
- ▶ Constructed examples, made up for teaching: the two classes of eight, the CGPA weighting, the pictogram, the 24-month series, and the “2 of 3” poll.

