



CLASS NOTES • CHAPTER 1

Data and Research Basics

From a question about student life to a dataset anyone can analyse

Statistics & Data Analysis Fundamentals
Stat.BD

Mentor: Muhtasim Munif Fahim, MSc

CHAPTER 1

Data and Research Basics

Statistics begins well before any calculation. This chapter follows a question from a worry to a list of variables, from a population to a sample, and from a questionnaire to a rectangle of numbers.

Tasnim is a third-year sociology student at the University of Rajshahi. She is convinced that phones are ruining her classmates' sleep, and she wants to turn that conviction into a thesis. Before she can compute anything she has to settle four matters that have nothing to do with formulas. Whose sleep is she asking about? What exactly will she record, and in what units? How will she choose the students she asks, and what will those students let her say about the ones she did not ask? And how will ticks on a paper form become numbers that software can read? Each of these decisions is made before the first calculation, and each limits what the last calculation can mean.

This chapter takes the four matters in that order. It begins with what statistics is for: describing what was measured, estimating what was not, and testing whether a pattern is bigger than chance would produce. It then turns a vague worry into a question with four parts, and the question into a list of variables arranged as a dataset. The third part separates the population a question is about from the sample that was measured, and compares the ways a sample can be drawn. The fourth sorts variables by type and scale of measurement, because the scale decides every later summary, chart and test. The last part follows a questionnaire through coding into a dataset, with a codebook, missing values and the checks that catch typing errors.

Two published surveys carry the argument. Islam and colleagues (reference 1) surveyed 476 Bangladeshi university students online in May 2020 about depression and anxiety symptoms. Khanam and colleagues (reference 2) measured blood pressure in a random sample of adults in rural Sylhet. The first found its respondents through a link passed from friend to friend; the second drew them at random from a census list. Placing the two side by side is most of what this chapter has to say about sampling.

A third source is the course's own class survey of fourteen questions about study, sleep, screen time and stress. Until the real answers are coded, the figures in this chapter come from a **simulated** stand-in of 120 students generated by a program from a fixed seed. They are not real students, and every figure drawn from them is labelled as simulated where it appears.

What this chapter establishes

1. That statistics exists because people differ, and that it does three jobs: describing, estimating and testing.

2. How a descriptive statement about the people measured differs from an inferential statement about people who were not.
3. How to check a published percentage against its count.
4. How to write a research question with four parts, and turn it into a list of variables arranged as a dataset.
5. How a population, a sampling frame and a sample differ, and how parameters differ from statistics.
6. How the four probability sampling methods work, and what each costs.
7. Why a large sample cannot repair a biased method, and why non-response matters.
8. How to classify any variable as categorical or numerical, and place it on one of four scales of measurement.
9. How to write questions and answer options that produce analysable data, and how to reverse-code an item worded the other way.
10. How a codebook turns answers into data, why “not applicable” and “missing” are coded differently, and which checks to run before any analysis.

Notation

N	the number of units in the population, or on the sampling frame
n	the number of units in the sample
μ	a population mean: a parameter, fixed and unknown
$\bar{x}$	a sample mean: a statistic, computed from the sample
π	a population proportion
$\hat{p}$	a sample proportion, read “p-hat”
a	a count: the number of units in the sample with some characteristic
k	the sampling interval in a systematic sample
N_b, n_b	the population and sample sizes in stratum b
x, x'	an item’s answer, and the same answer after reverse coding
K	the number of answer options on an item
C, F	a temperature in degrees Celsius and in degrees Fahrenheit

Contents

What this chapter establishes	1
Notation	2
1 What statistics is for	5
1.1 People differ	5
1.2 Three jobs: describe, estimate, test	5
1.3 Description and inference	6
1.4 An anecdote is one row	7
1.5 Checking a published number	7
1.6 The research cycle	8
2 From a question to variables	9
2.1 The four parts of an answerable question	9
2.2 Describe, compare, relate	10
2.3 Variables, units and the dataset	10
2.4 Outcome and explanatory variables	11
2.5 A construct and its measurement	11
2.6 The variable nobody measured	12
3 Population and sample	12
3.1 Parameters and statistics	12
3.2 Census or sample	13
3.3 The sampling frame	14
3.4 Four probability sampling methods	14
3.5 Non-probability sampling	16
3.6 Bias and sampling error	17
3.7 Non-response	18
3.8 What the class survey can and cannot say	19
4 Types of variable and scales of measurement	20
4.1 Categorical or numerical	20
4.2 Four scales of measurement	20
4.3 Discrete and continuous	21

4.4	Cases that cause arguments	22
4.5	Interval against ratio	22
4.6	Bands lose information	23
4.7	The scale decides the analysis	23
5	From a questionnaire to a dataset	24
5.1	A questionnaire is a measuring instrument	24
5.2	Answer options	24
5.3	Likert items and the reversed item	25
5.4	Consent and the least data that will do	25
5.5	The class survey	26
5.6	The codebook	26
5.7	Two kinds of missing	27
5.8	From a form to a row	28
5.9	Checking the data	29
6	Chapter summary	29
7	Key terms	31
8	Exercises	33
9	Solutions	36
10	Our arithmetic and constructed numbers	39
11	References and further reading	39
12	For students in the Stat.BD course	40

1 What statistics is for

1.1 People differ

Ask a class of students how many hours they slept last night and the answers will not agree. Some slept five hours, some nine, most somewhere between. If every student gave the same answer, a single number would describe the whole class and there would be nothing left to analyse. Statistics exists because people differ: in sleep, in income, in blood pressure, in how they answer a questionnaire. The discipline is a set of methods for saying useful things about groups whose members are not alike. Figure 1 shows the pattern for the simulated class. Most students cluster around seven hours, a few sleep much less, and a few sleep much more. That spread is and every method in this course is a way of summarising it, reasoning about it, or explaining part of it.

Variation

the differences between members of a group on the same characteristic

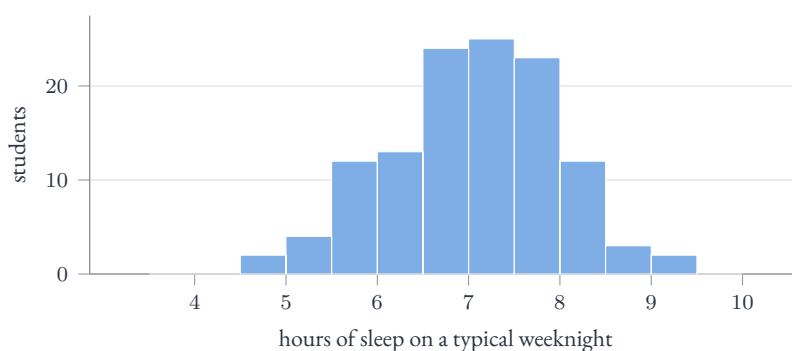


Figure 1: Weeknight sleep in the **simulated** class of 120 students. Each bar counts the students whose answer falls in a half-hour band. Generated by `data/class-data.R` from a fixed seed; not real students.

Intuition

If everyone slept seven hours, one number would say it all. Because people differ, describing a group takes at least two ideas: where the values sit, and how far apart they are. Chapter 2 makes both precise.

1.2 Three jobs: describe, estimate, test

Statistics does three jobs, and most confusion about a result comes from mistaking one for another. *Describing* summarises what was measured. The simulated class averages 6.80 hours of weeknight sleep; that is a fact about the 120 students who answered, and nothing more. *Estimating* reaches beyond the people measured to a larger group, and always carries a margin: students like these probably average between 6.63 and 6.97 hours. *Testing* asks whether a pattern in the data is larger than chance alone would produce: is the average below the seven hours often recommended, or could a gap that size appear by luck? The three jobs occupy the next five chapters: describing is Chapter 2, estimating is Chapter 3, and testing and modelling are Chapters 4 to 6.

Definition 1.1 • The three jobs of statistics

Description summarises the values that were measured, for the units that were measured. **Estimation** uses those values to state a likely range for a quantity in a larger group that was not measured in full. **Testing** judges whether a pattern in the measured values is larger than chance variation alone would be expected to produce.

The simplest description of a yes-or-no characteristic is a proportion: the count of units with the characteristic, divided by the number of units measured. Multiplied by 100 it is a percentage.

$$\hat{p} = \frac{a}{n}, \quad \text{percentage} = 100 \times \hat{p} \quad (1.1)$$

- $\hat{p}$ the sample proportion, a number between 0 and 1
 a the count of units in the sample with the characteristic
 n the number of units in the sample

Example 1.1 • Describing 476 students

Islam and colleagues (reference 1) asked 476 university students in Bangladesh to complete the PHQ-9, a nine-item screening questionnaire for depressive symptoms, between 6 and 12 May 2020. The paper's Table 1 reports how the 476 fell into five bands of PHQ-9 score:

PHQ-9 band	Students
None to minimal (below 5)	84
Mild (5 to 9)	136
Moderate (10 to 14)	133
Moderately severe (15 to 19)	72
Severe (20 or more)	51
Total	476

Mild to severe symptoms: $a = 136 + 133 + 72 + 51 = 392$. By equation (1.1), $\hat{p} = 392/476 = 0.824$, or 82.4%. The paper prints 392 and 82.4%; the sum is ours, and it agrees. This is a description: it says what was true of the 476 who answered.

1.3 Description and inference

The Islam figure can be written as two different sentences. “Of the 476 students who answered, 392 (82.4%) had mild to severe depressive symptoms” is a descriptive statement. It is true of those 476, provided they answered honestly, and it needs no further justification. “About four in five Bangladeshi university students had such symptoms in May 2020” is an inferential statement. It claims something about hundreds of thousands of students who were never asked. Newspaper headlines about surveys make the second kind of statement almost every time, and rarely announce the step from the first to the second. Whether the step is safe depends on one thing: how the 476 were found. Section 3 is about that route, and it is where most survey conclusions stand or fall.

Definition 1.2 • Descriptive and inferential statements

A **descriptive statement** is a claim about the units that were measured. An **inferential statement** is a claim about a larger group, the population, made from measurements on part of it. **Statistical inference** is the reasoning that justifies such a claim and states its uncertainty.

The leap is invisible in the sentence

Nothing in the wording of “four in five students” shows whether it rests on a random sample of the country’s students or on whoever clicked a link. The only way to judge the leap is to read the paper’s Methods section and ask how the people were chosen.

1.4 An anecdote is one row

“My cousin sleeps four hours a night and has a CGPA of 3.9, so sleep does not matter.” The cousin may well be real, and the story may be true. The fault lies in how the story reached the listener. Stories are retold because they are surprising, so the stories anyone hears are a biased selection of what happens: the cousin is mentioned *because* the cousin is unusual. There is also nothing to compare with. One person who sleeps little and does well says nothing about whether students who sleep little tend to do worse. Answering that needs sleep and CGPA recorded for every student in a group defined in advance, including the unremarkable ones nobody would tell a story about. That decision, to fix in advance who counts and to record all of them, is the first step from anecdote to data.

From anecdote to data

An anecdote is a single row chosen because it is memorable. Data are the whole column, for a group defined before anyone looked at the answers.

1.5 Checking a published number

A number in a peer-reviewed paper was typed by a person and can be wrong on the page. The quickest check a reader can make is also the commonest: when a paper prints a count beside a percentage, divide the count by the total and see whether the percentage follows. The check takes a few seconds, needs only a calculator, and tells the reader something about how carefully the rest of the table was assembled. It does not show that a paper is wrong in its conclusions; a slip in one cell is a slip in one cell. It shows only whether the number in front of the reader means what it says, which is the least a reader is entitled to know.

Example 1.2 • A percentage that does not follow from its count

The Results of Islam and colleagues state that 389 students had mild to severe anxiety symptoms on the GAD-7 and print the percentage as 87.7%.

Divide. $389/476 = 0.817$, which is 81.7%, not 87.7%.

Check the count. The paper’s Table 1 gives the GAD-7 bands as minimal 87, mild 185, moderate 118 and severe 86. Mild to severe is $185 + 118 + 86 = 389$, and all four bands

Example 1.2 • continued

sum to $87 + 389 = 476$. The count is right; the percentage beside it is not. Table 1 itself prints the four band percentages, and they add to $38.9 + 24.8 + 18.1 = 81.8$ for mild to severe, which agrees with 81.7 up to rounding.

The division, the sums and the conclusion are ours. The paper repeats 87.7% in its Discussion.

Not a sneer

The error is almost certainly a typing slip, and it survived peer review. The lesson is about readers, not authors: a reader who divides catches what reviewers missed. Chapter 2 returns to this paper and finds a second slip in a table.

1.6 The research cycle

A study moves through six stages, shown in Figure 2. It starts with a question about a defined group. It chooses a design: whom to ask and how to reach them. It measures, usually with a questionnaire or an instrument. It codes and cleans the answers into a dataset. It analyses the dataset by describing, estimating and testing. And it ends with an answer that states its own limits. The Islam survey runs the whole cycle: a question about student mental health during a lockdown; a design built on an online form; measurement with the PHQ-9 and the GAD-7; coding of ticked boxes into scores; frequency tables and a regression model; and a Discussion that lists what the study could not show. The arrow back to the start matters as much as the others. A good answer produces a sharper question.

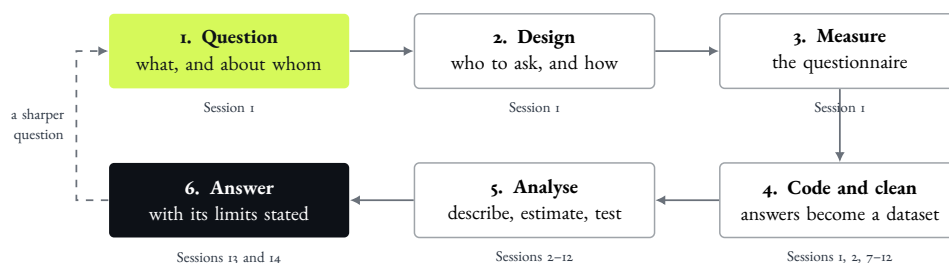


Figure 2: The research cycle, with the sessions of this course that teach each stage. This chapter covers the first three stages and the start of the fourth.

Intuition

Most failed student projects fail in the first three boxes, before any statistics is done. No analysis rescues a vague question, a sample that cannot represent anyone, or a questionnaire that measured the wrong thing.

2 From a question to variables

2.1 The four parts of an answerable question

“Is social media bad for students?” sounds like a research question and is not one. It does not say which students, so nobody can decide who belongs in the study. It does not say which platform or how much use counts, so the thing that might matter is undefined. It does not say compared with what: no use, less use, or a different platform. And “bad” names no outcome at all: bad for grades, for sleep, for mood, and recorded how? An answerable question fixes four things, shown in Figure 3. Health research calls the same structure PICO: population, intervention, comparison, outcome. The acronym is optional; the four ideas are not. Filling them in is what turns an argument into something data can settle.

W	Who: the people the question is about, precisely enough to decide who counts
X	What: the thing that might matter, such as daily screen time
C	Compared with what: more against less, or one group against another
M	Measured how: the outcome, in units someone else could record the same way

Figure 3: The four parts of an answerable research question.

Definition 2.1 • An answerable research question

A research question is answerable with data when it names (1) the population it is about, precisely enough to decide who belongs; (2) the characteristic or exposure of interest; (3) the comparison to be made; and (4) the outcome, with a method of measurement that another researcher could repeat.

Example 2.1 • From a worry to a question

Tasnim’s starting point is a worry: “Phones are ruining our sleep.” It fails all four parts. Rewritten:

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

Who: undergraduate and Masters students at one university. **What:** daily screen time. **Compared with what:** more screen time against less. **Measured how:** self-reported weeknight sleep, in hours.

The rewritten question does not claim that phones *cause* poor sleep. It asks whether the two go together, which is a smaller and answerable question. Whether one causes the other is harder, and Chapter 5 takes it apart.

2.2 Describe, compare, relate

Most questions a student will ask of a dataset are of three kinds, and naming the kind early saves weeks. A *describing* question asks about one variable: what share of students do private tuition? A *comparing* question asks whether an outcome differs between groups: do students living in halls, in messes and at home sleep the same amount? A *relating* question asks whether two measured quantities move together: is more screen time linked to less sleep? Table 1 lists what each needs. The distinction is practical. The kind of question decides the kind of analysis, and the flowchart at the end of Chapter 5, which chooses a test, begins from exactly this distinction.

Table 1: Three kinds of question, each asked of the class survey.

Kind	Example	Needs	Chapter
Describe	What share of students do private tuition or a part-time job?	one variable	2, 3
Compare	Do hall, mess and family-home students sleep the same amount?	a group and an outcome	4
Relate	Is more screen time linked to less sleep?	two measured variables	5, 6

2.3 Variables, units and the dataset

Every question turns into a list of variables. Tasnim's question needs weeknight sleep and daily screen time. It also needs the characteristics that might affect both: whether a student does tuition or a job, where the student lives, which year the student is in. A characteristic that cannot differ within a study is a constant, and a constant explains nothing. If every student in a study is a first-year, year of study is a constant there. Once the variables are listed and measured, they are stored in one shape, the same in every software package this course uses: a rectangle with one row for each unit and one column for each variable. Table 2 shows the first three rows of the simulated class data in that shape.

Variable

a characteristic that can differ from one unit to the next

Table 2: The first three rows of the **simulated** class dataset, as coded. The codes are explained by the codebook in section 5.6.

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 holds everything one student answered, in one line. A column holds everyone's answer to one question, top to bottom. The vocabulary varies between books and programs: a row is also called a case, a record or an observation; a column is a variable or a field; the whole rectangle is a dataset or a data frame.



Definition 2.2 • Unit, observation, variable, value

The **unit** is the thing each row describes. An **observation** is the set of values recorded for one unit, one row. A **variable** is a characteristic recorded for every unit, one column. A **value** is one unit's entry for one variable, one cell. A **dataset** arranges the observations as rows and the variables as columns.

The unit is not always a person

In a study of newspaper editorials the unit is an editorial. In a study of test plots the unit is a plot. In a study of districts the unit is a district, and a row holds the district's figures, never one resident's. Reading district-level data as though it described individual residents is an error with its own name, the *ecological fallacy*, which Chapter 5 illustrates.

2.4 Outcome and explanatory variables

In “does screen time go with less sleep?”, sleep is the outcome and screen time is the explanatory variable. In “does sleep go with study stress?”, sleep has changed sides: it is now the explanatory variable and the stress score is the outcome. The lesson is that a role belongs to a question, not to a variable, and the data do not decide it. The researcher does, by writing the question. Textbooks and software use several names for the two roles: dependent and independent, response and predictor, Y and X . The pair “dependent and independent” misleads beginners, who go looking for a variable that is independent of everything. “Outcome” and “explanatory”, or “response” and “predictor”, describe the roles more honestly.

Outcome

the variable a question is trying to explain; also response, dependent variable, Y

Intuition

The outcome is what the question wants to explain. The explanatory variable is what might explain it. Change the question and the roles can swap.

2.5 A construct and its measurement

A construct is the idea a study cares about: sleep, stress, academic success, depression. A measurement is what was actually written down. They are never quite the same thing. Self-reported sleep is not sleep, because people misremember and round. A CGPA band is not academic success; it is one institution's grading of one kind of work, coarsened into four groups. Table 3 lists the gap for each construct in the class survey. Every finding is strictly a finding about the measurement. Whether it is also a finding about the construct depends on how well the measurement captures it, a question Chapter 6 takes up for the class survey's five stress statements.

Construct

the idea a study cares about, as distinct from what it actually records

ISLAM 2020: A SCREENING SCORE IS NOT A DIAGNOSIS

The PHQ-9 and the GAD-7 are screening questionnaires. The paper describes the PHQ-9 as a screening tool and, in its limitations, notes that the method could not reach people with medically examined depression and anxiety. A headline reading “82% of students are depressed” has quietly replaced the measurement, a score above a cut-off, with the construct, a clinical condition.

Table 3: Four constructs and how the class survey measures them.

Construct	What the survey records
Sleep	the hours a student <i>says</i> they sleep on a typical weeknight
Screen use	the hours reported by the phone's own screen-time setting
Study stress	the total of five agree-or-disagree statements
Academic success	a CGPA band, one of four

2.6 The variable nobody measured

Suppose students with more screen time do sleep less. One explanation is that screens keep people awake. Another is that something else drives both. A student living in a crowded mess may spend more evenings on a phone and also sleep less, because of roommates and noise, with no direct effect of the phone on sleep at all. Figure 4 draws the second explanation. Chapter 5 names this problem confounding and Chapter 6 shows how a regression model adjusts for it. The practical lesson belongs here, at the start, because it is about design rather than analysis. An analysis can only allow for a variable that was measured. A variable that nobody thought to ask about is lost for good, and the time to think of it is while the questionnaire is still being written.

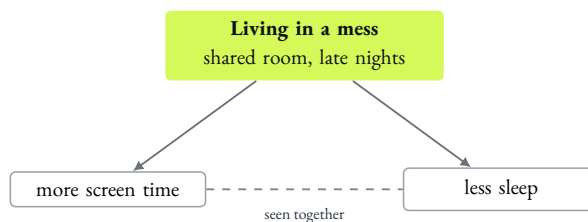


Figure 4: A third variable can produce a link between two others. Constructed to illustrate the idea; not a finding.

Measure it or lose it

The class survey asks where each student lives and whether each does tuition, although no question in this chapter is about either. They are there so that later analyses can allow for them.

3 Population and sample

3.1 Parameters and statistics

The population is the whole group a question is about: for Tasnim's question, every student enrolled at the University of Rajshahi. A number that describes a population, such as the average weeknight sleep of every enrolled student, is a *parameter*. It is fixed, and nobody will ever know it exactly, because nobody will ever measure every student on the same night without error. The sample is the part of the population that was actually measured. A number computed from the sample, such as the average sleep of the students who answered, is

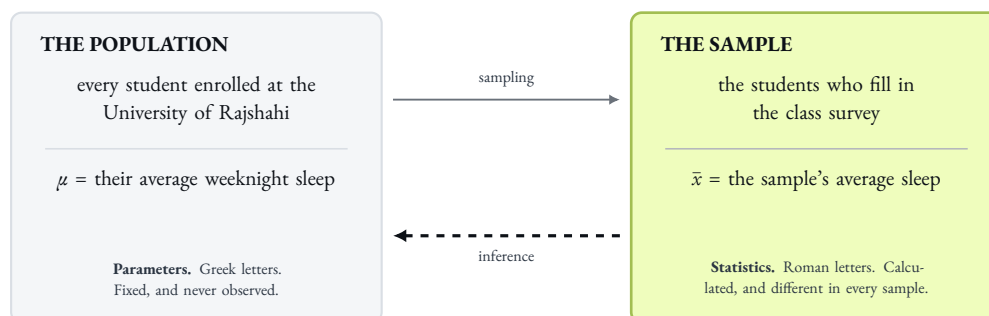
Population

the complete group a question is about

Sample

the units actually measured

a *statistic*. It can be computed exactly, and it would come out differently if a different set of students had answered. Figure 5 shows the two sides. Parameters are written with Greek letters and statistics with Roman ones: μ for the population mean, $\bar{x}$ for the sample mean.



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 5: Sampling runs from the population to the sample. Inference runs back, and is the only one of the two arrows that can go wrong.

Definition 3.1 • Population, sample, parameter, statistic

The **population** is the complete set of units a question is about. A **sample** is a subset of the population on which measurements are made. A **parameter** is a numerical characteristic of the population. A **statistic** is a numerical characteristic of a sample, computed from the data. The **sampling fraction** is the share of the population that was sampled:

$$f = \frac{n}{N} \quad (3.1)$$

- f the sampling fraction
- n the number of units in the sample
- N the number of units in the population

Intuition

The parameter is what the question wants. The statistic is what the data give. Chapters 3 to 6 are about the distance between them, and how to state it honestly.

3.2 Census or sample

A *census* measures every member of the population. Bangladesh's population and housing census attempts to count every household in the country. A sample measures some of them, chosen by a method. It seems natural that measuring everyone must be better, and often it is not. A census of millions must be done quickly by many fieldworkers with little training and little checking, and its errors are errors of execution: missed households, misheard answers, forms filled in carelessly. Those errors do not shrink as the count grows. A survey of a few thousand households can afford trained interviewers, repeat visits and careful checking. It is cheaper, faster, and frequently more accurate. The one condition is that the sample must be chosen so that it resembles the population, which is what the rest of this section is about.

The soup

A cook tastes a pot of soup with one spoon, and one spoonful is enough, provided the pot was stirred first. Random selection is the stirring.

3.3 The sampling frame

Between the population and the sample stands a list, and the list is never perfect. The sampling frame for a survey of University of Rajshahi students might be the registrar's enrolment list. Some students in the population are missing from it: those admitted late, those whose records are wrong, those on leave. Some entries on the list are not in the population: students who have left but were never removed. A sample can only represent the frame it was drawn from. A survey drawn from the members of a Facebook group can represent that Facebook group and nothing wider, however large it is. Good studies name their frame and say who it leaves out; the gap between population and frame is where many surveys fail before a single question is asked.

Sampling frame

the list or procedure from which a sample is actually drawn

Definition 3.2 • Sampling frame

The **sampling frame** is the list, map or procedure from which the sample is drawn. **Undercoverage** occurs when members of the population are absent from the frame; **overcoverage** when the frame includes units outside the population.

Example 3.1 • A named frame

Khanam and colleagues (reference 2) studied high blood pressure among adults aged 35 and over in Zakiganj and Kanaighat, two rural subdistricts of Sylhet. Their frame was the study site's own 2016 census database, and they selected participants with computer-generated random numbers. Fieldwork ran from August 2017 to January 2018.

Who the frame misses. The census list was not updated between censuses, and being on it was a condition of taking part, so an adult who moved into the area after 2016 could not be selected. The paper records some of the frame's other gaps as losses: people who had migrated out or died since the census.

Why it matters. Because the frame is named, a reader can judge exactly whom the results describe: adults on a 2016 list in two subdistricts, not rural Bangladesh as a whole.

3.4 Four probability sampling methods

A probability sample is one in which chance, not the researcher, decides who is chosen, and every unit's chance of selection is known in advance. That property is what later allows a margin of error to be calculated. Figure 6 shows the four standard methods on the same population of forty-eight. In a *simple random sample*, every unit is equally likely, as if names were drawn from a hat. In a *systematic sample*, the researcher picks a random starting point and then takes every k -th name on the list. In a *stratified sample*, the population is first split into groups known in advance, called strata, and a random sample is drawn inside each. In a *cluster sample*, the population is split into natural groups such as halls or tutorial groups, some groups are chosen at random, and everyone in the chosen groups is measured.

Probability sample

a sample in which chance decides who is chosen, and every unit's chance is known

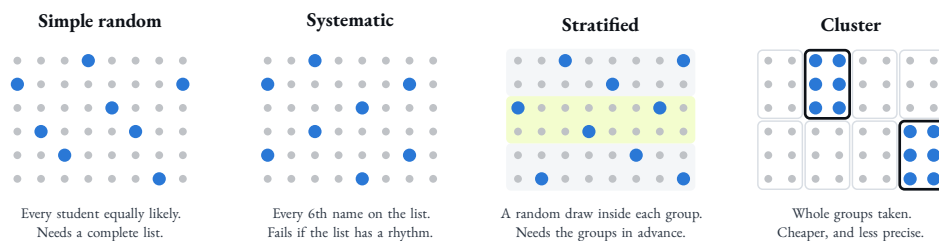


Figure 6: Four probability sampling methods on the same forty-eight units. Blue dots are the units chosen. Dots and picks constructed for the picture.

Definition 3.3 • Probability sample

A sample is a **probability sample** when every unit on the frame has a known, non-zero chance of selection and the selection is made by a random mechanism. In a **simple random sample** every possible set of n units is equally likely to be chosen.

Simple random sampling is the benchmark against which the others are judged. Its cost is the list: every unit must be numbered before the draw. Its quiet risk is bad luck with small groups. If one student in twenty is a Masters student, a random sample of forty can easily contain none.

Systematic sampling is the practical cousin of simple random sampling, easy to carry out by hand from a printed list. The interval k is the frame size divided by the sample size.

$$k = \frac{N}{n} \quad (3.2)$$

k the sampling interval: take every k -th unit, rounding k down to a whole number if necessary

N the number of units on the frame

n the number of units wanted in the sample

Example 3.2 • A systematic sample from a department list

A department list holds $N = 600$ students and a survey wants $n = 120$. By equation (3.2), $k = 600/120 = 5$. A random number between 1 and 5 is drawn for the start; suppose it is 3. The sample is the students in positions 3, 8, 13, 18, and so on up to 598: that is 120 students.

Where it breaks. Suppose the list is ordered by hall room, rooms come in blocks of five, and the fifth room of every block is a single room. Every selected position then sits at the same place in its block. A start of 3 picks no single rooms at all; a start of 5 picks nothing else. A list with a rhythm that matches k defeats the method. The department and its numbers are constructed for teaching.

Stratified sampling splits the frame by something known in advance, such as faculty or year, and draws at random inside each stratum. Every stratum is then certain to appear in the sample, and the result is usually more precise than a simple random sample of the same size.



The commonest rule for sharing out the sample is proportional allocation: each stratum gets the same share of the sample as it has of the population.

$$n_b = n \times \frac{N_b}{N} \quad (3.3)$$

n_b the number to sample from stratum b

n the total sample size

N_b the number of units in stratum b

N the number of units in the whole population

Example 3.3 • Proportional allocation by year

The same list of 600 has 160 first-years, 150 second-years, 140 third-years, 100 fourth-years and 50 Masters students. For a sample of 120, $n/N = 120/600 = 0.2$, so each year contributes a fifth of its students:

	1st	2nd	3rd	4th	Masters	Total
N_b	160	150	140	100	50	600
n_b	32	30	28	20	10	120

The Masters students, who might have been missed by bad luck in a simple random sample, are guaranteed ten places. If a study wanted more Masters students than their share, say 30, it could take them, but its analysis would then have to weight them back down to their true share. Numbers constructed for teaching.

Cluster sampling runs the other way. It is chosen for cost, not precision. Visiting three halls and surveying everyone in them is far cheaper than finding three hundred students scattered across the campus. The price is that students in the same hall share a canteen, a curfew and a Wi-Fi network, so their answers partly repeat each other. Three hundred students from three halls carry less information than three hundred from the whole campus. The loss has a name, the *design effect*, and large national surveys usually combine clusters and strata in several stages to balance cost against precision.

Every method has a cost

Simple random needs a complete list. Systematic fails on a list with a rhythm. Stratified needs every unit's stratum in advance, and weighting if strata are over-sampled. Cluster is cheaper and less precise. There is no method without a weakness, only a weakness the researcher has chosen and stated.

3.5 Non-probability sampling

Most student theses do not use a probability sample, and there is nothing shameful in that. The fault lies in not saying so, or in writing the results as though they described everyone. Table 4 lists the four common non-probability methods and the part of the population each cannot see. A *convenience* sample takes whoever is at hand. A *snowball* sample asks each

participant to pass the survey on. A *quota* sample fills a fixed number per group, by convenience within each group. A *purposive* sample hand-picks people for a reason, such as expertise. Purposive sampling is not a poor method; it is designed for a different job. Interviewing ten heads of department about a policy is purposive, and nobody would use it to estimate a percentage of all staff.

Table 4: Four non-probability methods and who each misses.

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	fixed numbers per group, filled by convenience	the hard-to-reach inside each group
Purposive	hand-picked for a reason	suited to depth, not to a percentage

ISLAM 2020: A SNOWBALL SAMPLE, STATED AS SUCH

The paper names its method as snowball sampling. An e-questionnaire built in Google Forms was shared, and participants were asked to pass it to friends through their personal and institutional Facebook and Messenger accounts. The authors explain in their limitations that, given limited resources and the time-sensitivity of the outbreak, snowball sampling was chosen instead of random sampling. The sample therefore over-represents students who were online, active on Facebook, and connected to the first students who received the link.

3.6 Bias and sampling error

A sample can miss the truth in two different ways, and they respond differently to more data. *Sampling error* is the scatter that comes from measuring some units rather than all: honest random samples each land a little away from the truth, in random directions. It cannot be avoided, it shrinks as samples grow, and Chapter 3 measures it. *Bias* is a lean in the method itself: every sample drawn the same way misses in the same direction. A larger sample does nothing to bias except make the wrong answer look more certain. Figure 7 draws the difference as four targets. The centre is the truth, each dot is the estimate from one sample, and the four combinations of low and high bias with low and high variability each have their own look.

Definition 3.4 • Sampling error and bias

Sampling error is the difference between a statistic and its parameter that arises because only a sample was measured; across repeated samples it averages out, and it shrinks as n grows. **Bias** is a systematic difference between what a method estimates on average and the parameter; it does not shrink as n grows. **Selection bias** is bias caused by the way units come to be in the sample.

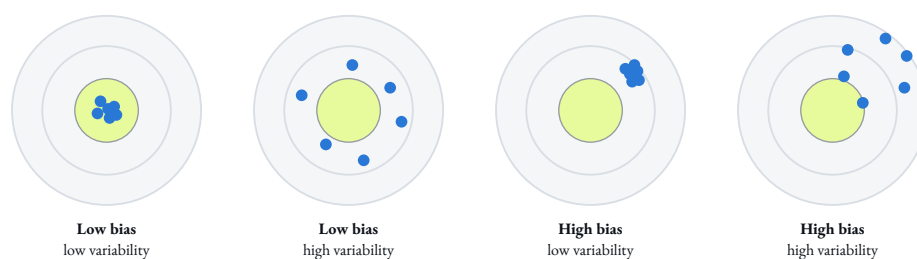


Figure 7: Bias against variability. Dot positions illustrative.

The Literary Digest, 1936

In 1936 the American magazine the *Literary Digest* mailed straw ballots in the millions and predicted that Alf Landon would defeat Franklin Roosevelt in the presidential election. Roosevelt won 523 electoral votes to Landon's 8. A later analysis by Squire (reference 3) concluded that both the sample and the response were biased, and that together they produced the wrong answer. The magazine's list over-represented some groups of voters, and the people who chose to return a ballot differed from those who did not. The size of the mailing did nothing to help: it made a biased estimate precise. The famous ballot counts are not quoted here, because they could not be checked against a primary source for these notes.

3.7 Non-response

A random draw makes a sample fair only until people begin deciding whether to answer. Suppose a perfect random sample of 500 students receives a survey on study stress and 250 complete it. Perhaps the most stressed students answered, because the topic mattered to them; perhaps the calmest did, because they had the time. Either way the 250 who answered probably differ from the 250 who did not, and not at random. That difference is non-response bias, and the second half of the *Literary Digest*'s failure. A study cannot usually tell which direction it runs, which is the point: a low response rate is a warning to be reported, not a verdict. Reporting the rate needs a clear denominator, the sampled units who were eligible to answer.

Non-response bias
bias arising because those who answer differ from those who do not

$$\text{response rate} = \frac{\text{completed}}{\text{sampled} - \text{ineligible}} \quad (3.4)$$

completed the number of sampled units who gave usable answers

sampled the number of units drawn from the frame

ineligible units drawn who turned out not to belong to the population, such as people who had died or moved away

Example 3.4 • Participation in the Sylhet survey

Khanam and colleagues (reference 2) approached 2,039 adults and completed 1,810. Their losses, as printed, were 36 refusals, 56 absent, 76 who had migrated out, 42 who had died, 5 with incomplete data, and 14 women excluded as pregnant: 229 in all.

Example 3.4 • continued

Crude rate. $1,810/2,039 = 0.888$, or 88.8%.

Excluding the ineligible. Those who had died, migrated or were excluded as pregnant were not part of the eligible population at fieldwork: $76 + 42 + 14 = 132$. By equation (3.4), $1,810/(2,039 - 132) = 1,810/1,907 = 0.949$, or 94.9%.

Who was lost among the eligible. $1,907 - 1,810 = 97$, which is the 36 refusals, 56 absent and 5 incomplete. Refused and absent together are 92, or 4.8% of the eligible.

The paper prints the counts. Every rate and the classification of losses as eligible or ineligible are ours.

A rate needs a denominator

The Islam survey cannot report a response rate at all. A link passed from friend to friend has no list of who received it, so nobody knows how many students saw the form and chose not to answer.

3.8 What the class survey can and cannot say

The same scrutiny applies to the course's own data. The students who fill in the class survey chose to join a statistics course run by Stat.BD. That is a convenience sample with a selection nobody can measure: it may hold more final-year students with a thesis to write, or more students from faculties that require statistics. The data can describe the students in the course exactly. They can show how every method in the course works, on answers that mean something to the people analysing them. They can suggest patterns worth studying properly. They cannot describe University of Rajshahi students in general. Every result from the class data in this course is a result about the class, and is labelled that way. Table 5 applies the same questions to the two published surveys.

Table 5: Two Bangladeshi surveys, two ways of finding people.

	Islam 2020	Khanam 2019
Who	university students across Bangladesh	adults aged 35 and over, two rural subdistricts of Sylhet
Frame	none: a link shared online	the site's own 2016 census list
Selection	snowball	computer-generated random numbers
Sample	476	1,810
Response rate	cannot be computed	94.9% of the eligible (ours)
Estimate for everyone?	only with strong assumptions	yes, for the two subdistricts, with a margin of error

Neither survey is simply good or bad

Each answers the question its method allows. Read the frame and the selection method before reading any result.

4 Types of variable and scales of measurement

4.1 Categorical or numerical

The first question to ask of any variable is whether its values are categories or quantities. A categorical variable records which group a unit is in: faculty, residence, blood group. A numerical variable records how much of something a unit has: hours of sleep, monthly income, number of siblings. A quick test separates them: does an average make sense? The average of sleep hours is a meaningful number. The average of faculty codes is not, even when the faculties have been coded 1 to 4. Each kind then splits in two, as Figure 8 shows. Categorical variables are *nominal* when the categories have no order and *ordinal* when they do. Numerical variables are *discrete* when they are counted and *continuous* when they are measured on a scale where any value in between is possible.

Categorical variable
one whose values are groups or labels

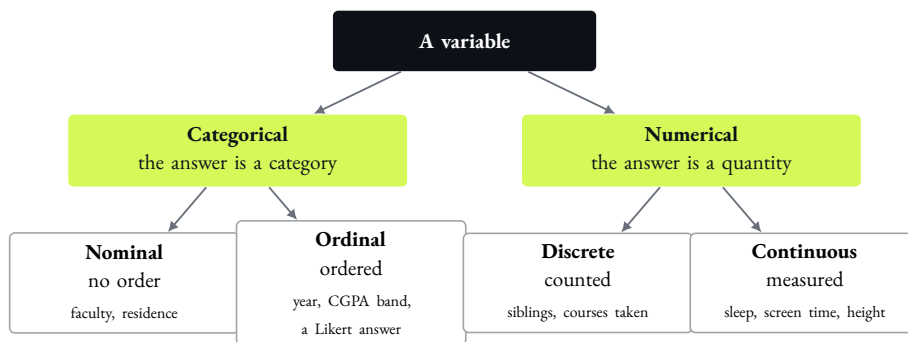


Figure 8: The first two questions to ask of any variable.

Intuition

Ask “category or quantity?” of every column before doing anything else with it. The answer decides which summaries, which charts and which tests are allowed.

4.2 Four scales of measurement

The psychologist S. S. Stevens proposed in 1946 that measurements fall on four scales (reference 4), and his names are still the vocabulary of every statistics package. Figure 9 draws them as a staircase: each step keeps every property of the one below and adds one more. A *nominal* scale only names: faculty, residence, tuition yes or no. An *ordinal* scale adds order: year of study, a CGPA band, an agree-or-disagree answer. An *interval* scale adds equal gaps, so that a difference of one unit means the same anywhere on the scale: temperature in degrees Celsius, the calendar year. A *ratio* scale adds a true zero, a zero that means “none”, so that ratios mean something: eight hours of sleep is twice four. The initials give the order to remember, NOIR. SPSS asks for this property of every variable and calls it the *measure*, with three choices: nominal, ordinal, and scale, where scale covers interval and ratio together.

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

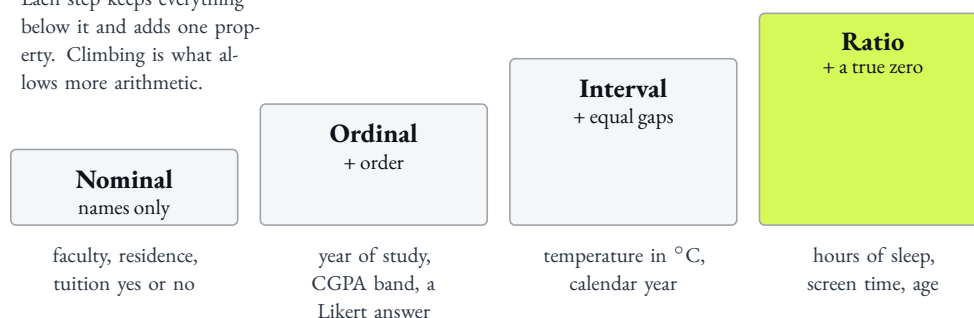


Figure 9: The four scales of measurement. Each step adds one property to the step below.

Definition 4.1 • Scales of measurement

A variable is measured on a **nominal** scale when its values are categories with no order; an **ordinal** scale when the categories have a natural order but the gaps between them are not equal or not known; an **interval** scale when values are numbers with equal gaps but zero does not mean the absence of the quantity; and a **ratio** scale when values are numbers with equal gaps and a true zero.

Three questions, asked in order, place any variable. Does it have an order? Are the gaps equal? Is there a true zero? Stop at the first “no”; the scale is the last “yes”. Table 6 runs the questions on the class survey.

Table 6: Three questions that place a variable on its scale.

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

Example 4.1 • Is year of study ordinal or interval?

Year of study is coded 1 to 5 in the class survey: first year to fourth year, then Masters. The codes are ordered, so the variable is at least ordinal. Are the gaps equal? The step from first to second year is one year of the same degree. The step from fourth year to Masters is a change of programme, often with a gap between them, and not the same kind of step. The gaps are not equal, so the answer to the second question is “no” and the variable is **ordinal**, even though it is written as a number.

4.3 Discrete and continuous

A *discrete* numerical variable takes whole-number values from counting: the number of siblings, the number of courses failed, the number of times a phone was checked before break-

fast. A *continuous* variable can take any value in a range, limited only by the precision of the instrument: sleep, height, screen time. Every continuous variable is recorded discretely in practice, because every recording is rounded. The class survey records sleep to the nearest half hour, so the recorded values are 6.0, 6.5, 7.0 and so on; the variable is still continuous, and only the recording is coarse. The distinction matters less than it first seems. A count with many possible values, such as the number of books read in a year, is usually analysed as though it were continuous. A count with only a few values, such as the number of siblings, often needs more care.

Counted or measured

Discrete values are counted; continuous values are measured. What matters is the underlying quantity, not the rounding used to write it down.

4.4 Cases that cause arguments

Three variables in the class survey cause genuine disagreement, and the disagreements are instructive. *CGPA itself* is numerical, but the survey records a band such as 3.00 to 3.49, and a band is ordinal: the bands are ordered, but a band is not a number, and the detail inside it is gone. *Year of study* is coded 1 to 5, and the codes are ordered but the gaps are unequal (Example 4.1). *A single Likert answer*, from strongly disagree coded 1 to strongly agree coded 5, is ordinal: nobody can say that the step from “neutral” to “agree” is the same size as the step from “agree” to “strongly agree”. The *total* of five such answers is another matter. It takes many values, spreads out, and is commonly treated as numerical. Whether that is justified is a long-running debate among methodologists, and Chapter 6 builds the class stress total and asks whether five statements measure one thing.

A number on a form is not a number

Codes are labels that happen to be digits. SPSS, R and Python will all compute a mean faculty code of 2.6 without complaint. The software does not know what the codes mean; the analyst must.

4.5 Interval against ratio

The difference between interval and ratio scales shows most clearly with temperature. Zero on the Celsius scale is the point where water freezes, not the absence of heat. So is a day at 30 °C twice as hot as a day at 15 °C? Converting both to Fahrenheit settles it.

$$F = 1.8 C + 32 \quad (4.1)$$

F the temperature in degrees Fahrenheit

C the temperature in degrees Celsius

Example 4.2 • A ratio that depends on the unit

By equation (4.1), 30 °C is $1.8 \times 30 + 32 = 86$ °F and 15 °C is $1.8 \times 15 + 32 = 59$ °F. In Celsius the ratio of the two days is $30/15 = 2$. In Fahrenheit it is $86/59 = 1.46$. A



Example 4.2 • continued

ratio that changes when the unit changes is a property of the scale's arbitrary zero, not of the weather.

The difference survives. The gap is 15 Celsius degrees, and $86 - 59 = 27 = 1.8 \times 15$ Fahrenheit degrees: the same gap in a different unit. On an interval scale differences mean something and ratios do not.

Contrast with sleep. Eight hours is twice four hours, and 480 minutes is twice 240 minutes. A true zero makes the ratio the same in every unit.

Very few variables in the social sciences are interval without also being ratio. The distinction earns its place mainly by exposing nonsense ratios. A student who scores 80 on a test is not “twice as knowledgeable” as one who scores 40, because a score of zero does not mean zero knowledge.

4.6 Bands lose information

Once a number is grouped into bands, the detail inside each band is gone for good, and the edges of the bands create false jumps. Two students with CGPAs of 3.01 and 3.49 land in the same band, while two with 2.99 and 3.01 land in different ones. Record sleep as “under 6 hours” or “6 or more” and a student sleeping two hours looks the same as one sleeping five and a half. Some analyses become impossible: no average sleep can be computed from two bands. The rule for questionnaire design follows directly. Record the finest detail that can ethically be collected, because it is always possible to group later and never possible to ungroup.

A deliberate exception

The class survey bands CGPA on purpose. An exact CGPA, combined with department and year, could identify a student in a small department. Banding trades information for privacy, knowingly. Sleep and screen time are recorded as numbers on purpose, for analysis.

4.7 The scale decides the analysis

Table 7 is a map of the next five chapters organised by the idea of this section. The scale on the left determines everything to its right: which summary describes the variable, which chart displays it, and which test compares or relates it. None of the tests needs to be understood yet. What matters is the shape of the table. A variable classified wrongly is summarised wrongly, drawn wrongly and tested wrongly, and every later step inherits the mistake.

Table 7: How the scale of measurement decides the analysis.

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

5 From a questionnaire to a dataset

5.1 A questionnaire is a measuring instrument

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, and a question that nudges people towards one answer is biased in exactly the same way. The most careful analysis of a leading question produces a precise estimate of the wrong thing, which is the Literary Digest’s lesson arriving by a different route: bias through wording rather than through selection. Students often treat the questionnaire as a formality to be finished the night before data collection. It is the instrument on which every later number depends. Table 8 lists five faults that recur in student questionnaires, each with a repair.

Table 8: Five ways to write a bad question.

Fault	Example	Repair
Double-barrelled	Are the library and the canteen satisfactory?	ask two questions
Leading	Don’t you agree that 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

The double-barrelled question is the commonest. A student who likes the library and dislikes the canteen has no honest answer to give. The last row matters most for data: a student who sleeps exactly six hours can tick two boxes, and the fault stays invisible until someone tries to analyse the answers.

5.2 Answer options

Every closed question must obey two rules. Its options must be *exhaustive*, so that every respondent has a box to tick, and *mutually exclusive*, so that no respondent has two. “Residence: hall or family home” fails the first rule, because a student living in a mess has nowhere to go; the class survey’s residence question was nearly written that way. “CGPA: 2.50–3.00

or 3.00–3.50” fails the second, because a CGPA of exactly 3.00 belongs to both. When in doubt, add “Other (please write)” and code the answers later. Closed questions produce answers that are quick to code. Open questions produce richer answers that must be coded by hand into categories decided afterwards, and the codebook must record how that was done so that another person could repeat it.

Definition 5.1 • Exhaustive and mutually exclusive options

A set of answer options is **exhaustive** when every possible respondent can choose at least one, and **mutually exclusive** when no respondent can truthfully choose more than one.

5.3 Likert items and the reversed item

A Likert item is a statement with ordered answers, usually five, from strongly disagree to strongly agree. The class survey has five such items about study stress, listed in Table 9. Four are worded so that agreeing means more stress: “I worry about my exam results.” The fourth item is worded the other way on purpose: “I feel confident I can handle my coursework.” Agreeing with it means *less* stress. The reversed wording catches respondents who tick “agree” all the way down without reading. A student who agrees both that they feel overwhelmed and that they feel confident is probably not reading. The cost of the design is a step in the data. Before the five answers are added into a total, the reversed item must be flipped.

Likert item

a statement answered on an ordered agree-or-disagree scale

$$x' = (K + 1) - x \quad (5.1)$$

x' the reverse-coded answer

x the answer as given

K the number of answer options; for a five-point item $K = 5$, so $x' = 6 - x$

Example 5.1 • Scoring one student

A student answers the five stress items 4, 3, 4, 2, 5. Item four is the reversed one. By equation (5.1), its answer becomes $6 - 2 = 4$: disagreeing that one feels confident is, on this scale, a sign of stress. The stress total is $4 + 3 + 4 + 4 + 5 = 20$.

Without the reversal the total is $4 + 3 + 4 + 2 + 5 = 18$. The item then works against the other four: a more stressed student, who disagrees more strongly that they feel confident, gets a *lower* total. The student is constructed for teaching.

Forgetting to reverse

Forgetting to reverse-code is one of the commonest errors in student analyses, and the software gives no warning. Chapter 6 measures the damage on the class data.

5.4 Consent and the least data that will do

Ethics is part of measurement, not paperwork beside it. People answer honestly when they trust that their answers cannot be traced back to them. Four practices follow. Tell respon-

dents what the survey is for and who will see the answers. Make taking part voluntary, and allow any question to be skipped. Collect no names, student identification numbers, phone numbers or email addresses unless the study genuinely needs them. Report groups, never individuals. The third practice is the one most often broken. Every extra identifying detail is a risk with no benefit, and details combine: a department, a year and an exact CGPA together could identify one student in a small department. That is why the class survey records faculty groups and CGPA bands. Most university research that surveys people also needs approval from a supervisor or an ethics committee before any data are collected.

ISLAM 2020: CONSENT AND ANONYMITY

The study was approved by the Ethical Clearance Committee of Khulna University. A consent form opened the e-questionnaire, respondents answered anonymously, and the paper notes that contact details were not collected because of ethical requirements on anonymity and confidentiality.

5.5 The class survey

Table 9 lists the fourteen questions of the class survey. Each design choice discussed in this chapter appears in it. The department question is open, and the coder assigns a faculty group afterwards. Sleep and screen time are recorded as numbers, not bands. CGPA is banded, for privacy. There are two sleep questions, one for weeknights and one for weekends, so that Chapter 4 can compare the two in the same students. And the stress items include one reversed item, s4. The five stress statements were written for this course. They are not a published, validated scale, and they are used here to teach how a scale is built and checked, not to measure stress in any clinical sense.

Table 9: The fourteen questions of the Stat.BD class survey.

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.

5.6 The codebook

A column of 1s, 2s and 3s is unreadable without a dictionary, and the codebook is that dictionary. For each variable it records the name used in the dataset, the question the variable

Codebook

the dictionary that says what every column and every code in a dataset means

came from, word for word, its type and scale, every code and its meaning, the units and the allowed range, and how missing answers are coded. Its value is practical. Six months after data collection nobody remembers whether residence code 2 was a mess or a family home: not the researcher, not the supervisor, and not the software. Students who skip the codebook discover at the analysis stage that they cannot say which way round a variable was coded, and the only remedy is to return to the paper forms, if they still exist. Table 10 is the codebook for the class survey.

Definition 5.2 • Codebook

A **codebook** lists, for every variable in a dataset, its name, its source question, its type and scale of measurement, its codes and their meanings, its units and allowed range, and its missing-value codes.

Table 10: Codebook for the class survey. Blank answers are coded 99 in every column. The faculty grouping is a simplification made for teaching.

Name	Question	Scale	Codes and units
faculty	1	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

The allowed ranges are not decoration. Nobody sleeps twenty hours on a typical weeknight, and a range written down in advance is what lets a checking rule catch a typing error later (section 5.9). In SPSS much of the codebook lives inside the data file, as variable labels and value labels; Session 7 sets them up from this table.

5.7 Two kinds of missing

A first-semester student has no CGPA. The CGPA question has no answer for that student, which is *not applicable*, coded 9. A third-year student who leaves the CGPA question blank does have a CGPA; the study simply does not know it. That is *missing*, coded 99. Both leave the cell empty for analysis, but they mean different things when results are reported. A result about CGPA describes only students who have one, and a reader needs to know how many students were left out for each reason. In the simulated class data, 15 students have no result yet and 3 did not know their screen time. The practical danger is arithmetic. A code of 99 left in a column of hours is read as 99 hours unless the software is told otherwise. Chapter 2 shows how far one wrong value moves an average, and Chapter 5 shows three undeclared codes of 99 reversing the sign of a correlation.

Definition 5.3 • Missing and not applicable

A value is **missing** when the unit has a true value that was not recorded. It is **not applicable** when the question has no value for that unit. A **missing-value code** is a number reserved to mark either case, and must be declared to the software so that it is never treated as a measurement.

A missing code is not a number

Every package must be told which codes mean “no value”. Sessions 7 to 12 show how in SPSS, R and Python. Until that is done, every mean, SD and chart that includes the column is wrong.

5.8 From a form to a row

Coding is the step where answers become data, and it is also the step where most errors enter. Figure 10 follows one constructed student from the paper form, through the codebook, to a row of the dataset. Two operations in it deserve attention. The department, Sociology, becomes faculty code 4: that step applies the codebook’s grouping rule, and a coder without the rule could reasonably choose differently. The screen time, 5 hours 12 minutes, becomes 5.2 hours, because 12 minutes is a fifth of an hour. Both operations are small, and both can go wrong in ways nobody notices later. Converting hours and minutes to decimal hours is the commonest arithmetic in coding.

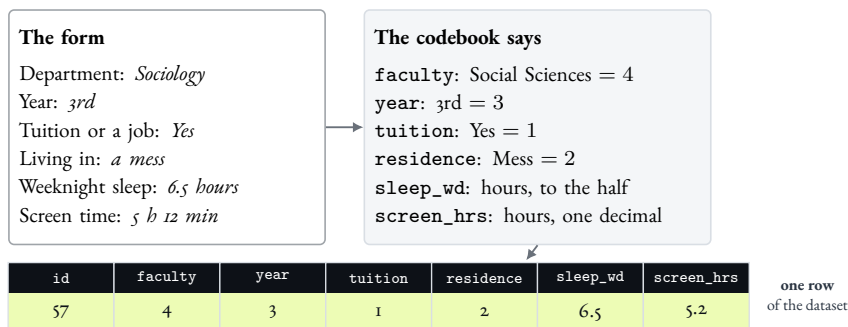


Figure 10: One student’s answers, the codebook’s rules, and the row that results. The student is constructed for the picture.

$$\text{decimal hours} = b + \frac{m}{60} \quad (5.2)$$

b the whole hours reported

m the minutes reported beyond the whole hours

Example 5.2 • Two conversions

Screen time of 5 hours 12 minutes. By equation (5.2), $5 + 12/60 = 5 + 0.2 = 5.2$ hours, which is already at one decimal.

Screen time of 5 hours 45 minutes. $5 + 45/60 = 5.75$ hours. The codebook asks for one decimal, so 5.75 is rounded to 5.8.

Example 5.2 • continued

A trap. A coder who writes 5.45 for “5 hours 45 minutes” has read the minutes as a decimal. The value is 0.3 hours short, about 18 minutes, and nothing in the dataset will ever reveal it.

5.9 Checking the data

Before any analysis, a dataset should be checked against its codebook. The checks are dull, mechanical, and the place where analyses are won or lost. Because they are mechanical, software can run all of them in seconds once the codebook exists.

1. **Range:** is every value inside the codebook’s allowed range?
2. **Consistency:** do answers contradict each other? A student who reports more week-end sleep than a night contains, or a first-semester student with a CGPA band, is a contradiction to investigate.
3. **Codes:** is every category code one that the codebook defines? A residence code of 4 is an error, because only 1, 2 and 3 exist.
4. **Missing:** are 9 and 99 declared as missing, not counted as values?
5. **Double entry:** for paper forms, type everything twice, by different people if possible, and compare the two files.

A typing error, planted on purpose

One row in the simulated class data records self-study of 25 hours a day. The form said 2.5; the typist missed the decimal point. A range check against “allowed 0 to 16” catches it at once. Chapter 2 opens with this row and shows how far one wrong value in 120 moves an average. The entry was planted in the simulated data deliberately, for that purpose.

6 Chapter summary

1. Statistics exists because people differ. It does three jobs: describing what was measured, estimating what was not, and testing whether a pattern is larger than chance would produce.
2. A descriptive statement is about the units measured; an inferential statement is about a population. The step between them is only as safe as the method that chose the sample.
3. When a paper prints a count and a percentage, divide. The Islam survey’s 87.7% should read 81.7%.
4. An answerable question names who, what, compared with what, and measured how. Questions describe, compare or relate, and the kind decides the analysis.
5. A dataset has one row per unit and one column per variable. The unit is not always a person. Roles such as outcome and explanatory belong to the question, not to the

variable.

6. A construct is not its measurement. A screening score is not a diagnosis.
7. An analysis can only allow for variables that were measured.
8. The population is who the question is about; the frame is the list drawn from; the sample is who was measured. Parameters describe populations, statistics describe samples.
9. Probability samples let chance choose, with known chances: simple random, systematic, stratified and cluster, each with its own cost. Non-probability samples are common and must be reported as what they are.
10. Sampling error shrinks as samples grow; bias does not. Non-response can bias a perfectly drawn sample, and a response rate needs a stated denominator.
11. Every variable is categorical or numerical, and sits on a nominal, ordinal, interval or ratio scale. The scale decides the summary, the chart and the test.
12. Good questions are single, neutral, specific and free of assumptions, with options that are exhaustive and mutually exclusive. Reversed items must be reverse-coded before scoring.
13. A codebook makes the data readable by anyone. Not applicable and missing are coded differently, and every missing code must be declared to the software.
14. Check range, consistency, codes and missing values before any analysis.

7 Key terms

Bias

A systematic difference between what a method estimates on average and the parameter.
It does not shrink as the sample grows.

Categorical variable

A variable whose values are groups: nominal or ordinal.

Census

A measurement of every member of the population.

Cluster sample

A sample of whole natural groups, chosen at random.

Codebook

The dictionary of a dataset: names, questions, scales, codes, units, ranges and missing codes.

Construct

The idea a study cares about, as distinct from its measurement.

Continuous and discrete

Measured on a scale where any value is possible, and counted in whole numbers.

Descriptive and inferential statements

Claims about the units measured, and claims about a population from a sample.

Exhaustive, mutually exclusive

Every respondent has an option; no respondent has two.

Interval scale

Ordered, with equal gaps, and no true zero.

Likert item

A statement answered on an ordered scale, usually five points from strongly disagree to strongly agree.

Missing and not applicable

A true value not recorded; a question with no value for that unit.

Nominal scale

Categories with no order.

Non-probability sample

Convenience, snowball, quota or purposive: chosen without known chances of selection.

Non-response bias

Bias arising because those who answer differ from those who do not.

Ordinal scale

Ordered categories whose gaps are not equal or not known.

Outcome and explanatory variable

What a question tries to explain, and what might explain it.

Parameter and statistic

A number describing a population (μ , π), and a number computed from a sample ($\bar{x}$, $\hat{p}$).

Population and sample

The group a question is about, and the part of it measured.

Probability sample

A sample chosen by a random mechanism, every unit having a known, non-zero chance.

Ratio scale

Ordered, with equal gaps and a true zero.

Response rate

Completed responses divided by eligible sampled units, equation (3.4).

Reverse coding

Flipping an item worded the other way, $x' = (K + 1) - x$.

Sampling error

The difference between a statistic and its parameter that arises because only a sample was measured.

Sampling frame

The list or procedure from which a sample is drawn.

Simple random sample

A sample in which every set of n units is equally likely.

Stratified sample

Random samples drawn separately inside groups known in advance.

Systematic sample

Every k -th unit from a random start, with $k = N/n$.

Unit, observation, variable, value

What a row describes, a row, a column, a cell.

Variation

The differences between members of a group on the same characteristic.

8 Exercises

Exercises in Part A need no calculation. Parts B and C need a calculator. Numbers not attributed to a paper are constructed for teaching.

Part A. Ideas.

1. Say whether each statement describes, estimates or tests. (a) “Of the 120 students who answered, 64 live in a hall or a mess.” (b) “Students at this university probably average between 6.6 and 7.0 hours of sleep.” (c) “The gap in sleep between hall and family-home students is larger than chance would readily produce.”
2. A survey of 300 students at one college, recruited through a class WhatsApp group, finds that 45% skip breakfast. Write one descriptive statement and one inferential statement from this result, and say what the inferential one would need to be safe.
3. Rewrite “Does private tuition hurt students’ studies?” as an answerable question. Label its four parts.
4. Name the unit in each study. (a) A comparison of rice yield across fifty test plots. (b) A count of English loanwords in 200 newspaper editorials. (c) A study relating literacy rate to average household income across the 64 districts of Bangladesh. (d) For study (c), explain why its result cannot be read as a statement about individual people.
5. In each question, name the outcome and the explanatory variable. (a) Does year of study predict CGPA? (b) Do students with higher stress scores sleep less? (c) Does sleep predict stress score? Comment on (b) and (c) together.
6. Classify each variable as categorical or numerical, and then as nominal, ordinal, interval or ratio. Where it is numerical, say also whether it is discrete or continuous. (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”, from strongly disagree (1) to strongly agree (5). (9) Calendar year of admission. (10) Distance from home district, in kilometres.

Part B. Samples.

7. Name the sampling method. (a) Every student in five randomly chosen departments is surveyed. (b) The registrar’s list is split by faculty and 50 students are drawn at random from each. (c) A researcher surveys students in the canteen at lunchtime. (d) Each respondent is asked to forward the link to three friends. (e) Starting from the 9th name, every 20th name on the hall register is taken. (f) Interviewers must find 40 men and 40 women, and approach anyone passing until each quota is full.

8. A hall register lists 1,250 residents and a survey wants 50. (a) Find the sampling interval. (b) With a random start of 17, list the first five positions chosen. (c) The register lists one resident per line, and every fifth line, positions 5, 10, 15 and so on, is the resident of a single room. Which random starts would give a sample made entirely of single-room residents, and which would give none?
9. Four faculties have 2,400, 1,800, 3,000 and 2,800 students. (a) Allocate a sample of 250 proportionally. (b) A researcher instead takes 80 from the smallest faculty. What must the analysis then do, and why?
10. A survey of students is drawn at random from the list of hall residents. Name two groups of students in the population who are missing from this frame, and say how each might differ from hall residents in sleep or study habits.
11. A student doubles the size of an online survey shared in a Facebook group, from 400 to 800 respondents. Which of the two kinds of error in Definition 3.4 does this reduce, and which does it leave alone? Use the Literary Digest to support the answer.
12. A random sample of 500 students is drawn from the enrolment list. Twenty turn out to have left the university before the survey began, 168 do not reply, and 312 complete the form. (a) Compute the crude completion rate and the response rate of equation (3.4). (b) Which is the right figure to report, and why?

Part C. Reading a paper.

13. Islam and colleagues report the PHQ-9 bands for their 476 students as 84, 136, 133, 72 and 51, from none-to-minimal up to severe. (a) Check that the bands add to 476. (b) Compute the percentage with moderate or worse symptoms, a score of 10 or more. (c) The abstract says “around 15%” had moderately severe depression. Check it.
14. Table 1 of the same paper reports 320 male and 156 female students. (a) Check the printed percentage of 32.8% female. (b) Give two reasons, from the way the sample was found, why this share might differ from the share of women among Bangladeshi university students.
15. Using Table 5, explain why Khanam and colleagues can put a margin of error on their prevalence for the two subdistricts and why Islam and colleagues cannot do the same for Bangladeshi students, even though both are large samples.
16. A newspaper headline reads “82% of university students are depressed”. Identify two separate problems with the headline, one about the population and one about the measurement.

Part D. From questions to data.

17. Name the fault in each question and rewrite it. (a) “How often and how long do you use social media?” (b) “Don’t you think the library should open longer?” (c) “Do you sleep enough?” (d) “Have you stopped copying assignments?”

- 18.** Repair each set of options. (a) Daily screen time: 0–2, 2–4, 4–6 hours. (b) “Where do you study most? Library / Home.” (c) Monthly spending: below 3,000; 3,000–5,000; above 6,000 taka.
- 19.** A student answers the five stress items s_1 to s_5 as 5, 4, 5, 2, 4. (a) Compute the stress total correctly. (b) Compute it with s_4 left unreversed. (c) Another student answers 5 to all five items. What is that student’s correct total, and what do the answers suggest?
- 20.** Code this form into a row, using Table 10. Department: Finance, which the codebook places in Business Studies. Year: Masters. Tuition or job: no. Lives: family home. Self-study: 3 hours. Weeknight sleep: 7 hours 20 minutes. Weekend sleep: 9 hours. Screen time: 3 hours 40 minutes. CGPA: 3.52. Stress items: 3, 4, blank, 2, 3.
- 21.** Missing or not applicable? Give the code. (a) A first-semester student and the CGPA question. (b) A third-year who leaves the CGPA question blank. (c) A student who ticks “don’t know” for screen time. (d) A student who leaves s_3 blank.
- 22.** Five rows have been typed from paper forms. Using Table 10, find every value that a checking rule would flag, and name the rule.

id	faculty	year	residence	study_hrs	sleep_wd	sleep_we	cgpa_band	s2
201	3	2	1	25	6.5	8.0	3	4
202	1	1	2	3.0	7.0	18	9	3
203	2	4	4	2.5	6.0	7.5	2	5
204	4	3	3	1.5	5.5	6.5	3	6
205	5	2	1	4.0	7.5	8.5	3	2

9 Solutions

Part A.

1. (a) Describes: a count of the students who answered. (b) Estimates: a range for students who were not all measured. (c) Tests: a judgement about whether a pattern exceeds chance. (Exercise 1.)
2. Descriptive: “Of the 300 students who answered, 45% skip breakfast.” Inferential: “About 45% of students at this college skip breakfast.” The inferential statement needs the 300 to resemble the college’s students as a whole. A WhatsApp group is a convenience sample, so it would need at least a check that the respondents’ years, departments and sex match the college’s figures, and it would still be open to selection the check cannot see. (Exercise 2.)
3. One version: “Among third- and fourth-year students at the University of Rajshahi, do students who do private tuition have lower CGPAs than students who do not?” **Who:** third- and fourth-year RU students. **What:** doing private tuition. **Compared with what:** students who do no tuition. **Measured how:** CGPA from the latest result. Other versions are equally good if all four parts are fixed. (Exercise 3.)
4. (a) A test plot. (b) An editorial. (c) A district. (d) Each row holds a district’s figures, so the study shows how literacy and income go together *across districts*. Richer districts might have higher literacy while, within every district, the richer households are no more literate than the poorer ones. Reading a district-level pattern as a fact about individuals is the ecological fallacy. (Exercise 4.)
5. (a) Outcome CGPA; explanatory year of study. (b) Outcome sleep; explanatory stress score. (c) Outcome stress score; explanatory sleep. Questions (b) and (c) use the same two variables with their roles swapped. The data would look the same for both; only the question decides which variable is the outcome. (Exercise 5.)
6. (1) Categorical, nominal. (2) Numerical, ratio; treated as continuous, since money comes in very small steps. (3) Categorical, ordinal. (4) Numerical, discrete, ratio. (5) Categorical, ordinal: the gap between first and second need not equal the gap between fiftieth and fifty-first. (6) Categorical, nominal. (7) Numerical, continuous, ratio. (8) Categorical, ordinal. (9) Numerical, discrete, interval: the gaps between years are equal, but year zero is not the beginning of time, so “twice as late” means nothing. (10) Numerical, continuous, ratio. (Exercise 6.)

Part B.

7. (a) Cluster: whole departments are taken. (b) Stratified, with faculty as the strata. (c) Convenience. (d) Snowball. (e) Systematic, with $k = 20$ and a start of 9. (f) Quota: the numbers per group are fixed, but the people within each group are found by convenience. (Exercise 7.)

8. (a) $k = 1,250/50 = 25$. (b) 17, 42, 67, 92, 117. (c) Because 25 is a multiple of 5, the start and every later position leave the same remainder when divided by 5. A start of 5, 10, 15, 20 or 25 therefore selects only lines that are multiples of 5, which are all single-room residents, and any other start selects none of them. (Exercise 8.)

9. (a) $N = 10,000$ and $n/N = 250/10,000 = 0.025$. By equation (3.3): $0.025 \times 2,400 = 60$; $0.025 \times 1,800 = 45$; $0.025 \times 3,000 = 75$; $0.025 \times 2,800 = 70$. Total 250. (b) The smallest faculty now has 80 places instead of its proportional 45, so it is over-represented. Any figure for all students combined must weight that faculty's answers down, here by $45/80$, or it will lean towards the smallest faculty's answers. (Exercise 9.)

10. Students living in messes and students living at home with family are both absent. Mess residents may share rooms with more people and keep later hours; students at home may commute, lose time to travel and sleep on a family timetable. Either difference could make the hall-only sample misstate sleep for students as a whole. (Exercise 10.)

11. Doubling the sample reduces sampling error: the estimate scatters less from one sample to the next. It does nothing to bias. The survey still reaches only members of that Facebook group who chose to answer, so it will estimate that group's figure more precisely and remain as far as before from the figure for all students. The Literary Digest mailed ballots in the millions and still predicted the wrong winner, because size cannot correct a biased list or a biased response. (Exercise 11.)

12. (a) Crude: $312/500 = 0.624$, or 62.4%. Response rate: the 20 who had left were not in the population, so the eligible sample is $500 - 20 = 480$, and $312/480 = 0.650$, or 65.0%. (b) Report 65.0%, with the denominator stated, because the rate is meant to describe how many eligible people answered. The 168 non-responders are the figure to worry about, since they may differ from those who replied. (Exercise 12.)

Part C.

13. (a) $84 + 136 + 133 + 72 + 51 = 476$. (b) Moderate or worse: $133 + 72 + 51 = 256$, and $256/476 = 0.538$, or 53.8%. (c) $72/476 = 0.151$, or 15.1%, which Table 1 prints and the abstract rounds to "around 15%". The sums and part (b) are ours. (Exercise 13.)

14. (a) $156/476 = 0.328$, or 32.8%, as printed. (b) The link spread from friend to friend through Facebook and Messenger, so the sample's make-up follows the networks of the first students who received it; if those were mostly men, the sample would be too. Men and women may also differ in how active they are on Facebook, or in how willing they are to answer a survey about mental health. Neither reason can be checked from the paper, which is itself the point. (Exercise 14.)

15. A margin of error measures sampling error, and it has a meaning only when chance chose the sample with known probabilities. Khanam and colleagues drew at random from a census list of the two subdistricts, so their sample's departures from the population are random and can be measured. Islam and colleagues had no frame and no random selection, so the departures of their sample from all students include selection bias of unknown size and direction, which no margin of error can express. Size does not change this. (Exercise 15.)

16. Population: the 82% describes 476 students reached by a snowball sample during a lockdown, not university students in general. **Measurement:** 82.4% had mild to severe *symptoms* on a screening questionnaire, a PHQ-9 score of 5 or more; the figure includes mild symptoms and is not a diagnosis of depression. A careful version: “In an online survey of 476 Bangladeshi students in May 2020, 82% reported at least mild depressive symptoms on a screening questionnaire.” (Exercise 16.)

Part D.

17. (a) Double-barrelled: “How many days in a typical week do you use social media?” and “On a day when you use it, how many hours?” (b) Leading: “Should the library’s opening hours be shorter, stay the same, or be longer?” (c) Vague: “How many hours do you sleep on a typical weeknight?” (d) Loaded, since it assumes the respondent copied: “In this semester, have you submitted an assignment that was mostly copied from someone else?” (Exercise 17.)

18. (a) Not mutually exclusive: 2 and 4 each belong to two options, and there is no option above 6. Use 0 to under 2, 2 to under 4, 4 to under 6, 6 or more; better still, ask for the number of hours. (b) Not exhaustive: add hall room, mess, and “Other (please write)”. (c) Not exhaustive: 5,001 to 6,000 has no option. Use below 3,000; 3,000 to 5,999; 6,000 or more. (Exercise 18.)

19. (a) s_4 reversed: $6 - 2 = 4$. Total $5 + 4 + 5 + 4 + 4 = 22$. (b) Unreversed: $5 + 4 + 5 + 2 + 4 = 20$. (c) s_4 reversed: $6 - 5 = 1$, so the total is $5 + 5 + 5 + 1 + 5 = 21$. The student strongly agrees both that they feel overwhelmed and that they feel confident, which is contradictory. The likeliest explanation is that the student ticked the same column without reading, which the reversed item exists to reveal. (Exercise 19.)

20. faculty 2; year 5; tuition 0; residence 3; study_hrs 3.0. Weeknight sleep: $7 + 20/60 = 7.33$ hours, and the nearest half hour is 7.5. Weekend sleep 9.0. Screen time: $3 + 40/60 = 3.67$, which is 3.7 to one decimal. cgpa_band 4, since 3.52 lies in 3.50 to 4.00. Stress items s_1 to s_5 : 3, 4, 99, 2, 3, with s_4 stored as answered and reversed only when the total is computed. (Exercise 20.)

21. (a) Not applicable, code 9. (b) Missing, code 99. (c) Missing, code 99: the student has a screen time but does not know it. (d) Missing, code 99. (Exercise 21.)

22. Row 201: study_hrs 25 fails the range rule (0 to 16), probably 2.5. Row 202: sleep_we 18 fails the range rule. Its cgpa_band of 9 needs no flag, since a first-year with no result yet is expected. Row 203: residence 4 fails the code rule; only 1, 2 and 3 exist. Row 204: s_2 of 6 fails the code rule; items run 1 to 5. Row 205: faculty 5 fails the code rule; only 1 to 4 exist. Each flagged value is checked against the paper form before it is corrected. (Exercise 22.)

10 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
392	136 + 133 + 72 + 51, checking the printed count	Islam 2020
81.7%	389/476, against the printed 87.7%	Islam 2020
389; 476; 81.8	GAD-7 band sums; printed band percentages summed	Islam 2020
256; 53.8%	moderate or worse depressive symptoms (Exercise 13)	Islam 2020
88.8%; 94.9%; 1,907; 97; 92, 4.8%	participation rates and losses among the eligible	Khanam 2019
86, 59; 1.46; 27	Fahrenheit conversions, their ratio and difference	none

Simulated, not real students: every figure from the class data, including Figure 1, Table 2, the mean of 6.80 hours, the interval from 6.63 to 6.97 hours (computed as Chapter 3 teaches), and the counts of 15 and 3. They are generated by data/class-data.R from a fixed seed and will be replaced by the real survey.

Constructed for teaching, and not from any study: the department list of 600 and its years (Examples 3.2 and 3.3); the student in Example 5.1 and in Figure 10; the third-variable picture in Figure 4; the faculty examples; and every number in the exercises that is not attributed to a paper.

Not quoted, because it could not be verified: the Literary Digest's ballot counts.

11 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.
- Khanam R, Ahmed S, Rahman S, Kibria GMA, Syed JRR, Khan AM, Moin SMI, Ram M, Gibson DG, Pariyo G, Baqui AH. 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. PMC6830635. doi:10.1136/bmjopen-2018-026722. Licensed CC BY-NC: its numbers are re-typeset here with attribution; read it at PMC.
- Squire P. Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly* 1988;52(1):125–133. doi:10.1086/269085. Cited for its conclusion only; subscription journal.
- Stevens SS. On the theory of scales of measurement. *Science* 1946;103(2684):677–680. PMID 17750512. doi:10.1126/science.103.2684.677. Cited for the four scales; subscription journal.



5. US National Archives and Records Administration. Electoral College results, 1936: Roosevelt 523, Landon 8. A government record; it has no DOI or PMID.
6. 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.

12 For students in the Stat.BD course

Before the next session. Read this chapter, and work the Part A exercises (Exercises 1 to 6). The rest are for practice at any point in the course. Bring a calculator.

What happens to the survey. The anonymous class survey completed in Session 1 is coded with the codebook in Table 10, with the groups' coding rules from the codebook activity folded in, and becomes the class dataset for every later session. Until it is ready, the simulated stand-in is used, and it is labelled as simulated wherever it appears.

What the next session does with this chapter. Session 2 describes the class dataset: frequency tables, averages and spread, box plots and histograms, and the charts that mislead. It opens with the row in which one student appears to study 25 hours a day, and asks how far one wrong value moves the average of 120.

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

Question	Where it is settled
How far does one wrong value, or an undeclared 99, move an average?	Chapter 2
How is each scale summarised and charted?	Chapter 2
The second slip in the Islam paper's tables	Chapter 2
How was the interval from 6.63 to 6.97 hours computed?	Chapter 3, the confidence interval
How is sampling error measured, and why does it shrink with n ?	Chapter 3, the standard error
Is a gap in sleep between groups larger than chance would produce?	Chapter 4
Weeknight against weekend sleep in the same students	Chapter 4, paired data
Does residence go with tuition? Does screen time go with sleep?	Chapter 5
Is a link a cause? Confounding and the ecological fallacy	Chapter 5
Allowing for a third variable	Chapter 6, multiple regression
Do the five stress statements measure one thing?	Chapter 6, survey scales
What does forgetting to reverse s4 cost?	
Value labels and missing codes in software	Session 7 onward