



STATISTICS & DATA ANALYSIS FUNDAMENTALS • STAT.BD • SESSION 1

Data and Research Basics

From a question about your own life to a dataset you can analyse

Muhtasim Munif Fahim, MSc

Stat.BD

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

Session 1 of 14 • Theory

Fourteen sessions, one dataset, three tools

Session	What it does
1	Data and research basics
2	Describing data
3	From sample to population
4	Testing hypotheses, comparing groups
5	Relationships between variables
6	Regression and survey scales
7–8	SPSS, raw data to results
9–10	R, raw data to results
11–12	Python, raw data to results
13	Project lab
14	Capstone, and choosing your tool

The one idea behind the course

Learn the statistics once. Then watch the same analysis run in three tools, on a dataset this room creates today.

What you need

Nothing. No statistics, no programming, any department, any year.

Course syllabus, Stat.BD. The certificate needs 11 of the 14 sessions and the capstone.



The test of this session

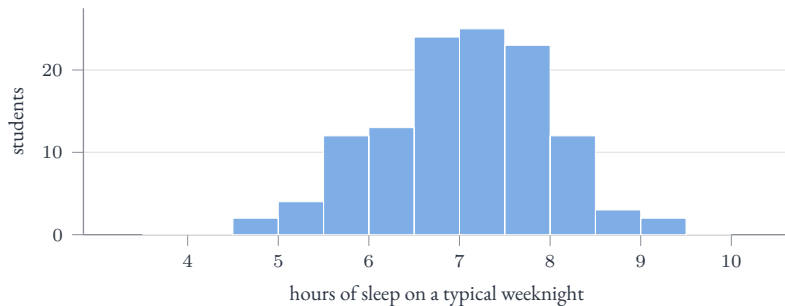
Take any question about people, and say who it is about, what must be measured to answer it, and how those measurements become a dataset.

- ▶ Write a research question and list the variables it needs
- ▶ Tell a population from a sample, and name four ways to draw one, each with a weakness
- ▶ Classify any variable by type and by scale of measurement
- ▶ Turn questionnaire answers into numbers with a codebook

And, before we leave, fill in the anonymous survey that becomes the course's data.



If everyone slept seven hours, there would be no statistics



People differ. Statistics is the discipline of saying useful things about groups of people who differ.

Simulated: a stand-in for a class of 120, generated from a fixed seed (data/class-data.R). Not real students. The real survey is filled in today.



Act I

What statistics is for



Stat.BD

Statistics does three jobs: describe, estimate, test

Describe

Summarise what was measured.

The students who answered slept 6.80 hours on average.

Estimate

Say something about people who were not measured, with a margin.

Students like these probably average between 6.63 and 6.97 hours.

Test

Ask whether a pattern is bigger than chance would produce.

Is the average below the recommended seven hours, or could that be luck?

Session 2 is describing. Session 3 is estimating. Sessions 4 to 6 are testing and modelling.

Hours and interval: **simulated** class data. The interval is ours, computed as Session 3 will teach.



Every faculty in this room already has a statistics problem

Faculty	A question that needs data to answer
Arts and Humanities	Do newspaper editorials use more English loanwords now than twenty years ago?
Business Studies	Do customers who get a follow-up call spend more in the next month?
Science and Engineering	Does a new fertiliser raise rice yield on the university's test plots?
Social Sciences and Law	Do first-generation students report more exam stress than their classmates?

What they share

Each question is about a group, each needs measurements on many members of it, and each answer will be uncertain because the members differ.



An anecdote is one row; data is the whole column

The anecdote

“My cousin sleeps four hours a night and has a CGPA of 3.9. Sleep doesn’t matter.”

One person, chosen because the story was memorable. Nothing to compare with.

The cousin is real. The problem is that you heard about the cousin *because* the cousin is unusual.

The data

Sleep and CGPA recorded for every student in a defined group, including the dull cases nobody would tell a story about.

Now a pattern can be seen, or seen to be absent.



One Bangladeshi student survey, start to finish

Question	How common are depression and anxiety symptoms among university students during COVID-19?
Who	476 university students in Bangladesh
When	6 to 12 May 2020
How found	A Google Form shared on Facebook
Measured	PHQ-9 (depression) and GAD-7 (anxiety)

82.4%

reported mild to severe depressive symptoms, 392 of 476

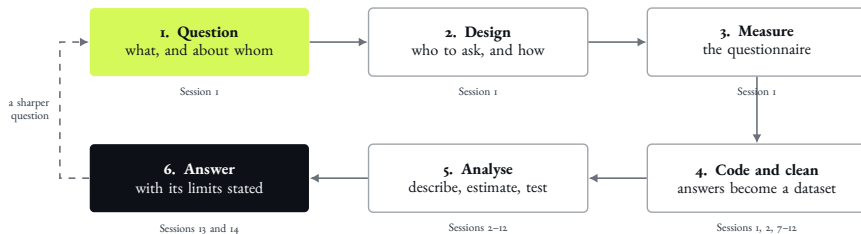
Why this paper

It is open access, it is about students like you, and it will come back in every theory session.

Islam MA et al. *PLoS One* 2020;15(8):e0238162. PMID 32845928. CC BY.



The research cycle runs from a question to a better question



Today covers the first four boxes. The rest of the course is box five, and the capstone is box six.



Describing the people you asked is not the same as inferring about everyone

Descriptive statement

“392 of the 476 students who answered, 82.4%, had mild to severe depressive symptoms.”

True of those 476, if they answered honestly.

Inferential statement

“About four in five Bangladeshi university students had such symptoms in May 2020.”

A claim about thousands who were never asked.

The leap

The second sentence is only as good as the route from the 476 to everyone else. That route is the sampling method, and it is Act III.

Counts and percentage: Islam 2020, Results. PMID 32845928. The inferential sentence is ours, written to show the leap.



A number in a published paper can be wrong on the page

The paper prints

389 students “(87.7%)” had mild to severe anxiety symptoms.

The arithmetic

$389 \div 476 = 0.817$, which is 81.7%.

The check

Mild 185 + moderate 118 + severe 86 = 389. The count is right; the percentage is not.

A habit for the whole course

When a paper gives a count and a percentage, divide. It takes ten seconds, and it is how you find out whether the rest of the table deserves your trust.

Printed figures: Islam 2020, Results and Table 1, PMID 32845928. The division and the category sum are ours.



Act II

From a question to variables



A research question you can answer has four parts

W

Who: the people the question is about, precisely enough to decide who counts

X

What: the thing that might matter, such as screen time

C

Compared with what: more against less, one group against another

M

Measured how: the outcome, in units someone else could record the same way

A question missing any one of the four cannot be answered with data.



Turn a worry into a question

The worry

“Phones are ruining our sleep.”

Who? Everyone? Ruining how? Compared with what?

The question

Among undergraduate and Masters students at the University of Rajshahi, do students with more daily screen time sleep fewer hours on a weeknight?

Who: RU students. **What:** daily screen time.

Compared: more against less. **Measured:** self-reported weeknight sleep, in hours.

This is the question our class survey is built to answer. We come back to it in Sessions 5 and 6.



Three kinds of question: describe, compare, relate

Kind	From our class survey	What it needs
Describe	What share of students do private tuition or a part-time job?	one variable
Compare	Do hall, mess and family-home students sleep the same amount?	a group variable and an outcome
Relate	Is more screen time linked to less sleep?	two measured variables

Why the kind matters

The kind of question decides the kind of analysis. Describe is Session 2. Compare is Session 4. Relate is Sessions 5 and 6.



Every question becomes a list of variables

The question

Do students with more screen time sleep fewer hours on a weeknight?

The variables it needs

- ▶ weeknight sleep, in hours
- ▶ daily screen time, in hours
- ▶ things that might affect both: tuition or a job, where you live, year

A variable

A characteristic that can differ from one person to the next. If it cannot differ, it is a constant, and it tells you nothing.



A dataset is a rectangle: one row per student, one column per variable

id	fac	year	tui	res	study	sl_wd	sl_we	screen	cgpa	s1	s2	s3	s4	s5
1	1	2	1	2	2.5	7.0	8.5	3.1	4	3	3	4	3	3
2	1	3	1	3	2.0	7.5	9.0	5.5	2	4	3	4	4	4
3	1	3	1	2	3.0	6.5	7.5	1.8	3	3	4	2	1	4
... and 120 rows in all														

A row

One student: every answer that person gave, in one line.

A column

One variable: everybody's answer to one question, top to bottom.

The first three rows of the **simulated** class data, as coded. The codes are explained in Act V.



Word	Meaning	In our survey
Unit	the thing each row is about	one student
Observation	all the values recorded for one unit	one row
Variable	a characteristic recorded for every unit	weeknight sleep
Value	one unit's entry for one variable	6.5 hours

The unit is not always a person

In a study of newspaper editorials the unit is an editorial. In a study of districts, the unit is a district, and a row holds the district's figures, never one resident's.



The same variable can be an outcome in one question and an explanation in another

Question A

Does screen time go with less sleep?

- explanatory: screen time
- outcome: **sleep**

Question B

Does sleep go with study stress?

- explanatory: **sleep**
- outcome: stress score

Roles belong to questions

The outcome is what the question is trying to explain. The explanatory variable is what might explain it. The data do not decide which is which; the question does.

Other names you will meet: dependent and independent; response and predictor; Y and X .



A construct is not the same thing as its measurement

Construct	How our survey measures it
Sleep	the hours a student <i>says</i> they sleep on a typical weeknight
Screen use	the hours the phone's own screen-time setting reports
Study stress	the total of five agree/disagree statements
Academic success	a CGPA band

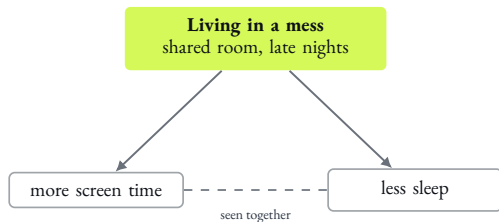
The gap

Every finding is about the measurement. Whether it is also about the construct depends on how good the measurement is.

Islam 2020 did not measure depression. It measured PHQ-9 scores, a screening questionnaire. The authors note they could not reach anyone with medically examined depression.



A third variable can sit behind any link



This is why our survey asks where you live and whether you tutor. Session 5 names it confounding; Session 6 adjusts for it.

Why this matters on day one

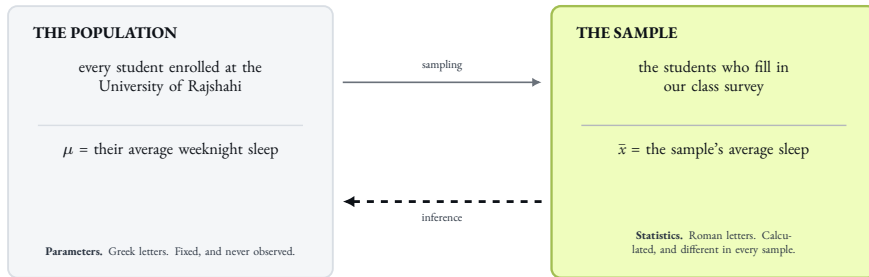
If you do not *measure* the third variable, no analysis can ever account for it. The time to think of it is when writing the questionnaire.



Act III

Who the question is about, and who was asked

The population is who the question is about; the sample is who was measured



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.



A census measures everyone; a good sample is usually the better choice

Census: every member of the population.

Bangladesh's population and housing census tries to count every household.

Sample: a subset, chosen by a method.

A survey of a few thousand households chosen at random.

A soup is tasted with one spoon. What matters is that it was stirred first.

Why sample at all

Cheaper and faster, and often more accurate: a small survey can afford trained interviewers and careful checking that a census of millions cannot.



The sampling frame is the list you actually draw from

Population	every enrolled RU student
Frame	the registrar's enrolment list
Sample	the students drawn from that list

Who is on the population but not the frame?
Students admitted late, students whose records are wrong, students on leave.

The frame decides the answer

A sample can only represent the frame it came from. A survey drawn from a Facebook group can only represent that Facebook group.



Four probability sampling methods, drawn on the same forty-eight students

Simple random



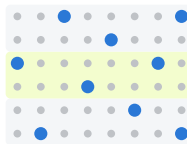
Every student equally likely.
Needs a complete list.

Systematic



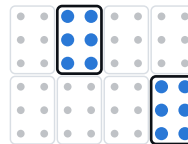
Every 6th name on the list.
Fails if the list has a rhythm.

Stratified



A random draw inside each group.
Needs the groups in advance.

Cluster



Whole groups taken.
Cheaper, and less precise.

In all four, chance decides who is chosen, and every student's chance is known. That is what makes a sample a **probability sample**.

Dots and picks constructed for the picture.



Simple random and systematic sampling, and where each breaks

Simple random

Number every student on the list. Let a computer draw, say, 200 numbers.

Its weakness

It needs a complete list. And by chance it can under-represent a small group, such as Masters students.

Systematic

Pick a random start between 1 and k , then take every k -th name.

Its weakness

If the list has a rhythm, the sample catches it. A list ordered by room, where every tenth room is a single room, gives a sample of single rooms.



Stratified and cluster sampling, and what each costs

Stratified

Split by something you know in advance, such as faculty. Draw at random inside each stratum.

Its cost

You must know every student's stratum before drawing. And if a small stratum is deliberately over-sampled, the analysis must weight it back down.

Cluster

Split into natural groups, such as halls. Pick some halls at random and survey everyone in them.

Its cost

Students in the same hall resemble each other, so 300 students from 3 halls carry less information than 300 from the whole campus.



Non-probability sampling is common, and every kind has a blind spot

Method	How people are found	Who it misses
Convenience	whoever is at hand: a class, a canteen	anyone not there that day
Snowball	participants pass the survey to friends	anyone outside those networks
Quota	fill fixed numbers per group, by convenience	the hard-to-reach inside each group
Purposive	hand-picked for a reason, such as expertise	fine for depth; not for a percentage

The Islam survey

A Google Form shared through Facebook and Messenger: a **snowball** sample. The authors chose it over a random sample because resources were limited and the outbreak would not wait, and they say so in their limitations.

Islam 2020, Methods and Discussion. PMID 32845928.



A very large sample predicted the wrong winner

1936, United States. The *Literary Digest* magazine mailed millions of straw ballots and predicted that Alf Landon would beat Franklin Roosevelt.

Roosevelt won 523 electoral votes to Landon's 8.

A later analysis concluded that *both* the sample and the response were biased, and that together they produced the wrong answer.

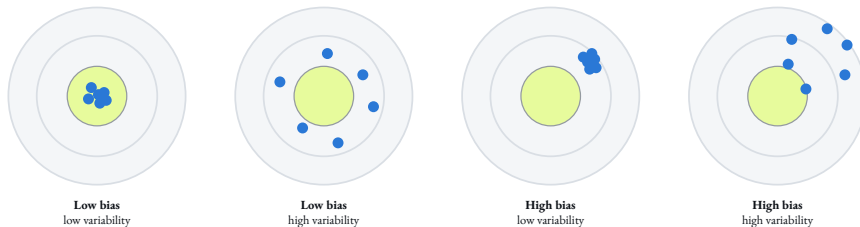
The lesson

Size cannot fix a biased method. A huge sample drawn the wrong way gives the wrong answer, precisely.

Squire P. Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly* 1988;52(1):125–133. doi:10.1086/269085. Electoral votes: US National Archives. Ballot counts are not quoted here: we could not verify them against a primary source.



Bias and sampling error are different failures



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.

Sampling error: samples scatter around the truth.
A bigger sample shrinks the scatter.

Bias: samples miss the truth in the same direction.
A bigger sample changes nothing.



The people who do not answer are rarely a random subset

A perfect random sample of 500 students.
250 fill in the form.

Who answers a survey about study stress? Perhaps the stressed, who care about the topic. Perhaps the calm, who have time. Either way, the 250 differ from the 250 who did not answer, and not at random.

Non-response bias

A random draw followed by a selective response is no longer a random sample.
Always report the response rate, and ask who is missing.



What our class survey can and cannot tell us

It can

Describe the students in this course, exactly.

Show how the methods work, on data that mean something to you.

Suggest patterns worth studying properly.

It cannot

Describe RU students in general. Everyone here chose to join a statistics course run by Stat.BD: a convenience sample, with a selection we cannot measure.

Every result from our class data in this course is a result about this room. We will say so, every time.



Two Bangladeshi surveys, two ways of finding people

	Islam 2020	Khanam 2019
Who	university students	adults aged 35 and over, two rural subdistricts of Sylhet
Frame	none: a link shared online	the study site's own 2016 census list
Selection	snowball	computer-generated random numbers
Sample	476	1,810
Can it estimate for everyone?	only with strong assumptions	yes, for the two subdistricts, with a margin of error

Neither is “good” or “bad”. Each answers the question its method allows.

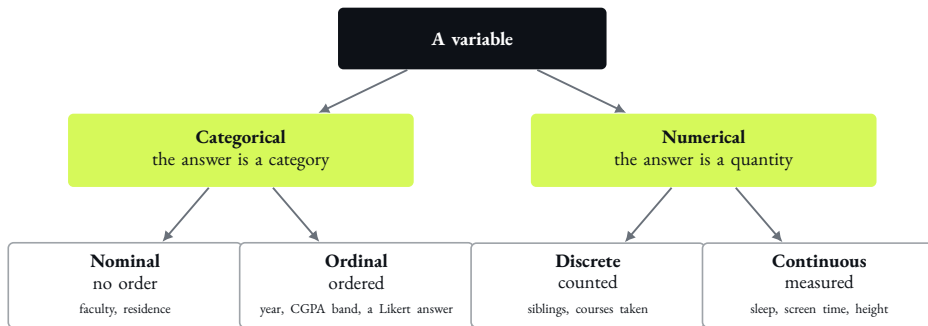
Islam 2020, PMID 32845928. Khanam R et al. *BMJ Open* 2019;9(10):e026722, PMID 31662350.



Act IV

What kind of variable is it?

Categorical or numerical: the first question to ask of any variable



Ask it of every column before doing anything else. The answer decides which summaries, which charts and which tests are allowed.



Four scales of measurement, each adding one property

Each step keeps everything below it and adds one property. Climbing is what allows more arithmetic.

Nominal
names only

faculty, residence,
tuition yes or no

Ordinal
+ order

year of study,
CGPA band, a Likert answer

Interval
+ equal gaps

temperature in °C,
calendar year

Ratio
+ a true zero

hours of sleep,
screen time, age

Nominal, ordinal, interval, ratio: **NOIR**. Stevens proposed the four in 1946, and they are still the vocabulary of every software package.

Stevens SS. On the theory of scales of measurement. *Science* 1946;103(2684):677–680. doi:10.1126/science.103.2684.677.



Three questions sort any variable onto its scale

Variable	Order?	Equal gaps?	True zero?	Scale
Faculty group	no	no	no	nominal
Tuition or job, yes or no	no	no	no	nominal
Year of study	yes	no	no	ordinal
CGPA band	yes	no	no	ordinal
Temperature in °C	yes	yes	no	interval
Hours of sleep	yes	yes	yes	ratio
Screen time	yes	yes	yes	ratio

Stop at the first “no”. The scale is the last “yes”.



Discrete values are counted; continuous values are measured

Discrete

Whole numbers from counting: number of siblings, courses failed, times you checked your phone before breakfast.

Continuous

Any value in a range, limited only by the ruler: sleep, height, screen time.

A continuous variable is always recorded discretely

Our survey records sleep to the nearest half hour. The variable is still continuous; the recording is coarse.



The tricky cases: CGPA band, year of study, a Likert answer

CGPA itself is numerical. **CGPA recorded as a band** (3.00–3.49) is ordinal: the bands have order, but a band is not a number.

Year of study is coded 1 to 5. The codes are ordered; the gaps are not equal.

A Likert answer (strongly disagree = 1 ... strongly agree = 5) is ordinal. **The total of five such answers** is often treated as numerical.

A number on a form is not a number

Codes are labels that happen to be digits. Averaging faculty codes gives a number that means nothing, and software will compute it without complaint.



Why 30°C is not twice as hot as 15°C

Zero on the Celsius scale is where water freezes, not “no heat”.

In Fahrenheit, the same two days are 86° and 59° : a ratio of 1.46, not 2.

A ratio that changes when the unit changes is not a property of the weather.

Interval against ratio

On an interval scale, differences mean something and ratios do not. On a ratio scale, both do: 8 hours of sleep is twice 4, in any unit.



Grouping a number into bands throws information away

Two students with CGPA 3.01 and 3.49 land in the same band.

Two with 2.99 and 3.01 land in different bands.

Record sleep as “under 6 hours / 6 or more” and a student sleeping 2 hours looks the same as one sleeping 5.5.

Our survey bands CGPA on purpose, for privacy, and records sleep and screen time as numbers on purpose, for analysis.

The rule

Record the finest detail you can ethically collect. You can always group later. You can never ungroup.



The scale decides the analysis

Scale	Summary	Chart	Typical test
Nominal	counts, percentages	bar chart	chi-square (Session 5)
Ordinal	median, percentages	bar chart	Mann–Whitney, Spearman (Sessions 4, 5)
Interval, ratio	mean and SD, or median and IQR	histogram, box plot	<i>t</i> -test, ANOVA, correlation, regression (Sessions 4 to 6)

Why today matters for the whole course

Classify the variable wrongly and every later step is wrong: the summary, the chart and the test.



Classify ten variables

1. Blood group
2. Monthly household income, in taka
3. Income band: low, middle, high
4. Number of courses taken this semester
5. Rank in a merit list
6. Mobile operator
7. Minutes walked to class
8. "I enjoy statistics", agree 1 to 5
9. Calendar year of admission
10. Distance from home district, in km

For each

Categorical or numerical? Then nominal, ordinal, interval or ratio?



Act V

From a questionnaire to a dataset



A questionnaire is a measuring instrument, and it can be badly made

A thermometer measures temperature.

A questionnaire measures what people know, do, think and feel.

Both can be badly calibrated. A thermometer that reads two degrees high is biased. A question that nudges people towards one answer is biased in exactly the same way.

No analysis can repair a bad question

The most careful statistics on a leading question give a precise estimate of the wrong thing.



Five ways to write a bad question

Fault	Example	Fix
Double-barrelled	Are the library and the canteen satisfactory?	ask two questions
Leading	Don't you agree phones harm students?	neutral wording
Vague	Do you study regularly?	hours on a typical day
Loaded	Have you stopped wasting time on Facebook?	remove the assumption
Overlapping options	Sleep: 4–6, 6–8, 8–10 hours	4 to under 6, 6 to under 8

Every one of these has appeared in a real student thesis.



Answer options must cover everyone, and each person only once

Not exhaustive

Residence: hall / family home

Where does a student in a mess tick?

Not mutually exclusive

CGPA: 2.50–3.00 / 3.00–3.50

Where does exactly 3.00 go?

Closed questions give answers that are quick to code. Open questions give richer answers that must be coded by hand, into categories decided afterwards.

The two rules

Exhaustive: every respondent has a box.

Mutually exclusive: no respondent has two.

When in doubt, add “Other (please write)”. Code it later.



Likert items, and the item worded the other way

	SD	D	N	A	SA
I worry about my exam results.	1	2	3	4	5
I feel confident I can handle my coursework.	1	2	3	4	5

For the second item, **agree means less stress**. Before adding it to the others it must be **reverse-coded**: $1 \rightarrow 5$, $2 \rightarrow 4$, $3 \rightarrow 3$, $4 \rightarrow 2$, $5 \rightarrow 1$, or simply $6 - \text{answer}$.

Why word one item backwards

To catch people who tick “agree” all the way down without reading.

Forget to reverse it

The item then cancels the others, and the scale falls apart. Session 6 shows by how much.



Ask for consent, and do not collect what you do not need

- ▶ Tell people what the survey is for and who will see it
- ▶ Make taking part voluntary, and skipping any question allowed
- ▶ Collect no names, student IDs, phone numbers or emails unless the study needs them
- ▶ Report groups, never individuals

Our survey

Anonymous. No names, IDs or contact details. CGPA in bands. Any question may be left blank.

The Islam survey was anonymous too, and received ethical approval from Khulna University before collecting data.

Islam 2020, Methods. PMID 32845928.



Our class survey, question by question

-
- 1 Your department (write it)
 - 2 Year of study: 1st, 2nd, 3rd, 4th, Masters
 - 3 Do you currently do private tuition or a part-time job?
 - 4 Where do you live: hall, mess, family home?
 - 5 Hours of self-study on a typical class day
 - 6 Hours of sleep on a typical weeknight
 - 7 Hours of sleep on a typical weekend night
 - 8 Daily screen time, from your phone's own setting
 - 9 CGPA band, or "no result yet"
-

10–14. Strongly disagree (1) to strongly agree (5):

- s1 I feel overwhelmed by my academic workload.
 - s2 I worry about my exam results.
 - s3 I find it hard to relax because of my studies.
 - s4 I feel confident I can handle my coursework.
 - s5 Academic pressure affects my sleep.
-

No names. No IDs. Any question may be left blank.



A codebook makes sure everyone reads the numbers the same way

For every variable, a codebook records:

- ▶ its name in the dataset: `sleep_wd`
- ▶ the question it came from, word for word
- ▶ its type and scale
- ▶ every code and what it means
- ▶ units, and the allowed range
- ▶ how missing answers are coded

Why

Six months later, you will not remember whether residence code 2 was mess or family home. Neither will your supervisor, or the software.



Our codebook, first part

Name	From question	Scale	Codes and units
faculty	1, department	nominal	1 Arts & Humanities, 2 Business Studies, 3 Science & Engineering, 4 Social Sciences & Law
year	2	ordinal	1 to 4 = year; 5 = Masters
tuition	3	nominal	0 No, 1 Yes
residence	4	nominal	1 Hall, 2 Mess, 3 Family home
study_hrs	5	ratio	hours, to the nearest half; allowed 0 to 16
sleep_wd, sleep_we	6, 7	ratio	hours, to the nearest half; allowed 0 to 16
screen_hrs	8	ratio	hours, one decimal; 99 = don't know
cgpa_band	9	ordinal	1 below 2.50, 2 2.50–2.99, 3 3.00–3.49, 4 3.50–4.00; 9 = no result yet
s1–s5	10–14	ordinal	1 strongly disagree to 5 strongly agree; s4 reversed before use

Blank answers are coded 99 in every column. The faculty grouping is ours, a simplification for teaching.



“No result yet” and “left blank” are different kinds of missing

A first-semester student has no CGPA. That is **not applicable**: the question has no answer for them. Code 9.

A third-year who skips the CGPA question has a CGPA; we just do not know it. That is **missing**. Code 99.

In the simulated class data: 15 students with no result yet, and 3 who did not know their screen time.

Counts: **simulated** class data.

A missing code is not a number

Leave 99 in the column and the average screen time jumps. Every package must be told that 99 means “no value”.



From a paper form to a row of numbers

The form

Department: *Sociology*
Year: *3rd*
Tuition or a job: *Yes*
Living in: *a mess*
Weeknight sleep: *6.5 hours*
Screen time: *5 h 12 min*



The codebook says

faculty: Social Sciences = 4
year: 3rd = 3
tuition: Yes = 1
residence: Mess = 2
sleep_wd: hours, to the half
screen_hrs: hours, one decimal



id	faculty	year	tuition	residence	sleep_wd	screen_hrs
57	4	3	1	2	6.5	5.2

one row
of the dataset

Coding is where answers become data. It is also where most errors enter.

The student is constructed for the picture.



Check the data before trusting any analysis of it

- ▶ **Range:** is every value inside the codebook's allowed range?
- ▶ **Consistency:** does anyone report more weekend sleep than hours in a night?
- ▶ **Codes:** is every category code one the codebook defines?
- ▶ **Missing:** are 9 and 99 declared as missing, not counted as values?
- ▶ **Double entry:** for paper forms, type everything twice and compare

A preview of Session 2

One row in the simulated data has self-study of 25 hours a day. The form said 2.5. The typist missed the decimal point.

The 25-hour entry is planted in the **simulated** data on purpose, for Session 2.



Activity: fill in the class survey

1. Take one form, or open the link.
2. Answer honestly. Skip anything you prefer not to answer.
3. Check your phone's own screen-time setting for question 8.
4. Do not write your name anywhere.

About eight minutes.

What happens next

Your answers are coded with our codebook, become the class dataset, and are shared with everyone for Sessions 2 to 14.



Activity: build the codebook for what you just answered

In groups of four:

1. Take three of the questions you were given.
2. For each, write the variable name, type, scale, codes, units, allowed range and missing code.
3. Code two forms from the pile using only your codebook.
4. Swap forms with another group, code again, and compare.

Where groups disagree

Every disagreement is a hole in the codebook. Write down the rule that would have prevented it.



What today established

1. Statistics describes, estimates and tests, because people differ.
2. A question needs a who, a what, a comparison and a measurement.
3. A sample represents its population only as well as the method that chose it; size cannot fix bias.
4. Every variable is categorical or numerical, and sits on one of four scales. The scale decides the analysis.
5. A codebook turns answers into data that anyone can read the same way; missing and not applicable are coded differently.



Session 2: Describing data

Frequency tables, averages and spread, box plots and histograms, and the charts that mislead. On our class dataset, by hand.

Before then

Read Chapter 1 of the class notes. Try the Part A exercises. Bring a calculator.

One question to think about

One student in the data says they study 25 hours a day. What happens to the class average?



- ▶ 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.
- ▶ Khanam R, Ahmed S, Rahman S, et al. Prevalence and factors associated with hypertension among adults in rural Sylhet district of Bangladesh: a cross-sectional study. *BMJ Open* 2019;9(10):e026722. PMID 31662350. doi:10.1136/bmjopen-2018-026722.
- ▶ Squire P. Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly* 1988;52(1):125–133. doi:10.1086/269085.
- ▶ Stevens SS. On the theory of scales of measurement. *Science* 1946;103(2684):677–680. doi:10.1126/science.103.2684.677.
- ▶ US National Archives, Electoral College results, 1936: Roosevelt 523, Landon 8.
- ▶ 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.

